ISSN 1001-5965
CODEN BHHDE8

北京航空航天大學学报

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

2019-12
Vol.45 No.12

北 京 航 空 航 天 大 学 学 报

第 45 卷　第 12 期　（总第 322 期）　2019 年 12 月

目　　次

期刊基本参数： CN 11-2625/V * 1956 * m * A4 * 198 * zh * P * ￥50.00 * 900 * 22 * 2019-12

第 45 卷终

（**编　辑**　张　嵘　李　晶　张欣蔚　孙　芳　王艳梅　贺　伟）

JOURNAL OF BEIJING UNIVERSITY OF
AERONAUTICS AND ASTRONAUTICS

Vol. 45 No. 12 (Sum 322) December 2019

CONTENTS

The End of Volume 45

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0363

基于生成对抗网络的零样本图像分类

魏宏喜*，张越
（内蒙古大学 计算机学院，呼和浩特 010021）

摘 要：在图像分类任务中，零样本图像分类问题已成为一个研究热点。为了解决零样本图像分类问题，采用一种基于生成对抗网络（GAN）的方法，通过生成未知类的图像特征使得零样本分类任务转换为传统的图像分类任务。同时对生成对抗网络中的判别网络做出改进，使其判别过程更加准确，从而进一步提高生成图像特征的质量。实验结果表明：所提方法在AWA、CUB和SUN数据集上的分类准确率分别提高了0.4%、0.4%和0.5%。因此，所提方法通过改进生成对抗网络，能够生成质量更好的图像特征，从而有效解决零样本图像分类问题。

关 键 词：深度学习；图像分类；零样本学习；生成对抗网络（GAN）；图像特征生成
中图分类号：TP391
文献标识码：A **文章编号**：1001-5965(2019)12-2345-06

随着人工智能的不断发展，深度学习已经广泛应用在图像分类领域，对于图像分类的研究也趋于成熟，并且有许多相关的研究成果也被成功应用在日常生活的各个方面。对图像进行分类需要图像拥有类标签，然而在进行人工标注时会出现有些图像样本没有标签的极端情况，对于这些没有标签的类别，传统的分类方法无法对其进行分类。这一问题与人类对物体的识别过程类似，对于已经见过的物体，人类根据大脑的记忆可以很快地判断出该物体属于哪一类，而对于从未见过的物体就很难判断出其所属的类。在现实生活中，如果想要对未见过的事物准确分类，就需要不断地进行学习，所以用较少的有标签样本来训练一个可以识别无标签样本的模型很有必要。

为了解决上述问题，Larochelle 等[1]于2008年提出了零样本学习这一概念，主要研究训练类与测试类互斥时的分类问题，即当测试类的样本数据没有在训练过程中出现时，利用有限有标签的训练样本对测试样本进行分类。尽管人们很难对于从未见过的事物准确分类，但是人类却可以根据以前学到过的知识通过其特征来描述该事物的类别，如从未见过“鸭嘴兽”，但是可以将其描述为“有嘴”“有尾巴”“有毛”“有蹼”等信息，这些信息即为样本的属性，而通过将样本的属性引入到分类过程中，便可以有效地解决零样本图像分类问题。

典型的零样本学习模型[2-4]的训练过程只包含已知类的样本和所有类的语义特征。由于已知类与未知类只在语义空间中存在联系，零样本学习通常采用视觉特征嵌入语义空间或语义特征嵌入视觉空间的方法[5-7]。近年来，由于生成对抗网络（Generative Adversarial Network，GAN）[8]的飞速发展，出现了一些利用生成对抗网络来解决零样本图像分类问题的方法[9-10]。这些方法的主要思想是通过生成未知类的图像特征使得零样本图像分类问题转化为经典的图像分类问题。

生成对抗网络由生成网络和判别网络2部分组成，这2个网络以互相对抗的方式进行训练。

收稿日期：2019-07-08；**录用日期**：2019-08-03；**网络出版时间**：2019-08-12 15:18
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190812.1500.002.html
基金项目：国家自然科学基金（61463038）；内蒙古自治区自然科学基金重大项目（2019ZD14）
*通信作者. E-mail：cswhx@imu.edu.cn

引用格式：*魏宏喜，张越. 基于生成对抗网络的零样本图像分类[J]. 北京航空航天大学学报，2019，45(12)：2345-2350.*
WEI H X，ZHANG Y. Zero-shot image classification based on generative adversarial network[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2345-2350 (in Chinese).

生成网络试图输入噪声产生虚假图像来欺骗判别网络,而判别网络则试图区分真实图像与虚假图像。通过 2 个网络间的互相对抗学习,使产生的图像更加真实。尽管生成对抗网络已经取得了较好的结果,但其仍然难以训练。而 WGAN(Wasserstein GAN)[11]的提出解决了以往生成对抗网络训练困难的情况,其能使网络的训练过程更加稳定,同时还提供了可以指导训练进程的损失函数。另外,文献[12]中的条件生成对抗网络可以把类标签和其他信息引入生成过程中,以此来生成想要获得的结果。

本文采用生成对抗网络解决零样本图像分类问题,主要思路是:利用随机噪声和未知类的语义描述生成未知类图像的特征,再训练一个分类器对生成的图像特征进行分类。虽然通过上述方法间接解决了零样本图像分类问题,但是该方法却引入了一个新问题,即如何训练一个能够准确生成图像特征的生成对抗网络。在一般生成对抗网络生成图像特征的过程中,现有方法仅以语义信息来判别生成图像特征的真实性,而这无疑会产生一定的偏差。因此,本文改进了生成对抗网络中的判别网络,将语义信息转换为图像特征,从而使得判别更加准确,得到更好的分类结果。

本文主要贡献如下:①引入图像特征生成方法,通过一种间接的方式解决了零样本图像分类问题;②对生成对抗网络中的判别网络做出改进,提出了一种更为合理的方式来判断生成图像特征的真实性,从而提高了生成图像特征的质量,进一步提高分类准确率。

1 零样本图像分类

人类能够根据一个事物的语义描述来识别物体,受此启发,零样本学习旨在通过语义描述来学习一个具有良好泛化能力的模型,从而对未知类进行分类。如图 1 所示,在训练样本中有“北极熊”“马”“狗”“熊猫”,而测试样本中的“斑点狗”“斑马”没有出现在训练过程中。虽然模型从未见过“斑点狗”“斑马”,但却可以对这 2 类的特征进行描述,随后便可以根据这些描述来区分不同的类别。

图 1 零样本学习示例

Fig. 1 An example of zero-shot learning

对于图像分类问题,通常训练的模型在测试时只能预测训练集中已有的类别。而在零样本图像分类中,测试集的类别与训练集的类别不存在交集。因此,需要寻找一种新的方法来解决零样本图像分类问题。文献[13]中给出了最初的解决方法,即类间属性迁移。在零样本图像分类问题中,虽然物体的类别不同,但是物体间存在相同的属性,可以挑选出每一个类别对应的属性并进行学习。在测试时,对未知类的属性进行预测,再将预测出的属性组合对应到类别,从而实现对未知类的分类。

虽然上述方法能够在一定程度上解决零样本图像分类问题,但其最终取得的分类效果却不理想。文献[9]给出了一种利用生成对抗网络来生成未知类图像特征的方式解决零样本图像分类问题,经过实验验证,其最终的分类结果要优于一般的零样本图像分类方法,其主要思路是:利用未知类的语义信息来生成未知类的图像特征,这一部分是通过一个生成网络来完成的;同时利用一个判别网络来对生成的图像特征进行判别,使最终获得的图像特征更加真实。

2 特征生成网络

零样本图像分类的主要难点在于无法区分未知类的类别,本文利用生成对抗网络来生成未知类的图像特征,从而间接地解决零样本图像分类问题,将其转换为一般的分类问题。本文对 f-CLSWGAN[9]模型中的判别网络做出改进,以进一步提高分类准确率。

2.1 特征生成模型 f-CLSWGAN

f-CLSWGAN 模型是一种生成对抗网络的变体,生成对抗网络由生成网络 G 与判别网络 D 组成,通过使 2 个网络相互博弈来进行学习。该模型的主要目的是:通过语义信息来生成图像特征,而不是直接生成未知类图像,以此来获得更准确的分类结果。

首先,f-CLSWGAN 模型对生成网络与判别网络添加条件变量,使原始的生成对抗网络扩展为条件生成对抗网络;其次,由于生成对抗网络对于图像本身的生成效果仍不是很理想,因此,f-CLSWGAN模型将会生成未知类的图像特征而不是未知类的图像,从而使得条件生成对抗网络变体为 f-GAN;再次,模型将 WGAN 引入 f-GAN

中，使模型的训练过程更加稳定，并且可以确保生成特征的多样性，使其进一步变体为 f-WGAN；最后，模型将生成的图像特征送入一个分类器，得到一个分类结果，由此便形成了最终的 f-CLSWGAN 模型。具体的网络结构如图 2 所示。图中：z 为随机噪声，$c(y)$ 为语义信息，x 为已知类的图像特征，$\tilde{x}$ 为生成的未知类图像特征，θ 为参数，y 为 $\tilde{x}$ 类的标签，$P(y|\tilde{x};\theta)$ 为用类别标签 y 预测 $\tilde{x}$ 的概率，L_{WGAN} 和 L_{CLS} 分别为生成对抗网络与分类器的损失函数。

f-CLSWGAN 模型的损失函数为

$$\min_G \max_D \{L_{\mathrm{WGAN}}\} + \beta L_{\mathrm{CLS}} \tag{1}$$

式中：β 为分类器的一个超参数。

L_{WGAN} 和 L_{CLS} 的损失计算如下：

$$L_{\mathrm{WGAN}} = E(D(x,c(y))) - E(D(\tilde{x},c(y))) - \lambda E((\| \nabla_{\hat{x}}(D(\hat{x},c(y)))\|_2 - 1)^2) \tag{2}$$

式中：等号右边第 3 项为惩罚项，λ 为惩罚系数，$\hat{x} = \alpha x + (1-\alpha)\tilde{x}$ 且 $\alpha \sim U(0,1)$。

$$L_{\mathrm{CLS}} = -E_{\tilde{x}\sim P_{\tilde{x}}}(\ln P(y|\tilde{x};\theta)) \tag{3}$$

相较一般的零样本图像分类方法，该方法采用了一种新的思路，即利用语义信息生成未知类的图像特征，而不是去学习语义与图像之间的映射关系。使用该方法，可以将目标转变为如何更好地通过语义信息生成未知类的图像特征与如何对图像特征进行分类。同时，该模型采用了 WGAN 这种生成对抗网络作为基础，相较于其他类型的生成对抗网络，WGAN 具有以下优点：①训练过程更加稳定；②能够通过损失函数来指导训练过程；③生成的样本具有多样性。由于 f-CLSWGAN模型具有以上优势，从而使生成的图像特征更加真实，以得到更好的分类结果。

2.2　改进模型 FD-fGAN

由于在 f-CLSWGAN 模型中的判别条件为原始的语义信息，而语义与图像特征之间存在着很大的差异。因此本文改进了 f-CLSWGAN 模型中判别网络的判别条件，将语义信息嵌入到视觉特征空间中，得到的新模型命名为 FD-fGAN。具体网络模型结构与图 2 类似，图 3 显示了改进后 FD-fGAN 模型的网络结构。首先，将语义信息 $c(y)$ 通过 2 个全连接层（FC），得到一个图像特征 x^c；然后，利用该图像特征对生成网络生成的图像特征进行判别。对于生成网络，则不对其进行修改。

通过本文方法，生成对抗网络中的损失函数 L_{WGAN} 的计算由式（2）变为了式（4），即将判别网

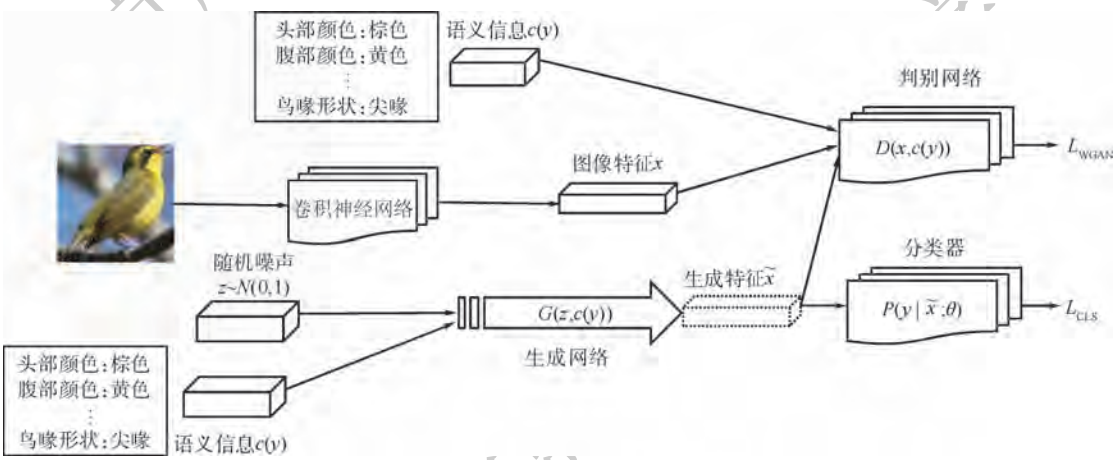


图 2　f-CLWSGAN 模型的网络结构

Fig. 2　Network structure of f-CLSWGAN model

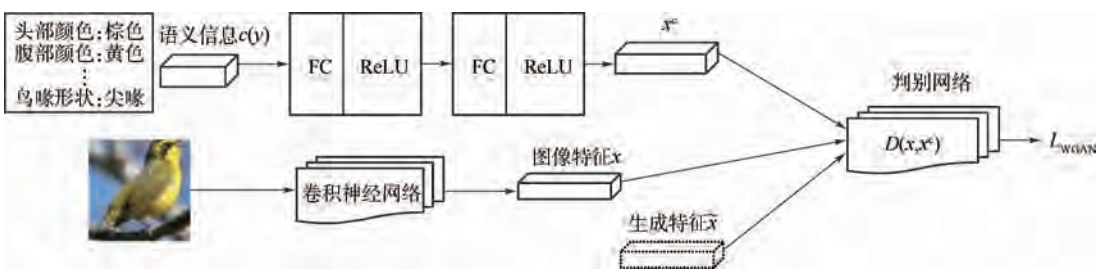


图 3　FD-fGAN 模型的网络结构

Fig. 3　Network structure of FD-fGAN model

络中的语义信息 $c(y)$ 替换为图像特征 x^c。

$$L_{\mathrm{WGAN}} = E(D(x,x^c)) - E(D(\tilde{x},x^c)) - \lambda E((\| \nabla_{\hat{x}}(D(\hat{x},x^c)) \|_2 - 1)^2) \quad (4)$$

本文方法将未知类的语义信息嵌入到视觉特征空间中,以获得一个图像特征。通过该图像特征可以减小判别特征与生成特征间的差异,使判别的过程更加准确,从而提高判别网络的判别性能,这也就使得生成网络生成的图像特征更加真实。最终能够得到一个更好的分类结果。

由于希望网络的训练过程足够稳定,且能够在训练过程中随时观察到最终的分类准确率,本文采用了 WGAN 作为网络结构的基础模型,构建了 FD-fGAN 模型。

3 实 验

3.1 实验数据集

本节实验采用了 3 种不同的经典数据集,分别为 AWA[13]、CUB[14] 及 SUN[15]。AWA 数据集是由 Lampert 等[13]在 2009 年构建的,由 50 个动物类别的 37 322 幅图像组成,每个图像都具有预先提取的特征表示,且每个类别都有 85 个不同的属性。CUB 数据集由加州理工学院在 2010 年构建,由 200 个不同鸟类的 11 788 幅图像组成,每个类别都有 312 个不同的属性。SUN 数据集是一个场景属性数据集,其是在细粒度的 SUN 分类数据集之上构建得到的,由 717 个场景的 14 340 幅图像组成,每个类别都有 102 个不同的属性。3 个数据集的具体划分情况如表 1 所示。将表 1 中的已知类作为训练集,未知类作为测试集,已知类与未知类没有交集。

表 1 三个数据集的划分情况

Table 1 Division of three datasets

数据集	类别数	已知类别数	未知类别数
AWA	50	40	10
CUB	200	150	50
SUN	717	645	72

3.2 实验结果与分析

参考 Xian 等的工作[9],实验中的图像特征是利用 ResNet-101 提取的,并且在 ImageNet 数据集上进行了预训练。而对于语义描述,使用数据集中包含的默认属性。实验参数配置也与 f-CLSWGAN模型相同。实验在 AWA、CUB 和 SUN 三个不同数据集上分别进行了 70 轮训练。图 4 为不同数据集上的分类准确率趋势。可以看出,在 AWA 数据集上,只用了 30 轮左右分类准确率就达到了最高水平,随后开始缓慢下降;在 CUB 数据集上,训练 20 轮以后分类准确率便开始在一个范围内波动,并在 43 轮时达到最高;在 SUN 数据集上,训练 25 轮以后分类准确率基本保持不变。分类准确率能很快达到稳定的主要原因是采用了 WGAN 作为 FD-fGAN 模型的基础。

表 2 给出了不同方法在 3 个数据集上的对比结果。可以看出,本文所设计的 FD-fGAN 模型取得的结果要优于其他模型。其中,基于生成对抗网络的方法都要优于其他零样本图像分类方法,证明了将生成对抗网络引入零样本图像分类问题的合理性。

原始 f-CLSWGAN 模型直接利用语义信息进行判别,而在语义信息与图像特征之间存在着巨大的差异。本文 FD-fGAN 模型将这一语义信息通过 2 个全连接层嵌入到视觉特征空间,以减小语义信息与图像特征之间的巨大差异,从而使得分类准确率有所提升。本文所设计的 FD-fGAN 模型在 AWA 数据集上取得了 68.6% 的分类准确率,比 f-CLSWGAN 模型提高了 0.4%;在 CUB 和 SUN 数据集上分别取得了 57.7% 与 61.3% 的分类准确率,比 f-CLSWGAN 模型提高了 0.4% 和 0.5%。经过实验证明,本文通过改进判别网络设

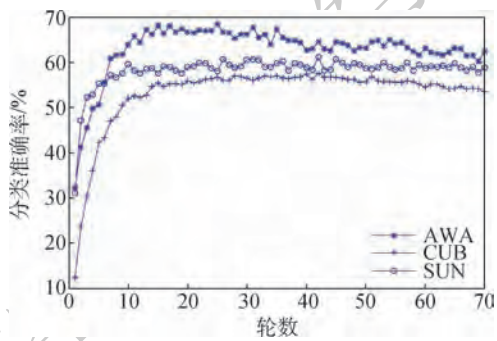


图 4 不同数据集上的分类准确率

Fig. 4 Classification accuracy on different datasets

表 2 不同方法在 3 个数据集上的分类准确率比较

Table 2 Comparison of classification accuracy of different methods on three datasets

方法	分类准确率/%		
	AWA	CUB	SUN
DAP[6]	44.1	40.4	39.9
CONSE[16]	45.6	34.3	38.8
SSE[17]	60.1	43.9	51.5
DeViSE[18]	54.2	52.0	56.5
SJE[19]	65.6	53.9	53.7
ESZSL[7]	58.2	53.9	54.5
ALE[20]	59.9	54.9	58.1
SYNC[2]	54.0	55.6	56.3
f-CLSWGAN[9]	68.2	57.3	60.8
FD-fGAN	68.6	57.7	61.3

计出的 FD-fGAN 模型能够取得更高的分类准确率，表明利用图像特征作为判别条件是可行的。

4　结　论

1）本文提出了一种利用图像特征进行判别的生成对抗网络 FD-fGAN，以此来解决零样本图像分类问题。经过实验验证，本文方法在 AWA、CUB、SUN 三个不同的数据集上分别提高了 0.4%、0.4% 与 0.5%。

2）本文所提出的利用图像特征作为判别条件来进行判别的方法，减少了语义信息与图像特征之间存在的巨大差异，加强了判别网络的判别能力。

本文的工作主要是针对生成对抗网络进行的，未来将研究如何使得由语义信息生成的图像特征更加准确，并对生成对抗网络作进一步改进。

参考文献（References）

[1] LAROCHELLE H, ERHAN D, BENGIO Y. Zero-data learning of new tasks[C]// Proceedings of the Association for the Advancement of Artificial Intelligence. Palo Alto: AAAI Press, 2008: 646-651.

[2] CHANGPINYO S, CHAO W, GONG B, et al. Synthesized classifiers for zero-shot learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 5327-5336.

[3] KODIROV E, XIANG T, GONG S. Semantic auto-encoder for zero-shot learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 4447-4456.

[4] XIAN Y, SCHIELE B, AKATA Z. Zero-shot learning-the good, the bad and the ugly[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 3077-3086.

[5] DING Z, SHAO M, FU Y. Low-rank embedded ensemble semantic dictionary for zero-shot learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 6005-6013.

[6] LAMPERT C H, NICKISCH H, HARMELING S. Attribute-based classification for zero-shot visual object categorization [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(3): 453-465.

[7] ROMERA-PAREDES B, TORR P. An embarrassingly simple approach to zero-shot learning[C]// Proceedings of the International Conference on Machine Learning. Madison: Omnipress, 2015: 2152-2161.

[8] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Proceedings of the Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[9] XIAN Y, LORENZ T, SCHIELE B, et al. Feature generating networks for zero-shot learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 5542-5551.

[10] ZHU Y, ELHOSEINY M, LIU B, et al. A generative adversarial approach for zero-shot learning from noisy texts[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 1004-1013.

[11] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein generative adversarial networks[C]// Proceedings of the International Conference on Machine Learning. Madison: Omnipress, 2017: 214-223.

[12] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06)[2019-07-01]. https://arxiv.org/abs/1411.1784.

[13] LAMPERT C H, NICKISCH H, HARMELING S. Learning to detect unseen object classes by between-class attribute transfer [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 951-958.

[14] WAH C, BRANSON S, WELINDER P, et al. The Caltech-UCSD Birds-200-2011 dataset: CNS-TR-2011-001 [R]. Pasadena: California Institute of Technology, 2011.

[15] PATTERSON G, HAYS J. SUN attribute database: Discovering, annotating, and recognizing scene attributes[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2012: 2751-2758.

[16] NOROUZI M, MIKOLOV T, BENGIO S. Zero-shot learning by convex combination of semantic embeddings[C]// Proceedings of the International Conference on Learning Representations. Brookline: Microtome Publishing, 2014: 10-19.

[17] ZHANG Z, SALIGRAMA V. Zero-shot learning via semantic similarity embedding[C]// Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 4166-4174.

[18] FROME A, CORRADO G S, SHLENS J, et al. DeViSE: A deep visual semantic embedding model[C]// Proceedings of the Neural Information Processing Systems. Cambridge: MIT Press, 2013: 2121-2129.

[19] AKATA Z, REED S, WALTER D, et al. Evaluation of output embeddings for fine-grained image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 2927-2936.

[20] AKATA Z, PERRONNIN F, HARCHAOUI Z, et al. Label-embedding for image classification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(7): 1425-1438.

作者简介：

魏宏喜　男，博士，教授，博士生导师。主要研究方向：机器学习与数据挖掘。

张越　男，硕士研究生。主要研究方向：图像分类。

Zero-shot image classification based on generative adversarial network

WEI Hongxi*, ZHANG Yue

(College of Computer Science, Inner Mongolia University, Hohhot 010021, China)

Abstract: The problem of zero-shot image classification has become a research focus in the field of image classification. In this paper, a method based on generative adversarial network (GAN) is used to solve the problem of zero-shot image classification. By generating image features of unseen classes, the zero-shot classification task is transformed into a conventional image classification task. At the same time, this paper makes modifications to the discriminant network in the generative adversarial network to make the discriminating process more accurate. The experimental results show that the performance of the proposed method has been increased by 0.4%, 0.4% and 0.5% on the datasets of AWA, CUB and SUN, respectively. Therefore, the proposed method can generate the better features by improving the generative adversarial networks, which results in solving the problem of zero-shot image calssification effectively.

Keywords: deep learning; image classification; zero-shot learning; generative adversarial network (GAN); image feature generation

Received: 2019-07-08; **Accepted**: 2019-08-03; **Published online**: 2019-08-12 15:18

URL: kns.cnki.net/kcms/detail/11.2625.V.20190812.1500.002.html

Foundation item: National Natural Science Foundation of China (61463038); Natural Science Foundation of Inner Mongolia Autonomous Region (2019ZD14)

* **Corresponding author**. E-mail: cswhx@imu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0372

基于自适应注入模型的遥感图像融合方法

杨勇[1,*]，卢航远[1]，黄淑英[2]，涂伟[1]，李露奕[1]
（1. 江西财经大学 信息管理学院，南昌 330032； 2. 江西财经大学 软件与物联网工程学院，南昌 330032）

摘 要：遥感图像融合的目的是融合高光谱分辨率、低空间分辨率的多光谱(MS)图像和高空间分辨率、低光谱分辨率的全色(PAN)图像，得到高光谱分辨率与高空间分辨率的融合图像。遥感图像的注入模型中如何确定注入细节及注入系数是该技术研究的关键。针对注入细节优化，先通过模拟MS传感器的特性来定义一种多尺度高斯滤波器，再用该滤波器卷积PAN图像以提取细节，得到与MS图像高度相关的细节。针对注入系数优化，综合考虑光谱信息与细节信息提出一种自适应的注入量系数。为更好地保留边缘信息，提出一种新的边缘保持权重矩阵，实现光谱信息与空间的双保真。将优化后的注入系数与注入细节相乘注入到上采样后的MS图像中，得到融合结果。对所提方法进行性能分析，并在各卫星数据集上进行大量测试，与一些先进的遥感图像融合方法进行对比，实验结果表明，所提方法在主观与综合客观指标上都能达到最优。

关 键 词：遥感图像融合；高斯滤波；注入系数；边缘保持；质量评价
中图分类号：TP391.41
文献标识码：A **文章编号**：1001-5965(2019)12-2351-13

遥感的目的是通过获取卫星的光谱测量，提取有关地球表面结构和内容的信息[1]。由于传感器技术的限制，同时在空间域和光谱域上获取高分辨率难以实现，因此通常利用图像处理技术融合现有的2类遥感图像，即高空间分辨率、低光谱分辨率的全色(Panchromatic, PAN)图像和低空间分辨率、高光谱分辨率的多光谱(Multispectral, MS)图像。对于在同一区域同时获得的PAN图像和MS图像，通过适当的算法，将这些数据结合起来，生成具有高空间分辨率和高光谱分辨率的图像，这一过程也称为PAN锐化[2]。遥感图像融合广泛应用于地理、军事、农业、生态环境等领域。

现有的传统PAN锐化方法主要分为2类：一类是成分替代法(Component Substitute, CS)，即将MS图像的成分替换成PAN图像成分，如亮度-色度-饱和度变换法(Intensity-Hue-Saturation, IHS)[3]、主成分分析法(Principle Component Analysis, PCA)[4]、基于施密特正交化方法(Gram-Schmidt, GS)[5]等。基于CS方法的融合结果具有较高的空间分辨率，但其光谱容易失真。另一类是多尺度分析法(Multi-Resolution Analysis, MRA)，即对PAN图像进行多尺度分解后将提取的细节加入到MS图像中。基于MRA的方法和基于CS的方法的主要区别在于如何提取空间细节[6]。对于MRA方法，通过PAN图像与其低通滤波之后的图像之间的差异来获得空间结构信息。传统的MRA方法主要包括基于金字塔分解

收稿日期：2019-07-09；**录用日期**：2019-08-18；**网络出版时间**：2019-08-26 16:13
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190826.1555.004.html
基金项目：国家自然科学基金(61662026,61862030)；江西省自然科学基金(20182BCB22006,20181BAB202010,20192ACB20002,20192ACBL21008)；江西省教育厅科学技术研究项目(GJJ170312,GJJ170318)；江西省研究生创新专项资金(YC2019-B094,YC2018-B065)
*通信作者. E-mail: greatyangy@126.com

引用格式：*杨勇，卢航远，黄淑英，等. 基于自适应注入模型的遥感图像融合方法[J]. 北京航空航天大学学报，2019，45(12)：2351-2363. YANG Y，LU H Y，HUANG S Y，et al. Remote sensing image fusion method based on adaptive injection model[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2351-2363 (in Chinese).*

或者小波变换的方法，如离散小波变换[7]、àtrous小波变换[8]、非抽样的轮廓波变换[9]、非抽样的剪切波变换[10]等。MRA方法可以较好地保存光谱信息，但可能会损失部分空间信息，因此文献[11]提出了一种基于引导滤波多尺度分解的方法，效果有所改进。最新的理论研究提出了稀疏表示(Sparse Representation，SR)的方法[12-13]，以及非线性的途径如卷积神经网络[14]等，取得了较好的效果。但是，这些算法需要大量的计算，运行时间较长，且需要大量的样本。由于遥感图像样本数量的限制，以及算法效率的要求，本文综合考虑了上述几种方法，对基于MRA方法的注入模型进行优化，以实现融合时对光谱信息与细节信息的双保真。

基于MRA方法中关键的一步在于如何提取PAN图像细节。一种较为先进的方法是利用线性时不变的滤波器去匹配MS传感器的点传播函数(Point Spread Function，PSF)。Aiazzi等[15]提出使用高斯滤波器去匹配MS传感器的调制传输函数，但需要使用MS传感器的出厂信息，而该信息不易获得。另一种较为可靠的方法是从现有图像中估计滤波器以模拟MS传感器的PSF。在此框架下，一幅低空间分辨率的MS(Low Resolution MS，LRMS)图像，可以通过滤波器对一幅高空间分辨率的MS(High Resolution MS，HRMS)图像进行空间滤波得到。该滤波器的冲击响应函数可模拟MS传感器的PSF，且该滤波器通常有与高斯模型相似的形状[2,15]。使用该滤波器卷积PAN图像，所得的高频分量即为HRMS图像中缺失的细节成分，且与MS图像高线性相关。此外，综合考虑光谱的保真和细节的注入，本文提出了一种基于相关性分析的高斯滤波估计和自适应细节注入的遥感图像融合方法。首先，通过引导滤波多尺度分解的注入模型得到初始融合图像；其次，用高斯滤波模拟MS传感器的PSF，并用估计的高斯滤波器卷积PAN图像得到优化的注入细节；然后，递归计算自适应的注入系数，使光谱保真与细节注入达到联合最佳；最后，通过一种新的边缘保持策略，得到融合图像。

本文主要创新点如下：①提出了一种基于高斯滤波估计的细节提取方法，使注入的细节得到优化；②提出了一种综合考虑光谱信息与细节信息的自适应注入系数模型，实现融合图像对光谱信息与细节信息的双保真；③提出了一种减少光谱信息损失的边缘保持模型，在增强边缘信息的同时保留光谱信息。

1 相关工作

1.1 注入模型的一般框架

基于MRA方法的注入模型是指通过滤波提取PAN图像细节，再利用注入系数将其注入到MS图像中的一种融合方法。该方法可根据实际需要，组合不同的融合技术，以发挥不同融合技术的优点。基于传统的MRA方法注入模型框架如图1所示，该注入模型的一般表示形式如下[16]：

$$\widehat{\mathbf{MS}}_k = \widehat{\mathbf{MS}}_k + \boldsymbol{G}_k .* (\boldsymbol{P}_I - \boldsymbol{P}_{\mathrm{L}}) \tag{1}$$

式中：$k = \{1, 2, \cdots, N\}$表示N个通道中的第k个子通道；$\widehat{\mathbf{MS}}_k$为估计的第k通道的HRMS图像；

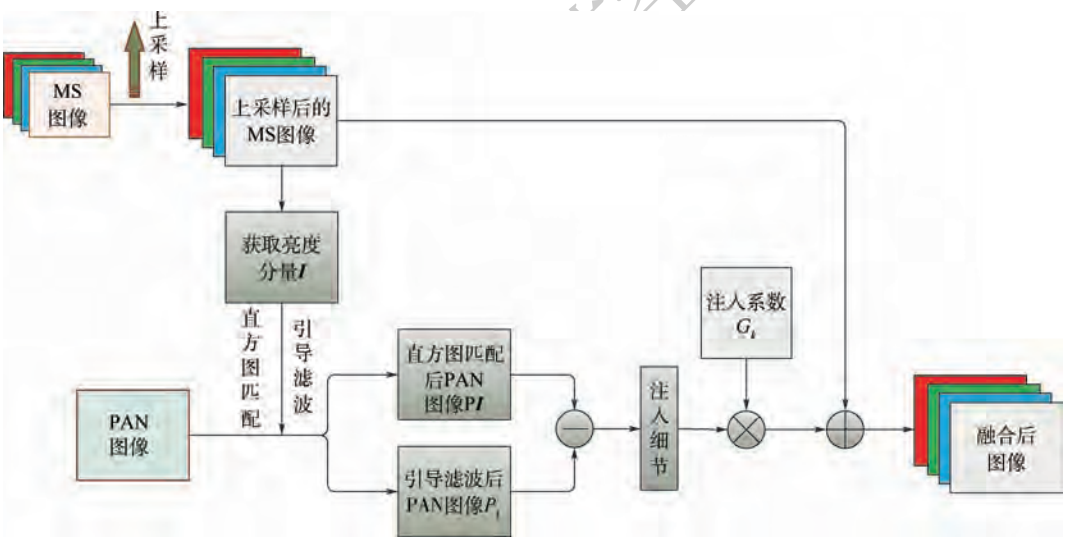


图1 注入模型一般框架

Fig. 1 Basic framework of injection model

$\widetilde{\mathbf{MS}}_k$ 为经过 bicubic 上采样之后的第 k 通道 MS 图像;$\boldsymbol{G}_k$ 为第 k 通道注入系数;".*"表示矩阵内积;$\boldsymbol{P}_I$ 为 PAN 图像与 MS 图像的亮度分量 $\boldsymbol{I}$ 进行直方图均衡化之后的图像;$\boldsymbol{P}_{\mathrm{L}}$ 为低分辨率版本的 PAN 图像。

对于注入系数 $\boldsymbol{G}_k$,文献[16]通过迭代算法,提出基于回归模型的通道间权重系数;文献[17]提出基于像素通道间比例的权重系数;文献[18]通过高通调制(High Pass Modulation,HPM)方法,用 MS 图像和 PAN 图像的乘法组合作为注入系数。其中,基于通道间比例的注入系数能更好地保持光谱信息,融合效果较好,通道间注入系数定义如下:

$$\boldsymbol{G}_k = \frac{\widetilde{\mathbf{MS}}_k}{\frac{1}{N}\sum_{k=1}^{N}\widetilde{\mathbf{MS}}_k} \tag{2}$$

1.2　线性加权法提取 $\boldsymbol{I}$ 分量

由于注入细节的同时不希望改变色度与饱和度,因此图像融合过程中往往仅对图像的亮度分量进行处理。MS 图像的亮度分量 $\boldsymbol{I}$ 可由如下公式计算:

$$\boldsymbol{I} = \sum_{k=1}^{N}\alpha_k\widetilde{\mathbf{MS}}_k \tag{3}$$

式中:α_k 为通道的组合系数。

文献[19-20]提出 $\boldsymbol{I}$ 分量可自适应地由各通道线性组合表示,即通过求解以下优化问题得到组合系数:

$$\begin{cases}\min\limits_{\alpha_1,\cdots,\alpha_N}\left\|\boldsymbol{P}-\sum\limits_{k=1}^{N}\alpha_k\widetilde{\mathbf{MS}}_k\right\|^2\\ \text{s.t. }\alpha_1\geqslant 0,\cdots,\alpha_N\geqslant 0\end{cases} \tag{4}$$

式中:$\boldsymbol{P}$ 为 PAN 图像。

在得到 MS 图像的 $\boldsymbol{I}$ 分量之后,将 $\boldsymbol{P}$ 与式(3)的 $\boldsymbol{I}$ 进行直方图匹配,得到的 $\boldsymbol{P}_I$ 可由如下公式计算:

$$\boldsymbol{P}_I \triangleq (\boldsymbol{P}-\mathrm{mean}(\boldsymbol{P}))*\mathrm{std}(\boldsymbol{I})/\mathrm{std}(\boldsymbol{P})+\mathrm{mean}(\boldsymbol{I}) \tag{5}$$

式中:mean($\boldsymbol{P}$)为 $\boldsymbol{P}$ 的均值;mean($\boldsymbol{I}$)为 $\boldsymbol{I}$ 分量的均值;std($\boldsymbol{I}$)为 $\boldsymbol{I}$ 分量的标准差;std($\boldsymbol{P}$)为 $\boldsymbol{P}$ 的标准差。

1.3　引导滤波

引导滤波最早由文献[21]提出,其可以保存输入图像的主要信息,同时获得引导图像的变化趋势。将 $\boldsymbol{P}_I$ 作为输入图像,$\boldsymbol{I}$ 作为引导图像,$\boldsymbol{P}_I$ 通过 $\boldsymbol{I}$ 引导滤波产生 $\boldsymbol{P}_{\mathrm{L}}$,获得了与 MS 图像线性相关的细节信息。引导滤波可由如下简化公式而得:

$$\boldsymbol{P}_{\mathrm{L}} = \mathrm{GF}(\boldsymbol{P}_I,\boldsymbol{I}) \tag{6}$$

式中:GF(·)表示引导滤波函数。

多尺度引导滤波[22]可表示为

$$\boldsymbol{P}_{\mathrm{L}}^{j} = \mathrm{GF}(\boldsymbol{P}_{\mathrm{L}}^{j-1},\boldsymbol{I}) \tag{7}$$

式中:j 为引导滤波层数;$\boldsymbol{P}_{\mathrm{L}}^{j}$ 为 $\boldsymbol{P}_I$ 经第 j 层引导滤波得到的图像。

2　图像融合方法

目前,遥感图像融合注入模型主要存在 2 类问题:因 PAN 图像与 MS 图像相关性不高导致的光谱失真;细节注入过多或者注入不足。针对上述问题,本文提出了一种基于高斯滤波估计的细节提取及自适应的注入系数优化方法。本文融合方法框架如图 2 所示。

首先,根据式(1),将上采样后的 MS 图像与 PAN 图像通过直方图匹配、引导滤波等传统注入模型得到初始融合图像 $\boldsymbol{F}^{(1)}$;其次,进行细节优化,即用多尺度高斯滤波器滤波该初始融合图像以模拟 MS 传感器的特性,将得到的高斯滤波估计滤波 PAN 图像,得到优化后的细节;然后,优化注入效益,即综合考虑光谱信息与细节信息计算自适应的注入量系数,具体流程如图 2 上虚线框所示,详细算法见 2.2 节;最后,通过一种自适应的边缘保持权重矩阵保持边缘信息,具体流程如图 2 下虚线框所示,详细算法见 2.3 节。将优化后的细节与注入系数及边缘权重系数相乘以后,注入到上采样后的 MS 图像中,得到最终的融合图像。

2.1　高斯滤波估计及细节提取

文献[15]提出使用高斯滤波器匹配 MS 传感器的调制传输函数,即 HRMS 图像经过 MS 传感器之后成为 LRMS 图像,使用高斯滤波模拟该过程,所得的高斯滤波器即为符合 MS 传感器 PSF 的滤波估计。使用估计的滤波器卷积 PAN 图像,提取的细节成分即为低分辨率 MS 图像所缺失的细节。文献[15]通过实验证明了该滤波估计有较强的鲁棒性,即使在初始融合图像中加入白噪声,也不影响细节的提取。另外,文献[17]提出了二步骤多尺度分解的方法来精细化 PAN 锐化。受此启发,本文提出使用多尺度高斯滤波模拟 MS 传感器的 PSF,即对初始融合图像 $\boldsymbol{F}^{(1)}$ 提取亮度分量,记为 $\boldsymbol{I}_1$,并迭代地进行高斯滤波,具体可表示为

$$\boldsymbol{I}_1^i = H_{\mathrm{G}}(\boldsymbol{I}_1^{i-1}) \qquad i=1,2,\cdots,n \tag{8}$$

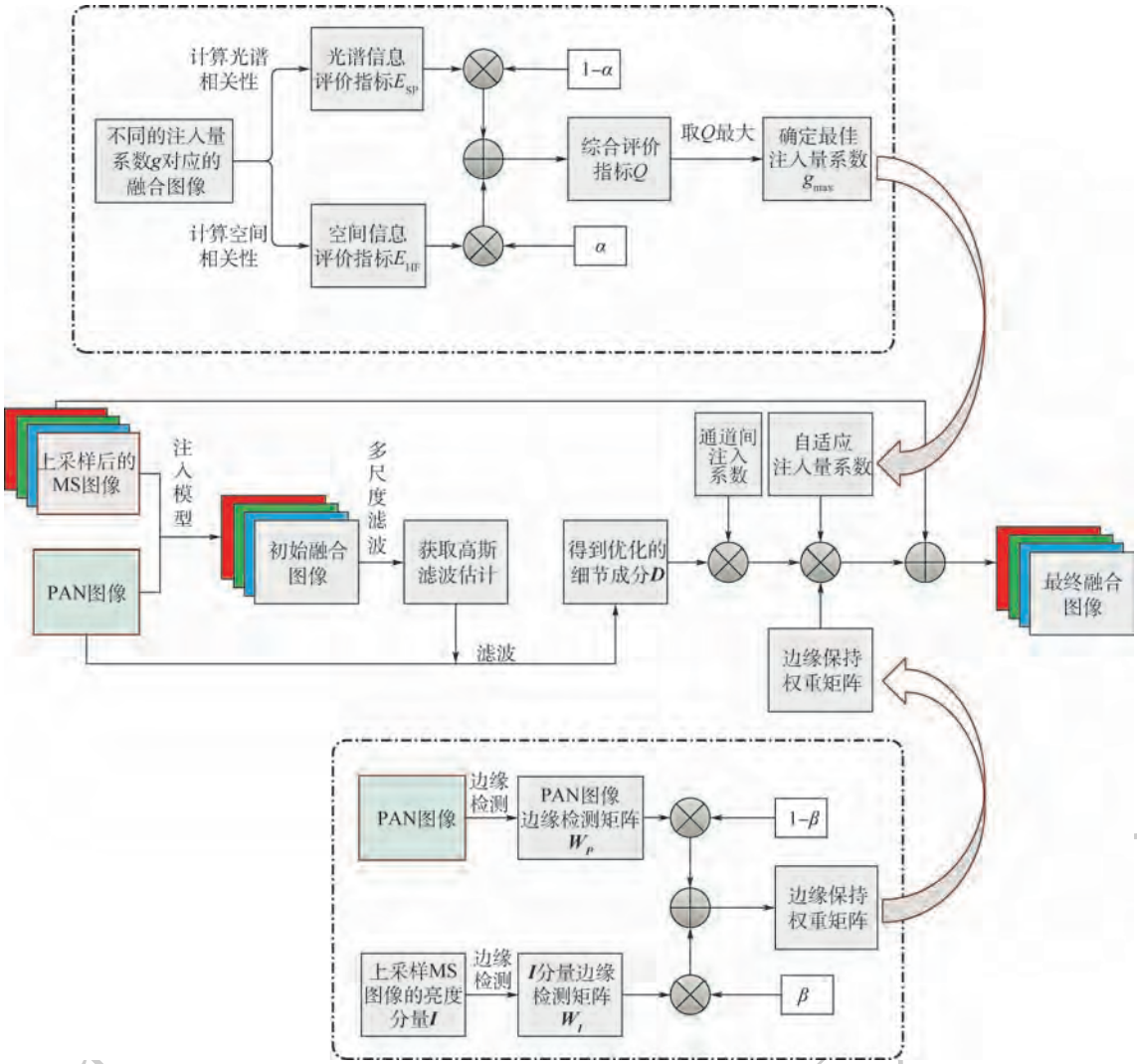


图 2 本文融合方法框架

Fig. 2 Framework of proposed fusion method

式中：$\boldsymbol{I}_1^i$ 为第 i 次滤波后的图像；$H_G(\cdot)$ 表示高斯滤波。

为了获取最佳高斯滤波估计，计算式(8)所得的结果 $\boldsymbol{I}_1^i$ 与上采样后的 MS 图像的亮度分量 $\boldsymbol{I}$ 的相关系数，相关系数最高时，估计的高斯滤波最准确。相关系数计算如下：

$$\mathrm{corr}(\boldsymbol{I}_1^i,\boldsymbol{I}) = \frac{\mathrm{cov}(\boldsymbol{I}_1^i,\boldsymbol{I})}{\mathrm{std}(\boldsymbol{I}_1^i)\cdot\mathrm{std}(\boldsymbol{I})} \tag{9}$$

式中：$\mathrm{cov}(\cdot)$ 表示协方差函数。

随着滤波层次的增加，滤波后的图像与上采样的 MS 图像越来越接近，达到极大值之后，随着滤波层次继续增加，滤波后图像与上采样的 MS 图像偏离越来越大。通过迭代计算式(8)和式(9)的值，取得式(9)最大时的迭代次数 m，即为所估计的高斯滤波器 H_m，表示如下：

$$\begin{cases} H_m(\cdot) = \underbrace{H_G(H_G(\cdots H_G(\cdot)))}_{m} \\ m = \arg\max_i \mathrm{corr}(\boldsymbol{I}_1^i,\boldsymbol{I}) \end{cases} \tag{10}$$

H_m 即表示一幅锐化图像经 MS 传感器模糊化的 PSF 最佳匹配，将滤波器 H_m 作用于 PAN 图像，所提取的细节成分也应符合 MS 传感器 PSF，这样就减少了光谱失真。细节提取 $\boldsymbol{D}$ 定义如下：

$$\boldsymbol{D} = \boldsymbol{P}_I - H_m(\boldsymbol{P}_I) \tag{11}$$

结合式(1)、式(7)和式(11)，细节优化后，第 k 通道融合图像 $\boldsymbol{F}_k^{(2)}$ 可更新为

$$\boldsymbol{F}_k^{(2)} = \widehat{\mathbf{MS}}_k + \frac{\widehat{\mathbf{MS}}_k}{\frac{1}{N}\sum_{k=1}^{N}\widehat{\mathbf{MS}}_k} .* \boldsymbol{D} \tag{12}$$

2.2 自适应注入量系数

取得优化的注入细节之后，对注入量进行

优化。不同的图像有不同的结构特征和光谱信息,从而对细节注入量有不同的要求,直接等比例注入细节容易导致注入过多从而引起光谱失真,或者注入不足导致图像模糊,因此需要通过一个注入量系数 g 来平衡光谱信息与细节信息。加入该系数后,结合式(12),新的融合图像 $\boldsymbol{F}_k^{(3)}$ 可以定义为

$$\boldsymbol{F}_k^{(3)} = \widehat{\mathbf{MS}}_k + g\frac{\widehat{\mathbf{MS}}_k}{\frac{1}{N}\sum_{k=1}^{N}\widehat{\mathbf{MS}}_k} .* \boldsymbol{D} \tag{13}$$

需要对不同结构特征的遥感图像确定不同的注入量系数 g。文献[23]认为注入模型中光谱的保真与空间分辨率要求需要保持平衡,可以将两者的评价指标进行加权求和作为融合图像指标,权重表示重视程度。设定细节注入量系数 g 的初值g和步长 r,对于 $g_0 \leqslant g \leqslant 1$,得到相应的融合后图像并计算加权评价指标,取得使该评价指标最高的$g_{\max}$作为最终的注入量系数。对于评价指标,本文提出了一种基于相关性分析的光谱信息与空间信息综合评价体系,将光谱信息与空间信息的综合评价指标 Q 定义为

$$Q = (1-\alpha)E_{\mathrm{SP}} + \alpha E_{\mathrm{HF}} \tag{14}$$

式中:E_{SP}和 E_{HF}分别为光谱信息评价指标和空间信息评价指标;α 为两者的权重。

对于一幅由式(13)融合后的图像 $\boldsymbol{F}_k^{(3)}$,其亮度分量记为 $\boldsymbol{I}_3$,则E_{SP}用融合后的图像与上采样的 MS 图像各通道 k 之间相关系数的平均值来表示,E_{HF}则用 $\boldsymbol{I}_3$与 PAN 图像的线性相关性表示,定义为

$$E_{\mathrm{SP}} = \frac{1}{N}\sum_{k=1}^{N}\mathrm{corr}(\boldsymbol{F}_k^{(3)}, \widehat{\mathbf{MS}}_k) \tag{15}$$

$$E_{\mathrm{HF}} = \mathrm{corr}(\boldsymbol{I}_3, \boldsymbol{P}_I) \tag{16}$$

文献[6]表明,在注入成分与注入量系数 g 的优化确定上,遵循的前提是 MS 图像的亮度成分与 PAN 图像相关性越大,融合效果越好。即若两者相关性高,注入细节的同时能较好地保存光谱信息,此时可以适当增加空间信息的权重,减少光谱信息的权重;反之,若相关性低,则注入的细节容易引起光谱失真,应当增加光谱信息权重而减少空间信息权重。设初值g_0产生的融合图像对应的亮度成分记为$\boldsymbol{I}_0$,则式(14)中的权重 α 可由下式确定:

$$\alpha = (\mathrm{corr}(\boldsymbol{I}_0, \boldsymbol{P}_I))^2 \tag{17}$$

平方项的作用为:①避免$\boldsymbol{I}_0$ 与$\boldsymbol{P}_I$高相关时光谱权重过小;②放大不同图像空间信息之间的区别。

对于不同的注入量系数 g,根据式(13)迭代计算 $\boldsymbol{F}_k^{(3)}$,并根据式(14)计算综合评价指标 Q,取得使指标 Q 最高的$g_{\max}$作为最终融合图像的注入量系数,并得到最后的融合结果。$g_{\max}$可表示为

$$g_{\max} = \arg\max_{g} Q(\boldsymbol{F}_k^{(3)}(g)) \tag{18}$$

结合式(13)和式(18),使用优化后的注入量系数,新的融合图像可更新为

$$\boldsymbol{F}_k^{(3)} = \widehat{\mathbf{MS}}_k + g_{\max}\frac{\widehat{\mathbf{MS}}_k}{\frac{1}{N}\sum_{k=1}^{N}\widehat{\mathbf{MS}}_k} .* \boldsymbol{D} \tag{19}$$

2.3　边缘信息保护

为保持边缘纹理信息,Leung 等[20]提出利用边缘信息作为各通道融合权重,Yang 等[11,24]提出利用改进的边缘信息作为注入各像素的权重。其中,改进的边缘信息计算 PAN 图像的边缘检测矩阵$\boldsymbol{W}_P$,以及上采样后的 MS 图像各通道的边缘检测矩阵$\boldsymbol{W}_{\widetilde{\mathbf{MS}}_k}$,并将两者进行加权和作为融合图像的边缘信息权重。该方法在边缘信息保护上起到了较好的效果。第 k 通道的边缘保持权重矩阵 $\boldsymbol{E}_k$ 定义如下[20]:

$$\boldsymbol{E}_k = (1-\beta_k)\boldsymbol{W}_P + \beta_k \boldsymbol{W}_{\widetilde{\mathbf{MS}}_k} \tag{20}$$

式中:β_k 为各通道的加权系数,可由式(21)计算得到[11];边缘检测矩阵[19]可由式(22)得到。

$$\begin{cases}\min\limits_{\beta_1,\cdots,\beta_N}\left\|\boldsymbol{W}_P - \sum\limits_{k=1}^{N}\beta_k \boldsymbol{W}_{\widetilde{\mathbf{MS}}_k}\right\|^2 \\ \text{s.t. } \beta_1 \geqslant 0, \cdots, \beta_N \geqslant 0\end{cases} \tag{21}$$

$$\boldsymbol{W}_A = \exp\left(-\frac{\lambda}{|\nabla \boldsymbol{A}|^4 + \varepsilon}\right) \tag{22}$$

式中:$\boldsymbol{A}$ 为某图像;$\nabla \boldsymbol{A}$ 为图像 $\boldsymbol{A}$ 的梯度;λ 和 ε 为调制参数。

式(21)获取各通道的边缘信息权重,在细节上有更好的表现,但作为注入系数时改变了各通道间的比例,光谱信息略有损失。为了尽量保持原有光谱信息,同时实现边缘信息的保持,本文提出一种新的边缘保持权重矩阵。针对式(21),将求取各通道的边缘检测矩阵$\boldsymbol{W}_{\widetilde{\mathbf{MS}}_k}$替换为求取 MS 图像亮度成分 $\boldsymbol{I}$ 的边缘检测矩阵 $\boldsymbol{W}_I$,并计算 PAN 图像的边缘检测矩阵 $\boldsymbol{W}_P$ 与 $\boldsymbol{W}_I$ 加权和,以综合考虑 PAN 图像和 MS 图像的边缘信息。边缘保持权重矩阵 $\boldsymbol{E}$ 可表示为

$$\boldsymbol{E} = (1-\beta_{\mathrm{new}})\boldsymbol{W}_P + \beta_{\mathrm{new}}\boldsymbol{W}_I \tag{23}$$

式中:$\boldsymbol{W}_P$ 和 $\boldsymbol{W}_I$ 计算方法同式(22);自适应的权重因子 β_{new} 由式(24)计算,并由式(25)归一化得到。

$$\begin{cases}\min\limits_{\beta}\left\|\boldsymbol{W}_P - \beta \boldsymbol{W}_I\right\|^2 \\ \text{s.t. } \beta \geqslant 0\end{cases} \tag{24}$$

$$\beta_{\mathrm{new}} = \beta/(1+\beta) \tag{25}$$

根据式(23)得到的边缘保持权重矩阵 $\boldsymbol{E}$,再结合式(19),可得到最终融合图像 $\boldsymbol{F}_k^{(4)}$,表示如下:

$$\boldsymbol{F}_k^{(4)} = \widehat{\mathbf{MS}}_k + g_{\max} \frac{\widehat{\mathbf{MS}}_k}{\frac{1}{N}\sum_{k=1}^{N} \widehat{\mathbf{MS}}_k} .* \boldsymbol{E} .* \boldsymbol{D} \quad (26)$$

与式(20)的边缘保持权重矩阵相比,本文得到的边缘保持权重矩阵 $\boldsymbol{E}$ 仅对细节成分 $\boldsymbol{D}$ 产生权重变化,以尽可能增强边缘信息,并不改变通道间的比例,从而减少光谱损失。

3 实验结果分析

3.1 实验设置

为评估本文方法的有效性,使用来自不同遥感卫星的四大数据集,包括 QuickBird、IKONOS、WorldView-2 及 pleiades,这些遥感图像在空间分辨率、光谱波长等方面都有不同的特征,且都包括红、绿、蓝、近红外 4 个通道。

参数设置中,设置注入量系数初值 $g_0 = 0.1$,步长 $r = 0.05$,以综合考虑计算效率与计算精度;高斯滤波器窗口设置为 5×5,参考文献[25]中的默认设置;式(22)中设置参数 $\lambda = 10^{-9}$,$\varepsilon = 10^{-10}$,同样参考文献[19]中的默认设置。

本文进行了 2 种类型的实验。一种是对真实图像进行下采样之后的仿真图像,也称为有参考遥感图像融合实验,即遵循 Wald 协议[26],原图像作为参考图像,仿真图像包含四大数据集共计 80 组图片。另一种是真实图像的实验,也称为无参考遥感图像融合实验,真实图像包括四大数据集共计 50 组图片。

用于对比的方法包括最新的几种方法,其中基于 MRA 模型的方法有基于调制传递函数模型的多尺度分析方法(CBD)[15]、加性小波亮度比的方法(Additive Wavelet Luminance Proportion, AWLP)[27]、基于双边滤波亮度比的方法(Bilateral Filter Luminance Proportional, BFLP)[28];将 MRA 和 CS 两种模型相结合的方法有基于多尺度引导滤波和改进的成分替代法(Improved IHS and Multi-Scale Guided Filter, IMG)[11];将注入模型与其他算法相结合的方法有基于抠图模型和多尺度变换的方法(Matting Model and Multi-scale Transform, MMMT)[24]、基于自适应光谱与亮度调制的方法(Adaptive Spectral Intensity Modulation Pansharpening, ASIMP)[25]。对这些方法分别进行主观评价与客观评价,也即定性和定量评价。

3.2 质量评价

评价包括肉眼视觉的主观评价和客观评价指标的定量评价。常用的有参考图像的客观评价指标如下:

1)相关系数(Correlation Coefficient, CC)[12]。主要用于评价融合图像与参考图像的空间相似度,其值区间为[0, 1],越接近 1 表示 2 幅图像越相近,融合效果越好。

2)通用图像质量指数(Universal Image Quality Indexes, UIQI)[29]。UIQI 联合相关度损耗、亮度失真和对比度失真 3 个因子来评价融合图像的效果,值区间为[0,1]。UIQI 的值越接近 1,说明结构失真越小。

3)均方根误差(Root Mean Square Error, RMSE)[30]。RMSE 计算融合图像与参考图像各像素之间均方根误差,用于表示两者的差异,其理想值为 0。

4)相对平均光谱误差(Relative Average Spectral Error, RASE)[31]。RASE 反映融合图像在光谱方面的误差,其理想值为 0。

5)光谱角映射(Spectral Angle Mapper, SAM)[32]。SAM 用于度量 2 幅图像之间的光谱信息接近程度,其值越小,光谱失真越小,理想值为 0。

6)全局相对光谱损失(ERGAS)[33]。ERGAS 用于评估融合图像的空间和光谱质量,其值越小说明总体性能越好,理想值为 0。

常用的无参考图像的客观评价指标包括 D_s、D_λ 和 QNR[34]。其中,QNR 值由 D_s 和 D_λ 共同计算而得,并基于 UIQI 测度评价融合图像和参考图像间的局部相关性、亮度和对比度。D_s 和 D_λ 越小越好,理想值为 0,而 QNR 值越大越好,理想值为 1。

3.3 算法性能分析

为了验证 2.1 节~2.3 节提出的优化方法的有效性,定量说明本文方法所带来的性能提升,分别计算各优化步骤的平均客观评价指标。4 个优化步骤对应的融合图像表示为:F1 表示初始融合图像;F2 表示经高斯滤波优化细节后图像;F3 表示经自适应优化注入量系数后的图像;F4 表示加入自适应边缘保持后的图像。并分别计算 80 组图像融合结果的平均客观评价指标。这 4 个步骤是依次递进的过程。为了更直观地展示各指标的相对变化,各客观评价指标分别进行最大值归一化处理,结果如图 3 所示。可知,对于初始融合图像 F1,注入细节是通过 PAN 图像与 MS 图像的 $\boldsymbol{I}$ 分量进行直方图匹配再由引导滤波得到,细节存

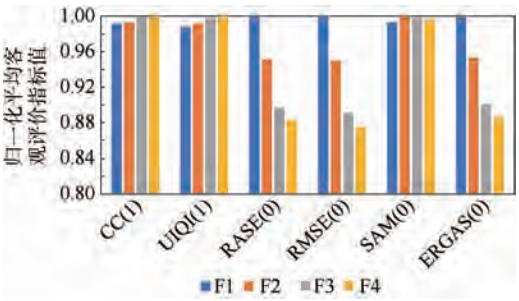


图 3　优化过程平均客观评价指标对比

Fig. 3　Average objective evaluation indicator comparison of optimization steps

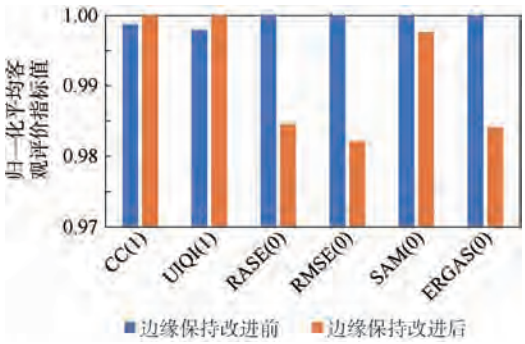


图 4　边缘保持方法改进前后平均客观评价指标对比

Fig. 4　Comparison of average objective evaluation indicators before and after improvement of edge preserving method

在与 MS 图像线性非相关的问题,且未考虑注入量,从而导致融合结果与参考图像的光谱相关度和空间相关度较低;通过估计的高斯滤波器优化细节后,所得的细节与 MS 图像有更高的相关性,各项指标明显提升;而优化注入量系数后,有效防止了注入过多或注入不足,所有的指标进一步提升;进行边缘保护后,改进的边缘保持方法在保护边缘信息的同时,减少了光谱失真。随着 F2、F3、F4 优化步骤的依次进行,总体指标也越来越优。其中 SAM 指标基本保持平稳,随着优化的进行,SAM 指标先小幅上升再小幅降低,说明本文方法对光谱信息损失的影响很小。

如 2.3 节所述,本文提出的一种新的自适应边缘保持方法是对文献[11]的改进。为了验证本文方法的有效性,在保持前 3 个步骤不变的前提下,对改进前[11]与改进后的方法在 80 组遥感图像上进行平均客观评价指标对比,指标同样进行最大值归一化处理,结果如图 4 所示。可知,改进后的边缘保持方法在所有指标上都有所提升,其中,在 RASE、RMSE 及 ERGAS 等指标上提升效果明显,可知本文方法能有效地提升图像融合的质量,减少光谱损失。

3.4　仿真实验

本文采用 3 组来自不同卫星的遥感图像作为展示,分别进行主观效果和客观评价指标的评价,然后对 80 组图像测量平均客观评价指标。

第 1 组图像来自 QuickBird 数据集,参考图像大小为 256×256,经不同方法融合后主观效果如图 5 所示。小红色矩形框区域经放大后在大矩形框中显示。可知,AWLP 方法、CBD 方法、IMG 方法森林区域存在明显光谱失真,与参考图像相比,颜色偏亮;BFLP 方法、MMMT 方法有稍微的光谱失真,且存在过度注入问题,引入了不必要的噪声。ASIMP 方法与本文方法在海滩细节与森林的光谱信息上与参考图像较为接近,难以用肉眼区别,用表 1 所示的客观评价指标加以区分。表中:黑体数据表示最优值,下划线数据表示次优

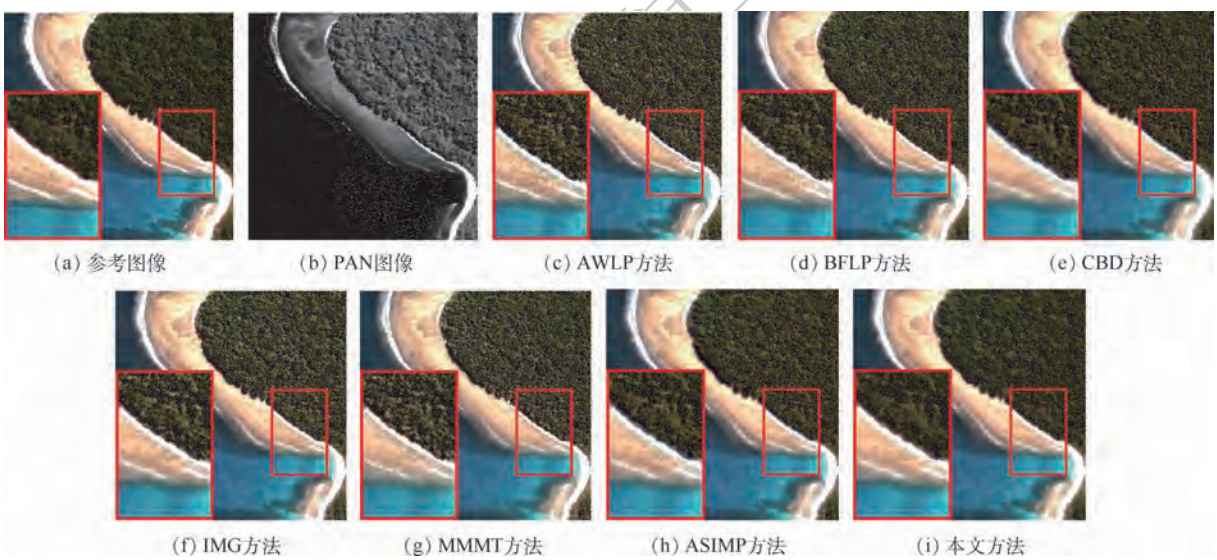


图 5　QuickBird 数据集遥感图像融合结果

Fig. 5　Fusion results of remote sensing images from QuickBird dataset

值。可知,本文方法除了在 SAM 指标上比 BFLP 方法略低以外,在所有其他指标上都达到最优。

第 2 组图像来自 WorldView-2 数据集,参考图像大小为 256×256,经不同方法融合后主观效果如图 6 所示,客观评价指标如表 2 所示。由图中红色放大矩形框可知,IMG 方法有较为明显的光谱真失,亮度偏亮;AWLP 方法、BFLP 方法和 CBD 方法有轻微的光谱失真,表现为屋顶区域颜色更浅,且在屋顶、中间白色地面有较为明显的噪声;MMMT 方法、ASIMP 方法与本文方法能较好地保持光谱信息,但 MMMT 方法在边缘上更加模糊,而本文方法与参考图像不论在空间信息上还是光谱信息上都更为接近。为更好地区分融合效果,可通过表 2 查看融合结果客观评价指标,可知,本文方法在所有指标上都达到最优。

第 3 组图像来自 pleiades 数据集,经不同方法融合后主观效果如图 7 所示,客观评价指标如表 3 所示。从放大的红色矩形区域可知,AWLP

表 1 图 5 对应的遥感图像融合结果客观评价指标

Table 1 Objective evaluation indicators of remote sensing image fusion results in Fig. 5

方法	CC(1)	UIQI(1)	RASE(0)	RMSE(0)	SAM(0)	ERGAS(0)
AWLP	0.954 3	0.842 5	24.337 7	19.664 9	4.835 5	7.141 5
BFLP	0.961 5	0.848 1	21.473 9	17.350 9	**2.944 6**	6.763 3
CBD	0.974 0	0.861 9	18.325 4	14.806 9	5.320 4	5.191 5
IMG	0.946 8	0.842 5	26.540 4	21.444 7	3.506 1	8.677 0
MMMT	0.962 9	0.844 2	21.438 4	17.322 2	5.585 2	5.609 1
ASIMP	0.977 3	0.862 4	17.356 5	14.024 1	5.107 4	4.722 1
本文	**0.982 9**	**0.884 3**	**14.008 4**	**11.318 8**	2.951 5	**3.964 6**

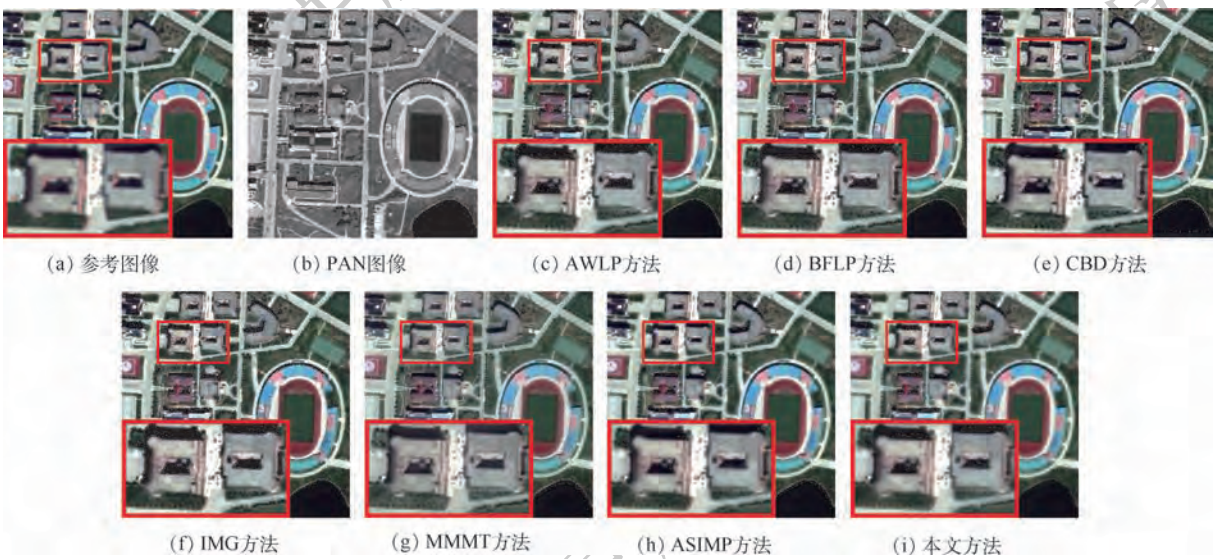


图 6 WorldView-2 数据集遥感图像融合结果

Fig. 6 Fusion results of remote sensing images from WorldView-2 dataset

表 2 图 6 对应的遥感图像融合结果客观评价指标

Table 2 Objective evaluation indicators of remote sensing image fusion results in Fig. 6

方法	CC(1)	UIQI(1)	RASE(0)	RMSE(0)	SAM(0)	ERGAS(0)
AWLP	0.908 7	0.846 3	23.071 4	25.239 6	5.531 1	5.629 8
BFLP	0.919 0	0.850 9	20.849 9	22.809 3	5.185 0	5.137 5
CBD	0.948 3	0.855 1	17.116 3	18.724 8	5.660 9	4.210 8
IMG	0.916 7	0.855 8	20.678 2	22.621 5	5.179 1	5.177 6
MMMT	0.936 8	0.840 1	17.873 8	19.553 6	6.018 0	4.468 8
ASIMP	0.957 6	0.861 9	14.912 1	16.313 6	4.966 9	3.712 6
本文	**0.957 9**	**0.862 4**	**14.592 5**	**15.963 8**	**4.775**	**3.639 7**

方法、BFLP 方法和 CBD 方法存在较为明显的光谱失真,且引入了较多的噪声;IMG 方法、MMMT 方法和 ASIMP 方法有较轻微的光谱失真,引入少量的噪声;而本文方法能在保持空间细节的同时,较好地保留光谱信息,色彩更加真实。对于难以用肉眼明显区分的方法,可通过表 3 查看融合结果客观评价指标,可知,本文方法除了在 UIQI 指标上比 ASIMP 略低以外,在所有其他指标上都达到最优。

为了更客观地评价各方法的性能,对来自 QuickBird、IKONOS、WorldView-2 及 pleiades 四大数据集的共计 80 组有参考遥感图像融合后计算平均客观评价指标,结果如表 4 所示。可知,本文方法在所有的平均客观评价指标上达到了最优。

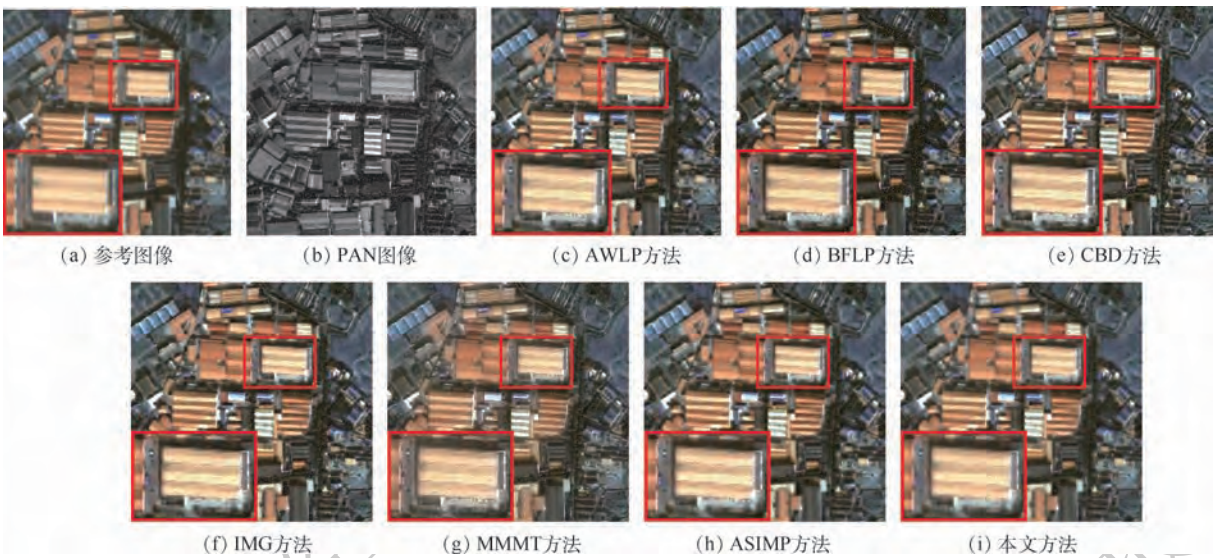


图 7　pleiades 数据集遥感图像融合结果

Fig. 7　Fusion results of remote sensing images from pleiades dataset

表 3　图 7 对应的遥感图像融合结果客观评价指标

Table 3　Objective evaluation indicators of remote sensing image fusion results in Fig. 7

方法	CC(1)	UIQI(1)	RASE(0)	RMSE(0)	SAM(0)	ERGAS(0)
AWLP	0.922 7	0.907 3	19.585 9	20.150 3	4.060 0	4.847 1
BFLP	0.926 7	0.903 4	21.395 9	22.012 5	3.726 1	5.906 0
CBD	0.907 5	0.892 4	21.936 6	22.568 8	4.207 2	5.509 9
IMG	0.927 8	0.909 1	19.256 1	19.811 0	3.222 4	4.840 1
MMMT	0.923 2	0.909 2	17.370 1	17.870 6	4.153 9	4.347 1
ASIMP	0.939 0	**0.918 1**	16.884 7	17.371 3	3.573 3	4.251 1
本文	**0.940 5**	0.913 3	**15.644 5**	**16.095 3**	**3.170 0**	**3.934 0**

表 4　80 组遥感图像仿真实验的平均客观评价指标

Table 4　Average objective evaluation indicators of simulation experiments from 80 groups of remote sensing images

方法	$\overline{\text{CC(1)}}$	$\overline{\text{UIQI(1)}}$	$\overline{\text{RASE(0)}}$	$\overline{\text{RMSE(0)}}$	$\overline{\text{SAM(0)}}$	$\overline{\text{ERGAS(0)}}$
AWLP	0.910 5	0.865 7	21.373 0	14.326 7	3.718 4	5.261 6
BFLP	0.905 4	0.827 9	29.250 6	18.965 0	3.900 0	6.973 7
CBD	0.903 3	0.858 1	21.788 5	14.848 6	4.289 0	5.579 6
IMG	0.904 9	0.853 8	22.586 3	15.437 3	3.394 0	5.667 3
MMMT	0.914 7	0.873 3	18.287 9	12.496 2	3.841 8	4.679 7
ASIMP	0.917 4	0.871 7	19.916 1	13.270 7	3.845 1	5.072 7
本文	**0.921 5**	**0.881 5**	**18.244 7**	**12.333 0**	**3.303 5**	**4.638 0**

注:"——"表示平均值。

3.5　真实图像实验

真实图像的实验采用 2 组图像进行主观和客观评价,并对所有来自四大数据集的 50 组真实图像计算平均客观评价指标。

第 1 组图像来自 IKONOS 数据集,图像大小为 780×608。不同方法融合后的主观效果如图 8

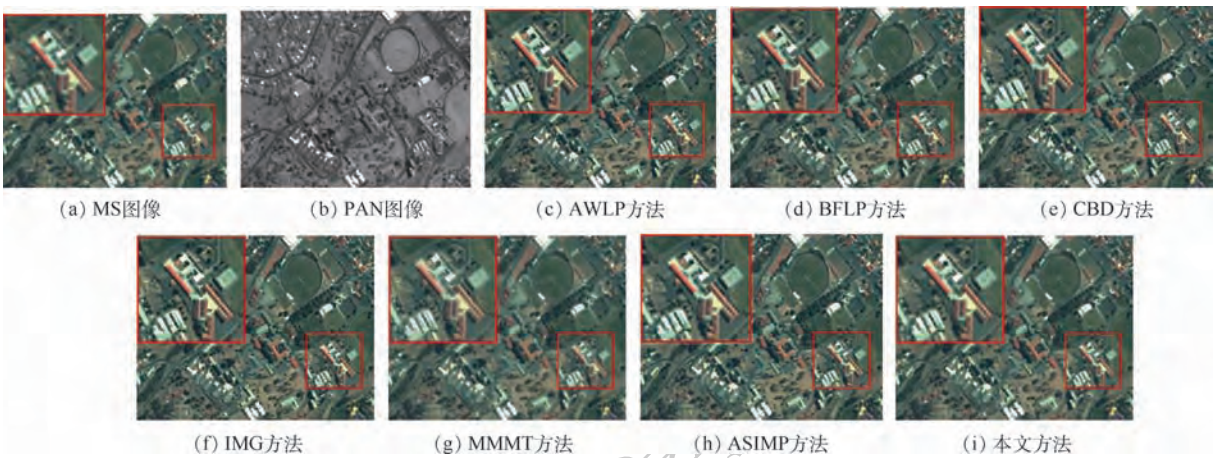


图 8 IKONOS 数据集真实图像融合结果

Fig. 8 Fusion results of real images from IKONOS dataset

所示,相应的客观评价指标如表 5 所示。由图 8 可知,BFLP 方法和 MMMT 方法在橙色屋顶区域有明显光谱失真,其他方法与本文方法难以用肉眼直接判别,可通过表 5 中图 8 的客观评价指标来评价,可知,本文方法除了在 D_s 指标上略低于 ASIMP 方法以外,其他指标都达到最优。

第 2 组图像来自 QuickBird 数据集,图像大小为 512 × 512,融合后视觉效果如图 9 所示。可知,AWLP 方法、CBD 方法和 IMG 方法亮度偏亮,存在过度注入的问题,其他方法难以用肉眼直接观察,可由表 6 中图 9 的客观评价指标测量,可知,本文方法在所有客观评价指标上都达到最优。

为更加客观评价不同方法的有效性,本文对来自四大数据集的 50 组真实图像进行实验,并测算平均客观评价指标,如表 7 所示。可知,本文方法所有指标都达到最优。

表 5 图 8 对应的真实图像定量评价结果

Table 5 Quantitative evaluation results of real image in Fig. 8

方法	$D_\lambda(0)$	$D_s(0)$	QNR(1)
AWLP	0.040 3	0.052 8	0.909 1
BFLP	0.039 4	0.055 8	0.906 9
CBD	0.039 6	0.051 3	0.911 1
IMG	0.050 5	0.056 2	0.896 1
MMMT	0.043 9	0.071 0	0.888 2
ASIMP	0.035 5	**0.048 1**	0.918 1
本文	**0.028 1**	0.049 0	**0.924 3**

图 9 QuickBird 数据集真实图像融合结果

Fig. 9 Fusion results of real images from QuickBird dataset

表 6　图 9 对应的真实图像定量评价结果

Table 6　Quantitative evaluation results of real image in Fig. 9

方法	$D_\lambda(0)$	$D_s(0)$	QNR(1)
AWLP	0.022 0	0.040 9	0.938 1
BFLP	0.013 9	0.042 4	0.944 3
CBD	0.026 0	0.043 9	0.931 2
IMG	0.020 4	0.046 6	0.934 0
MMMT	0.019 6	0.043 4	0.937 8
ASIMP	0.019 4	0.040 1	0.941 2
本文	**0.011 7**	**0.038 8**	**0.949 9**

表 7　50 组真实图像的平均定量评价结果

Table 7　Average quantitative evaluation results of 50 groups of real images

方法	$\overline{D_\lambda(0)}$	$\overline{D_s(0)}$	$\overline{\mathrm{QNR}(1)}$
AWLP	0.018 4	0.043 1	0.939 4
BFLP	0.018 0	0.042 1	0.940 8
CBD	0.020 9	0.043 0	0.937 1
IMG	0.024 5	0.048 4	0.928 4
MMMT	0.018 6	0.047 3	0.935 1
ASIMP	0.024 9	0.047 8	0.930 6
本文	**0.013 8**	**0.040 8**	**0.946 0**

注:"——"表示平均值。

4　结　论

1)本文方法在优化注入细节的过程中使用高斯滤波估计去匹配 MS 传感器的特性,并用该估计的高斯滤波器去卷积 PAN 图像,得到优化后的细节。

2)对于注入效益的优化,本文方法综合考虑光谱信息和细节信息,并建立综合评价指标,以确定自适应的注入量系数。

3)为加强边缘细节,本文方法提出一种光谱信息保真的自适应边缘提取方法,以对注入效益进一步优化。

4)通过对来自 QuickBird、IKONOS、WorldView-2 及 pleiades 四大数据集的 80 组仿真图像与 50 组真实图像进行融合实验,并将本文方法与一些先进的融合方法进行对比,结果表明本文提出的方法在主观和平均客观评价指标上能达到最优。

在未来的工作中,可能需要精细化高斯滤波估计,以得到更准确的滤波器。这些工作将在以后的研究中进一步尝试。

参考文献(References)

[1] GHASSEMIAN H. A review of remote sensing image fusion methods[J]. Information Fusion,2016,32:75-89.

[2] GEMINE V,PAOLO A,ROCCO R,et al. Pansharpening based on deconvolution for multiband filter estimation[J]. IEEE Transactions on Geoscience and Remote Sensing,2019,57(1):540-552.

[3] MALPICA J A. Hue adjustment to IHS pan-sharpened IKONOS imagery for vegetation enhancement[J]. IEEE Geoscience and Remote Sensing Letters,2007,4(1):27-31.

[4] PSJR C,SIDES S C,ANDERSON J A. Comparison of three different methods to merge multiresolution and multispectral data:Landsat TM and SPOT panchromatic[J]. Photogrammetric Engineering & Remote Sensing,1991,57(3):265-303.

[5] LABEN C A,BROWER B V. Process for enhancing the spatial resolution of multispectral imagery using pan-sharpening:U. S. 6011875 A[P]. 2000-01-04.

[6] MENG X C,SHEN H F,LI H F,et al. Review of the pansharpening methods for remote sensing images based on the idea of meta-analysis:Practical discussion and challenges[J]. Information Fusion,2019,46:102-113.

[7] LI S,KWOK J T,WANG Y. Using the discrete wavelet frame transform to merge Landsat TM and SPOT panchromatic images[J]. Information Fusion,2002,3(1):17-23.

[8] NUNEZ J,OTAZU X,FORS O,et al. Multiresolution-based image fusion with additive wavelet decomposition[J]. IEEE Transactions on Geoscience and Remote Sensing,1999,37(3):1204-1211.

[9] CUNHA A L D,ZHOU J,DO M N. The nonsubsampled contourlet transform:Theory,design,and applications[J]. IEEE Transactions on Image Processing,2006,15(10):3089-3101.

[10] YIN M,LIU W,ZHAO X,et al. A novel image fusion algorithm based on nonsubsampled shearlet transform[J]. Optik-International Journal for Light and Electron Optics,2014,125(10):2274-2282.

[11] YANG Y,WAN W,HUANG S Y,et al. Remote sensing image fusion based on adaptive HIS and multiscale guided filter[J]. IEEE Access,2016,4:4573-4582.

[12] YANG Y,WU L,HUANG S Y,et al. Compensation details-based injection model for remote sensing image fusion[J]. IEEE Geoscience and Remote Sensing Letters,2018,15(5):734-738.

[13] YANG Y,WU L,HUANG S Y,et al. Remote sensing image fusion based on adaptively weighted joint detail injection[J]. IEEE Access,2018,6:6849-6864.

[14] GIUSEPPE M,DAVIDEC,LUISA V,et al. Pansharpening by convolutional neural networks[J]. Remote Sensing,2016,8(7):594.

[15] AIAZZI B,ALPARONE L,BARONTI S,et al. MTF-tailored multiscale fusion of high-resolution MS and pan imagery[J]. Photogrammetric Engineering and Remote Sensing,2006,72(5):591-596.

[16] VIVONE G,RESTAINO R,CHANUSSOT J. Full scale regression-based injection coefficients for panchromatic sharpening[J]. IEEE Transactions on Image Processing,2018,27(7):3418-3431.

[17] YIN H,LI S. Pansharpening with multiscale normalized nonlocal means filter:A two-step approach[J]. IEEE Transactions on

Geoscience and Remote Sensing,2015,53(10):5734-5745.

[18] VIVONE G,ALPARONE L,CHANUSSOT J,et al. A critical comparison among pansharpening algorithms[J]. IEEE Transactions on Geoscience and Remote Sensing,2015,53(5):2565-2586.

[19] RAHMANI S,STRAIT M,MERKURJEV D,et al. An adaptive IHS pan-sharpening method[J]. IEEE Geoscience and Remote Sensing Letters,2010,7(4):746-750.

[20] LEUNG Y,LIU J,ZHANG J. An improved adaptive intensity-hue-saturation method for the fusion of remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters,2013,11(5):985-989.

[21] HE K,SUN J,TANG X. Guided image filtering[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(6):1397-1409.

[22] 杨勇,阙越,黄淑英,等. 多视觉特征和引导滤波的鲁棒多聚焦图像融合[J]. 计算机辅助设计与图形学学报,2017,29(7):1324-1333.
YANG Y,QUE Y,HUANG S Y,et al. Robust multifocus image fusion via multiple visual features and guided filtering[J]. Journal of Computer-Aided Design & Computer Graphics,2017,29(7):1324-1333(in Chinese).

[23] 敬忠良,肖刚. 图像融合:理论与应用[M]. 北京:高等教育出版社,2007:160-172.
JING Z L,XIAO G. Image fusion:Principle and application[M]. Beijing:Higher Education Press,2007:160-172(in Chinese).

[24] YANG Y,WANG W G,HUANG S H,et al. Anovel pan-sharpening framework based on matting model and multiscale transform[J]. Remote Sensing,2017,9(4):391-750.

[25] YANG Y,WU L,HUANG S Y,et al. Pansharpening for multiband images with adaptive spectral-intensity modulation[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,2018,11(9):3196-3208.

[26] WALD L,RANCHIN T,MANGOLINI M. Fusion of satellite images of different spatial resolutions:Assessing the quality of resulting images[J]. Photogrammetric Engineering and Remote Sensing,1997,63(6):691-699.

[27] OTAZU X,GONZÁLEZ-AUDÍCANA M,FORS O,et al. Introduction of sensor spectral response into image fusion methods. Application to wavelet-based methods[J]. IEEE Transactions on Geoscience and Remote Sensing,2005,43(10):2376-2385.

[28] KAPLAN N H,ERER I. Bilateral filtering-based enhanced pansharpening of multispectral satellite images[J]. IEEE Geoscience and Remote Sensing Letters,2014,11(11):1941-1945.

[29] WANG Z,BOVIK A C. A universal image quality index[J]. IEEE Signal Processing Letters,2002,9(3):81-84.

[30] YANG Y,TONG S,HUANG S Y,et al. Multifocus image fusion based on NSCT and focused area detection[J]. IEEE Sensors Journal,2015,15(5):2824-2838.

[31] RANCHIN T,WALD L. Fusion of high spatial and spectral resolution images:The ARSIS concept and its implementation[J]. Photogrammetric Engineering and Remote Sensing,2000,66(1):49-61.

[32] ALPARONE L,WALD L,CHANUSSOT J,et al. Comparison of pansharpening algorithms:Outcome of the 2006 GRSS data fusion contest[J]. IEEE Transactions on Geoscience and Remote Sensing,2007,45(10):3012-3021.

[33] WALD L. Quality of high resolution synthesized images:Is there a simple criterion? [C]// Proceedings of the Third Conference on Fusion of Earth Data,2000:99-103.

[34] ALPARONE L,AIAZZI B,BARONTI S,et al. Multispectral and panchromatic data fusion assessment without reference[J]. Photogrammetric Engineering and Remote Sensing,2008,74(2):193-200.

作者简介:

杨勇 男,博士,教授,博士生导师。主要研究方向:图像处理、模式识别。

卢航远 男,博士研究生。主要研究方向:图像处理。

黄淑英 女,博士,副教授。主要研究方向:图像处理、机器学习。

涂伟 男,博士研究生。主要研究方向:图像处理。

李露奕 女,硕士研究生。主要研究方向:图像处理。

Remote sensing image fusion method based on adaptive injection model

YANG Yong[1,*], LU Hangyuan[1], HUANG Shuying[2], TU Wei[1], LI Luyi[1]

(1. School of Information Technology, Jiangxi University of Finance and Economics, Nanchang 330032, China;
2. School of Software and Internet Things Engineering, Jiangxi University of Finance and Economics, Nanchang 330032, China)

Abstract: The purpose of remote sensing image fusion is to fuse high-spectral-resolution low-spatial-resolution multispectral (MS) images and high-spatial-resolution low-spectral-resolution panchromatic (PAN) images, so as to obtain the fusion images with high spectral resolution and high spatial resolution. How to determine the injection details and injection coefficients in the injection model is the key to image fusion research. For detail optimization, a multi-scale Gaussian filter is defined by simulating the characteristics of MS sensor, and then the filter is used to convolve with PAN image to extract the details and obtain the details highly related to MS image. In order to optimize the injection coefficient, an adaptive injection coefficient is proposed based on spectral information and detail information. To better preserve the edge information, a new edge preserving weight matrix is proposed to achieve the dual fidelity of spectral information and space. Finally, the optimized injection coefficient is multiplied by details and injected into the up-sampled MS image to obtain the final fusion result. Performance analysis of the method proposed in this paper has been carried out and a large number of tests have been conducted in each satellite dataset. The experimental results show that, compared with the most advanced methods, the proposed method performs best in both subjective and comprehensive objective assessment.

Keywords: remote sensing image fusion; Gaussian filter; injection coefficient; edge preserving; quality evaluation

Received: 2019-07-09; **Accepted**: 2019-08-18; **Published online**: 2019-08-26 16:13

URL: kns.cnki.net/kcms/detail/11.2625.V.20190826.1555.004.html

Foundation items: National Natural Science Foundation of China (61662026, 61862030); Natural Science Foundation of Jiangxi, China (20182BCB22006, 20181BAB202010, 20192ACB20002, 20192ACBL21008); Science Research Foundation of Education Bureau of Jiangxi Province, China (GJJ170312, GJJ170318); Knowledge Innovation Fund for the Graduate Students of Jiangxi Province (YC2019-B094, YC2018-B065)

* **Corresponding author**. E-mail: greatyangy@126.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0370

基于 DCNN 和全连接 CRF 的舌图像分割算法

张新峰*，郭宇桐，蔡轶珩，孙萌
（北京工业大学 信息学部，北京 100124）

摘 要：针对中医舌诊中舌体分割不准确、分割速度较慢且需要人工标定候选区域等问题，提出了一种端到端的舌图像分割算法。与传统舌图像分割算法相比，所提算法可以得到更为准确的分割结果，并且不需要人工操作。首先，使用孔卷积算法，可以在不增加参数的条件下扩大网络的特征图谱。其次，使用孔卷积空间金字塔池化（ASPP）模块，令网络通过不同的感受野学习舌图像的多尺度特征。最后，将深度卷积神经网络（DCNN）和全连接的条件随机场（CRF）相结合，细化分割后的舌体边缘。实验结果表明：所提算法优于传统舌图像分割算法和主流的深度卷积神经网络，具有较高的分割精度，平均交并比达到了 95.41%。

关 键 词：深度学习；卷积神经网络（CNN）；语义分割；舌图像；条件随机场（CRF）
中图分类号：TN911.73
文献标识码：A **文章编号**：1001-5965(2019)12-2364-11

舌诊在中医中占有重要地位，是了解人体状态的重要手段之一。在舌诊中，医生通过观察舌质和舌苔的颜色、形状等特征，为诊断患者的健康状况提供重要依据。在传统的舌诊望诊过程中，医生通常通过自己的眼睛来观察舌头，从而进行判断。然而视觉感知和临床经验的变化往往会导致诊断结果的差异，即诊断结果是主观的。此外，环境因素（如不同的光源和亮度）也会影响医生做出准确的诊断。传统舌诊的评价指标不能科学、定量地描述，诊断结果不能系统地存储和标准化。因此，建立更加客观、准确的中医舌诊辅助系统具有重要的意义和必要性。

舌图像分析系统[1]通常分为以下步骤：①采集带有舌体区域的人体面部图像；②对面部图像进行分割，获取舌体区域；③提取舌体的颜色、纹理、裂纹、红刺等特征，对舌色和舌苔进行定量分析和定性描述；④将分割后的舌图像、提取的特征和评价结果保存在患者数据库中，协助医生进行后续诊断。在上述步骤中，舌体区域的分割是一个关键步骤。在采集到的面部图像中有许多干扰物，包括背景、面部器官，有时还包括颜色标准色卡等，因此对舌体区域分割的准确度将直接影响到最终舌诊的分析结果。

传统的舌图像分割算法可分为以下几类：基于阈值的方法[2-3]、基于边缘检测的方法[4-6]、基于图论的方法[7-9]和基于主动轮廓模型[10-11]。这些方法在一定条件下可以得到较好的分割效果。但是，它们有以下缺点：①自动化程度较低，往往需要一些人工操作，如边界框或标定轮廓点[4,7]；②分割精度较差，并且舌体边缘存在较多齿痕。由于颜色相似，算法无法准确区分舌头和嘴唇[12]。

近年来，深度卷积神经网络（DCNN）在计算机视觉方面取得良好的表现，包括图像分类[13-14]、目标检测[15-17]和语义分割[18-21]。通过端到端的方式，DCNN 比手工选取特征得到了更好

收稿日期：2019-07-09；**录用日期**：2019-08-14；**网络出版时间**：2019-08-27 09:40
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190826.1424.002.html
基金项目：国家重点研发计划（2017YFC1703300）
* **通信作者**. E-mail：zxf@bjut.edu.cn

引用格式：*张新峰，郭宇桐，蔡轶珩，等. 基于DCNN和全连接CRF的舌图像分割算法[J]. 北京航空航天大学学报，2019，45(12)：2364-2374. ZHANG X F，GUO Y T，CAI Y H，et al. Tongue image segmentation algorithm based on deep convolutional neural network and fully conditional random fields [J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2364-2374 (in Chinese).*

的结果。尽管如此,利用 DCNN 进行舌部分割的算法还很少。最新的算法[22-23]使用 DCNN 进行舌部分割,优于一些传统的分割算法。然而,它们还需要一些额外的预处理,如亮度辨别和图像增强,这使得整个过程复杂。

本文提出了一种基于 DCNN 和全连接条件随机场(CRF)的舌图像分割算法,主要工作如下:①提出了一种基于 DCNN 的舌图像分割网络;②使用孔卷积算法在不增加网络参数的条件下扩大了特征图谱的尺寸;③使用孔卷积空间金字塔池化(ASPP)模块提取舌图像的多尺度特征;④将 DCNN 与全连接 CRF 相结合,细化分割边缘。

1　舌图像分割算法

本节提出了一种舌图像分割算法,算法的网络结构如图 1 所示。该算法由 3 部分组成:①使用孔卷积的 DCNN;②ASPP 模块;③全连接的 CRF。首先,将舌图像输入到 DCNN 进行特征提取,在网络层中使用孔卷积算法扩大感受野。然后,利用 ASPP 模块通过结合多个不同大小的感受野来提取多尺度信息。最后,利用全连接 CRF 对网络输出图像进行处理,对分割边缘进行细化。

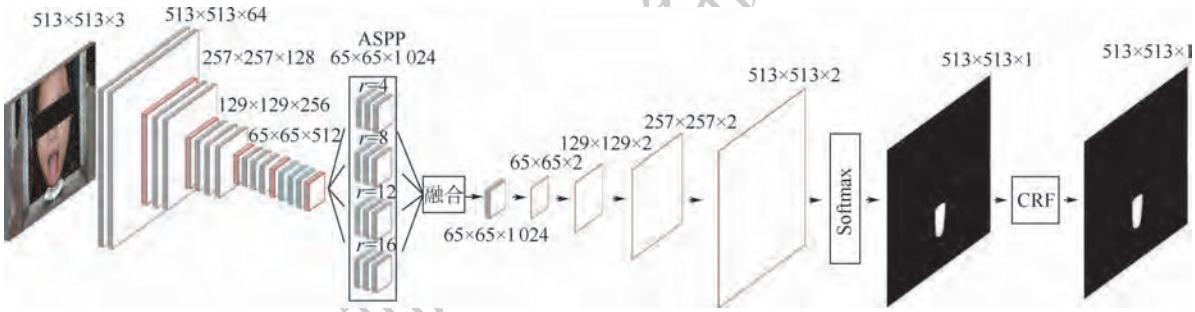


图 1　舌图像分割算法网络结构

Fig. 1　Network structure of tongue image segmentation algorithm

1.1　孔卷积

在计算机视觉领域,DCNN 通常会使用较多的卷积层和池化层,如在分类任务中,将卷积层和最大池化层在连续层上面重复组合使用,可以提取物体高维特征,分类效果表现良好。然而,这种重复使用卷积和池化的操作在分割任务中存在很严重的问题。在对图像中的舌体区域进行分割时,舌体的边界信息和语义信息是至关重要的。经过反复的卷积和池化操作后,网络的特征图谱会逐渐缩小,损失了较多的边界信息,不利于舌体边缘的精确分割。

为了减少边界信息的丢失,应避免在卷积和池化操作中使特征图谱缩减过小。如图 1 所示,算法的网络结构共有 5 个最大池化层,它们在网络中分别为第 3、6、10、14、18 层。将网络中的第 4池化层和第 5 池化层的步长参数设置为 1,而不是前 3 个池化层中的 2,这样可以使特征图谱不会过度减小,保持为一个较大尺寸。对于用于图像分类任务的 DCNN[13],最终的特征图(全连接层或全局池化层之前)比输入图像尺寸小 32 倍。使用上述操作,特征图谱只会减少为原尺寸的 8 倍。

上述方法可以增大网络的特征图谱,获得更好的边界信息,但同时会导致网络中较深层的感受野尺寸减小,不利于语义信息的获取。在语义分割任务中,边界信息和语义信息都是至关重要的,语义信息的损失同样会影响分割的准确率。所以,如果能在增大网络特征图谱的同时保证感受野尺寸不会缩小,分割结果会较为精确。感受野的公式如式(1)所示,相邻特征的距离如式(2)所示,其与步长 s 成正比。当步长固定时,扩大感受野的唯一方法是增加卷积核尺寸的大小。然而增大卷积核的尺寸会增加参数的数量,导致网络效率大幅降低。

$$r_{\text{out}} = r_{\text{in}} + (k-1)j_{\text{in}} \tag{1}$$

$$j_{\text{out}} = j_{\text{out}}s \tag{2}$$

式中:r_{in} 为前一层的感受野尺寸;r_{out} 为当前层感受野尺寸;k 为核大小;j_{in} 为前一层相邻特征的距离;j_{out} 为当前层相邻特征的距离。

为了解决此问题,本文使用了孔卷积算法,其最初是为了高效计算非抽取小波变换而开发的[24],可以在保持参数个数不变的情况下,增大卷积核的大小。算法在一维运算中如式(3)所示:

$$y[i] = \sum_{k=1}^{n} x[i + r \times k]w[k] \tag{3}$$

式中:$x[i]$ 为输入信号;$y[i]$ 为输出信号;$w[k]$ 为卷积核;参数 r 控制卷积的间隔。

在进行标准卷积时,即参数 $r=1$,从图 2(a)可以看到,在较高一层的神经元中,相邻神经元的感受野具有较高的重合性,虽然能够包含低一层

神经元中所有的信息,但是由于对像素进行了连续的遍历卷积,图像中的许多特征会被重复地进行提取,从而产生大量的冗余信息。如果将像素点间断地进行卷积运算(间隔由参数 r 控制),如图 2(b)所示,感受野的交集将会被最小化,保证了在不损失图像信息的前提下,减少冗余信息。合理的组合不同层中孔卷积的参数 r,可以使卷积层既能提取到图像的全部特征,又能在不增加参数的情况下扩大感受野,提高语义信息的提取能力。在图 1 中,网络中的孔卷积层被标记为每个通道的第一层,且参数 $r=2$。通过改变池化层的步长,以及使用孔卷积的方法,网络提取边缘信息和语义信息的能力有所提升,这有利于提高舌图像分割的准确率。

使用孔卷积还可以显式地控制在 DCNN 中计算特征响应的密度。用输出比率来表示输入图像空间分辨率与最终输出分辨率的比值,对于用于图像分类任务的 DCNN[13],最终的特征响应(在全连接层或全局池化层之前)比输入图像尺寸小 32 倍,因此输出比率为 32。如果想将 DCNN 中计算的特征响应的空间密度提高一倍(即输出比率为 16),则将最后一个缩小特征图谱的卷积层或池化层的步长设置为 1,以避免信号抽取。所有后续的卷积层都替换为 $r=2$ 的孔卷积层即可。通过调整参数 r,可以在任意分辨率下计算最终的 DCNN 响应。

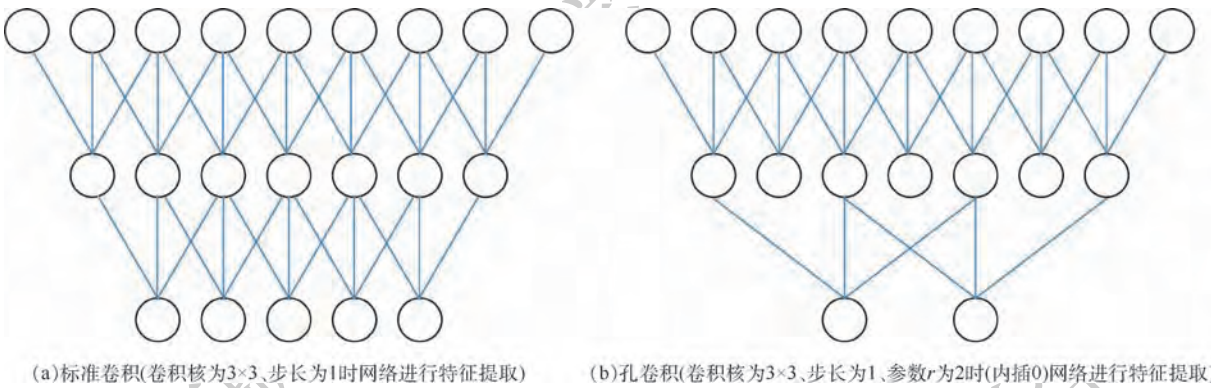


图 2 一维孔卷积示意图

Fig. 2 Schematic diagram of 1D atrous convolution

1.2 孔卷积空间金字塔池化模块

在 DCNN 中通常使用固定大小的卷积核进行卷积运算,如 3×3 或 5×5。通过重复的卷积运算,网络可以提取图像中所包含的深层特征。然而,具有固定尺寸的卷积核只能提取单一尺寸的特征,不能对图像在多个尺度上进行特征提取。

孔卷积能够控制感受野的大小,通过设置不同的卷积参数 r,感受野的大小也会发生变化。单独使用一个孔卷积层,网络可以在这个指定尺度下提取特征。同时使用多个具有不同感受野的孔卷积层,网络即可以对物体提取多尺度特征。孔卷积 ASPP 模块并行地排列 4 种不同参数的孔卷积层,通过多种感受野提取多尺度特征。每个通道提取到的特征通过 2 个卷积层进一步处理(卷积核为 1×1)。将 4 个通道提取到的特征通过一个卷积核大小为 1、深度为 4 的卷积层进行融合,结果作为最终的特征。ASPP 模块如图 3 所示。

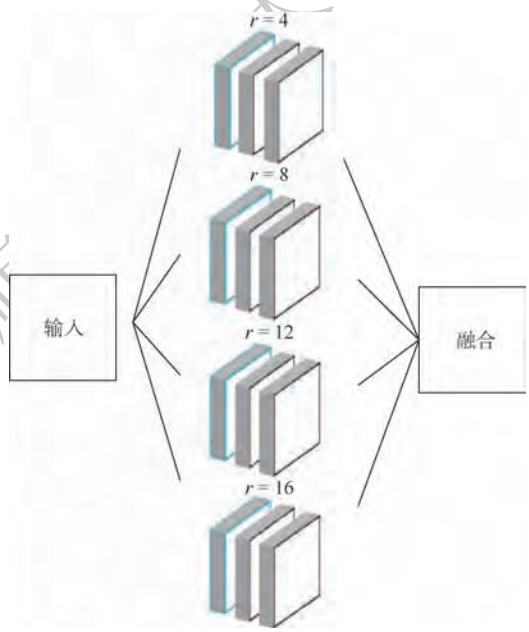


图 3 ASPP 模块

Fig. 3 ASPP module

除此之外,由于拍摄设备、拍摄距离和患者年龄的不同,图像中舌体区域尺寸的大小可能会存在差异。因此,需要分割算法能够准确地分割不同大小的舌体。为了解决此问题,一般的分割算法通常是使用包含不同尺寸物体的数据集来训练网络,确保模型对不同大小物体分割的鲁棒性。

然而,由于舌图像数据获取困难,舌图像的数量有限,没有足够的各种尺度的带标签的图像来对网络进行训练。ASPP 模块由于使用了不同大小的感受野,可以对单尺寸舌图像在多个尺度上进行特征提取,在一定程度上解决了样本不同尺寸数据量不足的问题。

1.3　全连接条件随机场

定位精度和分类性能之间的权衡是 DCNN 固有的争论,具有多个卷积层和最大池化层的深层网络模型在分类任务中被证明是成功的,但是顶层节点的不变性和较小的特征图谱的网络只能产生平滑的响应。DCNN 的不变性对语义分割有一定的限制,因为不变性意味着无论如何平移图像,最终的分类都是相同的。在语义分割任务中,物体的位置是有用的信息,DCNN 的不变性导致网络中位置信息的丢失。1.1 节中提到了 DCNN 可以可靠地预测图像中物体的存在和粗略位置,但不能精确地定位物体轮廓。因为重复地进行下采样操作,导致大量的边缘信息丢失,使得物体边缘处的像素点较为平滑。但语义分割任务需要清晰的边缘信息。除此之外,卷积运算是局部连接的,其只能提取一个像素周围矩形区域的信息,重复卷积运算虽然能使矩形面积逐渐变大,但即使到最后一个卷积层,也无法提取整个图像中一个像素与其他所有像素之间的相关性。为了解决这些问题,提高分割的准确率,将 DCNN 与全连接 CRF 相结合[25],通过计算任意 2 个像素之间的相似性来判断它们是否属于同一个类。通过这种方式利用整幅图像的信息,而不是在 DCNN 中只提取局部信息。全连接 CRF 可以突出物体的边缘,弥补了 DCNN 带来的边界平滑问题。

传统的 CRF 通常被用来平滑噪声[7],通过将相邻的节点耦合在一起,使得空间上邻近的像素获得相同的标签。这些短程 CRF 的主要功能是清除基于手工特征构建的弱分类器的虚假预测。

然而,DCNN 不同于这些较弱的分类器。DCNN 获得的边缘是十分平滑的,在这种情况下,使用短程 CRF 是不适合的[25],因为目标是得到清晰的物体边缘,而不是进一步平滑。为了解决短程 CRF 的局限性,用全连接 CRF 代替。将 DCNN 模型与全连接 CRF 相结合,用来精确地分割边界。模型能量函数如下:

$$E[i] = \sum_i \theta_i(x_i) + \sum_{ij} \theta_{ij}(x_i, x_j) \tag{4}$$

式中:x_i代表像素 i 的标签;一元势为 $\theta_i(x_i) = -\lg P(x_i)$,$P(x_i)$代表由 DCNN 预测出的像素 i 为舌体或者背景的概率;二元势[25]为以下表达:

$$\theta_{ij}(x_i, x_j) = \mu(x_i, x_j)\left[\omega_1 \exp\left(-\frac{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2}{2\sigma_\alpha^2} - \frac{\|\boldsymbol{I}_i - \boldsymbol{I}_j\|^2}{2\sigma_\beta^2}\right) + \omega_2 \exp\left(-\frac{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2}{2\sigma_\gamma^2}\right)\right] \tag{5}$$

其中:若 $x_i \neq x_j$,则 $\mu(x_i, x_j) = 1$,否则为 0,这意味着只惩罚具有不同标签的节点。公式的后半部分为 2 个在不同特征空间中的高斯核。第 1 个核同时依赖于像素位置(2 个不同的像素点坐标向量分别表示为 $\boldsymbol{p}_i$和 $\boldsymbol{p}_j$)和 RGB 颜色(2 个不同像素点的颜色向量分别表示为 $\boldsymbol{I}_i$和 $\boldsymbol{I}_j$),第 2 个核只依赖于像素位置。超参数 ω_1、ω_2 控制高斯核的权重,σ_α、σ_β 和 σ_γ控制高斯核的尺度。

式(5)的作用是判别相似的像素点是否属于同一类。如果像素点属于同一类,则能量函数值相对较小。反之,如果像素点不属于同一类,则能量函数相对较大。在舌图像分割中,唇部区域往往被错误地划分为舌形区域,影响了后续的分析。利用该能量函数,使与唇部连接的舌体分割更加精确。舌体部分像素的 RGB 值相似,舌体与唇部像素的 RGB 值存在差异。当对相似区域的像素点判别为不同类时,会产生较大的能量函数值。当存在差异的区域判别为同一类时,也会产生较大的值。通过多次迭代,使能量函数的值最小化来获得最终结果。通过这种方式,利用整个图像的信息来细化舌体边缘,提高分割的准确性。

2　实验结果分析

为了评估本文提出的舌图像分割算法的性能,进行了一系列的实验和分析。首先,比较了在网络上使用不同大小卷积核和参数 r 的孔卷积,对网络性能产生的影响。然后,比较了使用不同参数 r 的 ASPP 模块对网络性能产生的影响。最后,将本文算法与传统的舌图像分割算法及常用的 DCNN 算法进行了比较,验证了本文算法的有效性。

2.1　数 据 集

2.1.1　PASCAL VOC 2012

PASCAL VOC 2012 是一个语义分割数据集,共分为 21 个类别,其中包括 20 个前景物体和 1 个背景。原始数据集分别包含用于训练(1 464 幅)、验证(1 449 幅)和测试(1 456 幅)的像素级标记图像。数据集通过提供的方法进行扩充[26],得到 10 582 幅训练图像。数据集每个部分的数量如表 1 所示。

表 1 三个数据集的图片数量

Table 1 Number of images on three datasets

数据集	类型	图片数量
PASCAL VOC 2012	训练	10 582
	验证	1 449
	测试	1 456
Tongue dataset 1	训练	1 440
	测试	250
Tongue dataset 2	训练	160
	测试	40

2.1.2 舌图像数据集

Tongue dataset 1 采用佳能 EOS 1100D 相机拍摄,图像分辨率为 4 272 × 2 848。数据集中共有 1 690幅图像,其中包含训练集(1 440 幅)和测试集 (250 幅)。

Tongue dataset 2 采用尼康 E4500 相机拍摄,灯光和背景条件与 Tongue dataset 1 不同,拍摄的场景也存在差异,图像分辨率为 1 600 × 1 200。数据集中共有 200 幅图像,其中包含训练集(160 幅)和测试集 (40 幅)。

数据集原始图像和人工标注的标签如图 4 所示。所有受试者在参与研究前均已知情同意纳入研究。

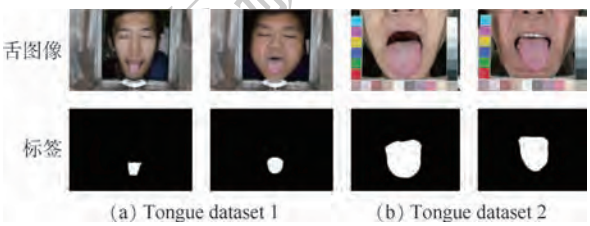


图 4 数据集原始图像和人工标注的标签

Fig. 4 Original images in dataset with corresponding artificial segmentation marks

2.2 实验设置

为了训练舌图像分割模型,使用 Caffe 框架搭建网络。所有实验都是在一台配置为 Intel Xeon E5-2623 v3 CPU @ 2.4G Hz,64 GB RAM 的电脑上进行的,GPU 型号为 NVIDIA GeForce GTX TITAN。

在实验中,mini-batch 设置为 20,初始学习率为 0.001,每 2000 次迭代将学习率乘以 0.1,动量固定为 0.9,对应的权值衰减为 0.0005。

2.3 评判标准

本文采用像素精度(PA)、平均像素精度(MPA)和平均交并比(MIOU)3 个标准来评价模型对舌图像的分割精度。PA、MPA、MIOU 的公式如下:

$$\mathrm{PA} = \frac{\sum_{i=0}^{m} p_{ii}}{\sum_{i=0}^{m}\sum_{j=0}^{m} p_{ij}} \tag{6}$$

$$\mathrm{MPA} = \frac{1}{m+1}\sum_{i=0}^{m}\frac{p_{ii}}{\sum_{j=0}^{m} p_{ij}} \tag{7}$$

$$\mathrm{MIOU} = \frac{1}{m+1}\sum_{i=0}^{m}\frac{p_{ii}}{\sum_{j=0}^{m} p_{ij} + \sum_{j=0}^{m} p_{ji} - p_{ii}} \tag{8}$$

式中:p_{ii} 为将第 i 类正确地分割为第 i 类的像素点数量,p_{ij} 为将第 i 类错误地分割为第 j 类的像素点数量,p_{ji} 为将第 j 类错误地分割为第 i 类的像素点数量。

2.4 结果分析

先将网络在 PASCAL VOC 2012 数据集上进行预训练,再在舌图像数据集上对网络进行微调。在对 DCNN 进行微调后,使用交叉验证来优化 CRF 参数。初始化参数 $\omega_2 = 3$,$\sigma_\gamma = 3$,通过对验证集的小部分数据进行交叉验证(本文使用20 幅图像)来寻找最佳参数 ω_1、σ_α 和 σ_β。

采用由粗到精的搜索方案,参数的初始搜索区间为:ω_1 为[2:6],σ_α 为[10:10:100],σ_β 为[10:10:100]。算法共迭代 10 次。

2.4.1 不同尺寸和参数的孔卷积

本节讨论使用不同尺寸和参数的孔卷积对网络性能的影响。为了便于实验结果的观察,在 fc6 中使用了一个分支,来代替在 ASPP 模块中使用的 4 个分支。如表 2 所示,在孔卷积层中使用大小为 7 × 7,参数 $r = 2$ 的卷积核。由于卷积核尺寸较大,网络的感受野较大,可以提取更多的特征,分割效果较好,该模型在 CRF 后的性能为 93.29%。但是 7 × 7 的卷积核导致网络参数庞大,运行缓慢(训练时每秒 1.57 幅图像)。通过将核的尺寸减小到 5 × 5($r = 2$),将模型速度提高到每秒 2.47 幅图像,将核大小减小到 3 × 3($r = 2$),将模型速度提高到每秒4.92幅图像。由于减小了

表 2 不同尺寸和参数的孔卷积对网络性能的影响

Table 2 Effect of atrous convolution with different size and parameters on network performance

核尺寸	r	参数数量/10^6	时间/s	MIOU/%
7 × 7	2	134.3	1.57	93.29
5 × 5	2	94.6	2.47	89.16
3 × 3	2	20.5	4.92	83.47
3 × 3	4	20.5	4.92	86.13
3 × 3	6	20.5	4.92	90.06
3 × 3	8	20.5	4.92	93.27

卷积核的大小,参数数量大幅减少,运行速度更快。然而感受野的减小导致网络性能下降,分割精度降低。当使用 3×3 的卷积核($r=2$)时,CRF 后的模型性能为 83.47%。不同参数下的模型结果如图 5 所示。

为了在不增加参数的情况下提高模型的精度,在 fc6 层增大参数以扩大网络的感受野。当使用 3×3 卷积核,参数设置为 $r=8$ 时,模型的性能可以与之前的结果(卷积核大小 7×7,$r=2$)相媲美。使用这种方法,能够令网络在不增加参数的条件下,增大网络的感受野。

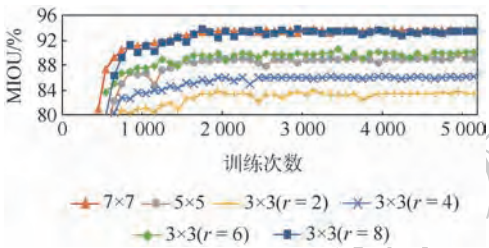


图 5　不同尺寸和参数的孔卷积 MIOU 结果

Fig. 5　MIOU of atrous convolution with different sizes and parameters

2.4.2　不同参数的孔卷积空间金字塔池化模块

本节讨论在 ASPP 模块中使用不同孔卷积参数 r 的效果。如图 3 所示,ASPP 模块在 fc6-fc7-fc8 中使用多个并行分支,最后融合生成结果。实验中卷积核的大小均为 3×3,每个分支的参数 r 不同。表 3 展示了几种参数的实验结果。表中:单分支是基准模型,即 fc6 为单分支,$r=8$。另外 4 个实验使用的是 ASPP 模块,有 4 个分支,参数不同。ASPP-2 代表 $r=(2,4,6,8)$,ASPP-4 代表 $r=(4,8,12,16)$,ASPP-6 代表 $r=(6,12,18,24)$,ASPP-8 代表 $r=(8,16,24,32)$。实验结果如表 3 所示,使用 ASPP 模块的性能优于单个分支。在使用 ASPP 模块进行的 4 个用于舌图像分割的实验中,ASPP-4(4,8,12,16)的性能最好,CRF 后分割的 MIOU 达到 95.41%。从图 6 中可以观察到不同参数下模型的分割结果。

表 3　不同参数的 ASPP 模块对网络性能的影响

Table 3　Effect of different parameters of ASPP module on network performance

方法	各通道参数	MIOU/%
单分支	8	93.27
ASPP-2	(2,4,6,8)	94.18
ASPP-4	(4,8,12,16)	95.41
ASPP-6	(6,12,18,24)	94.79
ASPP-8	(8,16,24,32)	94.57

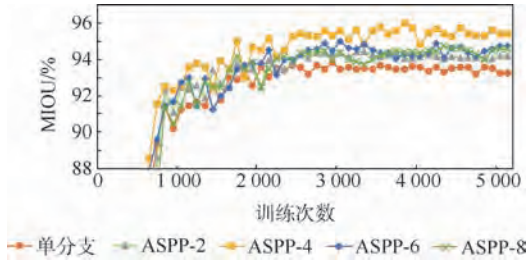


图 6　不同参数的 ASPP 模块 MIOU 结果

Fig. 6　MIOU of ASPP module with different parameters

2.4.3　不同模块效果

为证明网络结构的有效性,本节比较使用不同模块对分割结果的影响。实验中使用了 2 个不同的舌图像数据集。使用 Tongue dataset 1 中的训练集对模型进行训练,并用测试集对模型结果进行测试,实验结果如表 4 所示。结果表明,在网络中使用 ASPP 模块和 CRF 对分割结果有 3% 的提升。使用 ASPP 模块与全连接 CRF 的方法取得了较好的效果,在舌图像数据集中测试 MIOU 达到了 95.41%。可视化分割结果示例如图 7 所示。

在实际应用中,舌图像采集时的环境通常会有所不同,采用训练好的模型对新环境下的舌形图像进行分割,分割效果会下降。因此,当设备应用于新的环境时,可以使用少量的新数据对网络参数进行微调,提高模型的泛化能力。

基于在 Tongue dataset 1 中训练的模型,使用 Tongue dataset 2 中的 40 幅图像进行微调,并使用测试集进行测试。Tongue dataset 2 的分割结果略低于 Tongue dataset 1,但相差不多,最佳分割效果可达 93.75%。

表 4　不同模块对网络性能的影响

Table 4　Effect of different modules on network performance

单分支	ASPP	CRF	MIOU/%	
			Tongue dataset 1	Tongue dataset 2
√			92.48	90.87
√		√	93.27	91.21
	√		94.36	92.54
	√	√	95.41	93.75

2.4.4　与现有算法对比

表 5 为 GrabCut[7]、Snake[4]、FCN-8s[18]、U-net[27]、SegNet[20] 及本文算法在 2 个舌图像数据集上分割结果的对比。在 GrabCut 模型中,手动选择一个边界框,来标定舌体区域,然后迭代算法 10 次,得到最终的分割结果。在 Snake 模型中,手工将舌体边缘标定 15 个标记以此作为分割初

始边界,通过算法计算得到最终的分割结果。

由于 GrabCut[7] 和 Snake[4] 算法的自动化程度较低,需要手工标记操作,分割时间分别为 7.25 s和 6.33 s,分割效率较低。本文算法不需要人工干预,对模型进行训练后,在 2 个不同的数据集上进行测试,运行时间分别为 0.83 s 和0.75 s,速度优于 GrabCut 和 Snake 算法。然而,由于在网络的末端使用全连接 CRF,计算复杂度较高,因此速度低于 FCN-8s、U-net 和 SegNet 网络。

在精度方面,传统算法为了缩短计算时间,通常对图像进行降采样,降低图像的分辨率,再对图像进行分割。由于在降低图像分辨率时会丢失大量的图像信息,使用这种方法会导致分割后的舌体区域存在较为明显的棱角,舌体区域不完整,分割效果较差。本文算法对原始图像进行分割,并且网络的特征图谱尺寸较大,舌体分割后存在的齿痕较少,舌体区域更加完整。分割结果的示例如图 8 所示。使用 GrabCut 和 Snake 算法得到的分割结果如图8(b)和图8(e)所示,舌体的边缘没有分割完整,并且有一些缺失的部分会影响舌体的纹理特征。图 8(c)和图 8(f)是本文算法的分割结果,与传统算法相比,其得到的舌体边缘更加精确和完整。

此外,由于舌体和嘴唇区域存在交集,并且 2 种颜色较为相似,传统算法在这种情况下有时分割错误,将部分嘴唇区域判断为舌体区域。由于 DCNN 可以自动提取抽象特征,并利用全连接 CRF 对边缘进行细化,在这种情况下,本文算法可以更加准确地分割舌体区域,分割精度更高。如图 9 所示,图 9(b)和图 9(e)分别为 GrabCut 和 Snake 算法处理后的分割结果。传统的分割算法有时将嘴唇误判为舌体,分割效果较差。在这种情况下,本文算法可以比传统算法更好地区分舌体和嘴唇,获得良好的分割结果。图 9(c)和图 9(f)是使用本文算法得到的分割结果。在分割精度方面,本文算法在 PA、MPA、MIOU 三个方面都优于其他分割算法。不同算法的分割结果示例如图 10 所示。

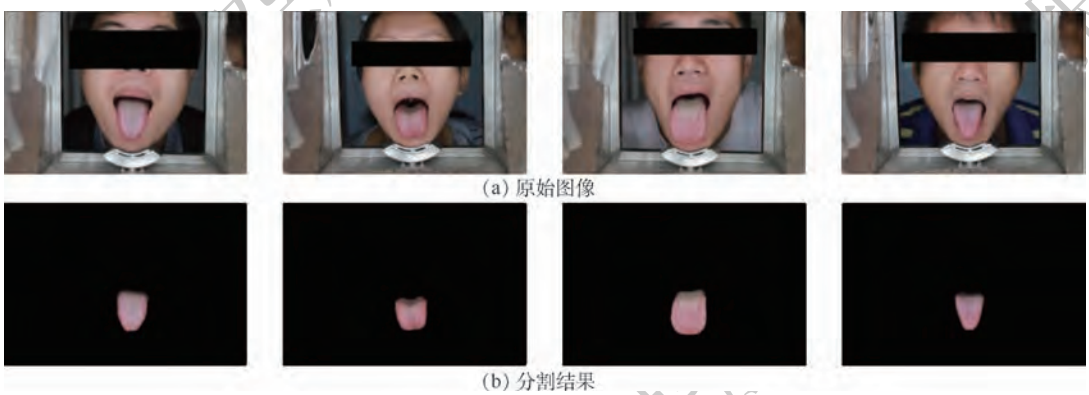


图 7 舌图像的分割结果示例

Fig. 7 Examples of tongue image segmentation results

表 5 不同算法在舌图像数据集上的分割结果

Table 5 Segmentation results on tongue image dataset by different algorithms

算法	Tongue dataset 1				Tongue dataset 2			
	PA/%	MPA/%	MIOU/%	时间/s	PA/%	MPA/%	MIOU/%	时间/s
GrabCut[7]	96.22	95.47	83.89	7.25	96.63	95.38	86.81	6.87
Snake[4]	97.15	96.39	90.43	6.33	98.50	96.71	93.54	6.57
FCN-8s[18]	97.36	96.04	91.37	0.37	96.84	95.27	90.36	0.31
U-net[27]	98.65	96.88	93.69	0.68	97.83	96.19	92.17	0.64
SegNet[20]	99.71	98.09	94.81	0.48	98.02	96.51	92.63	0.42
本文算法	99.85	98.29	95.41	0.83	98.97	97.60	93.75	0.75

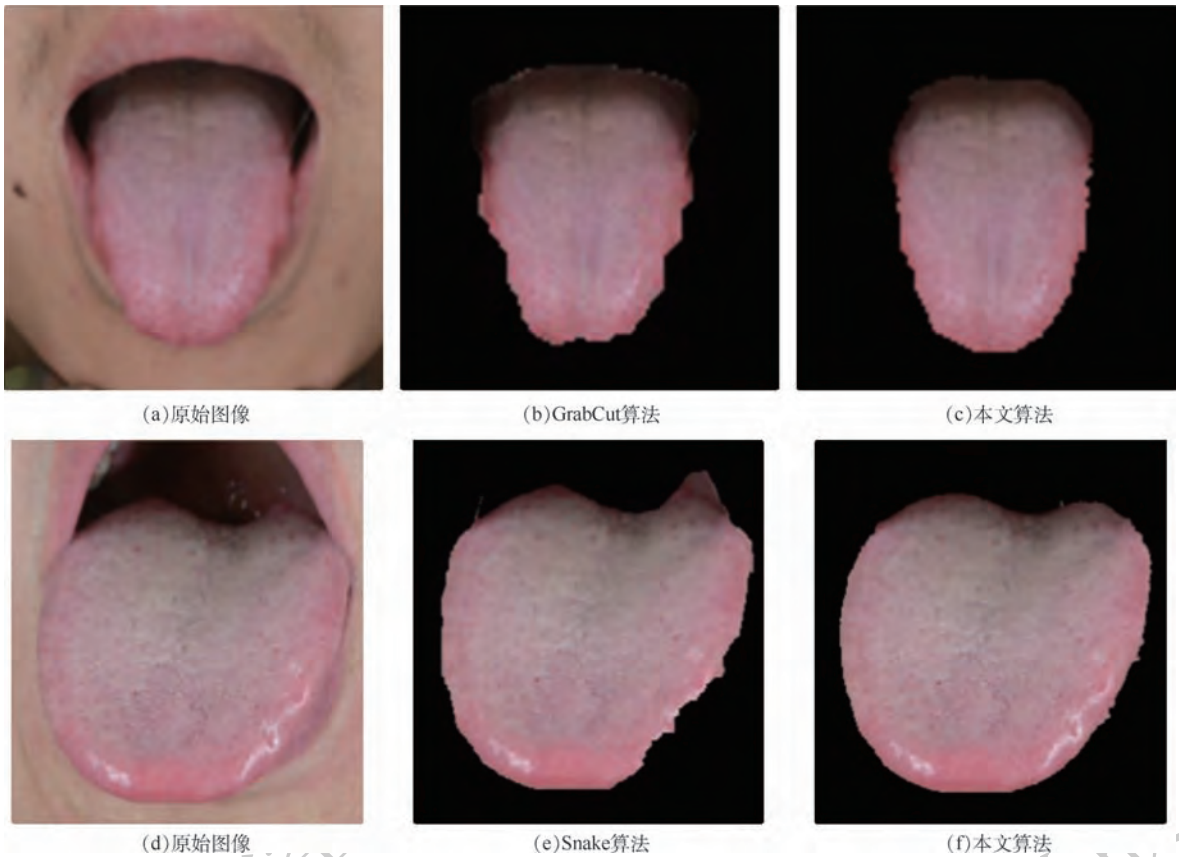


图 8　舌体不完整的分割结果示例

Fig. 8　Examples of segmentation results of tongue defect

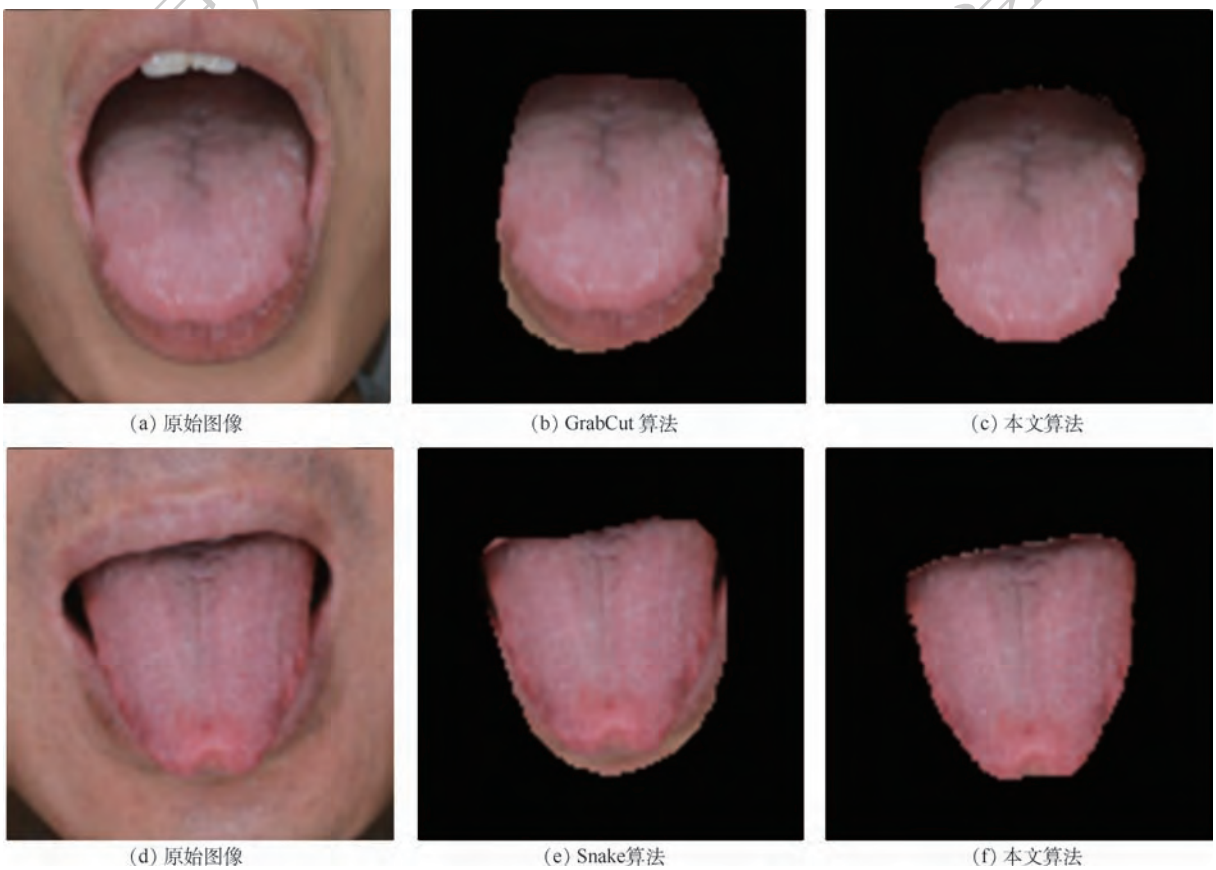


图 9　包含嘴唇部分的舌图像分割结果示例

Fig. 9　Examples of segmentation results of tongue image containing lips

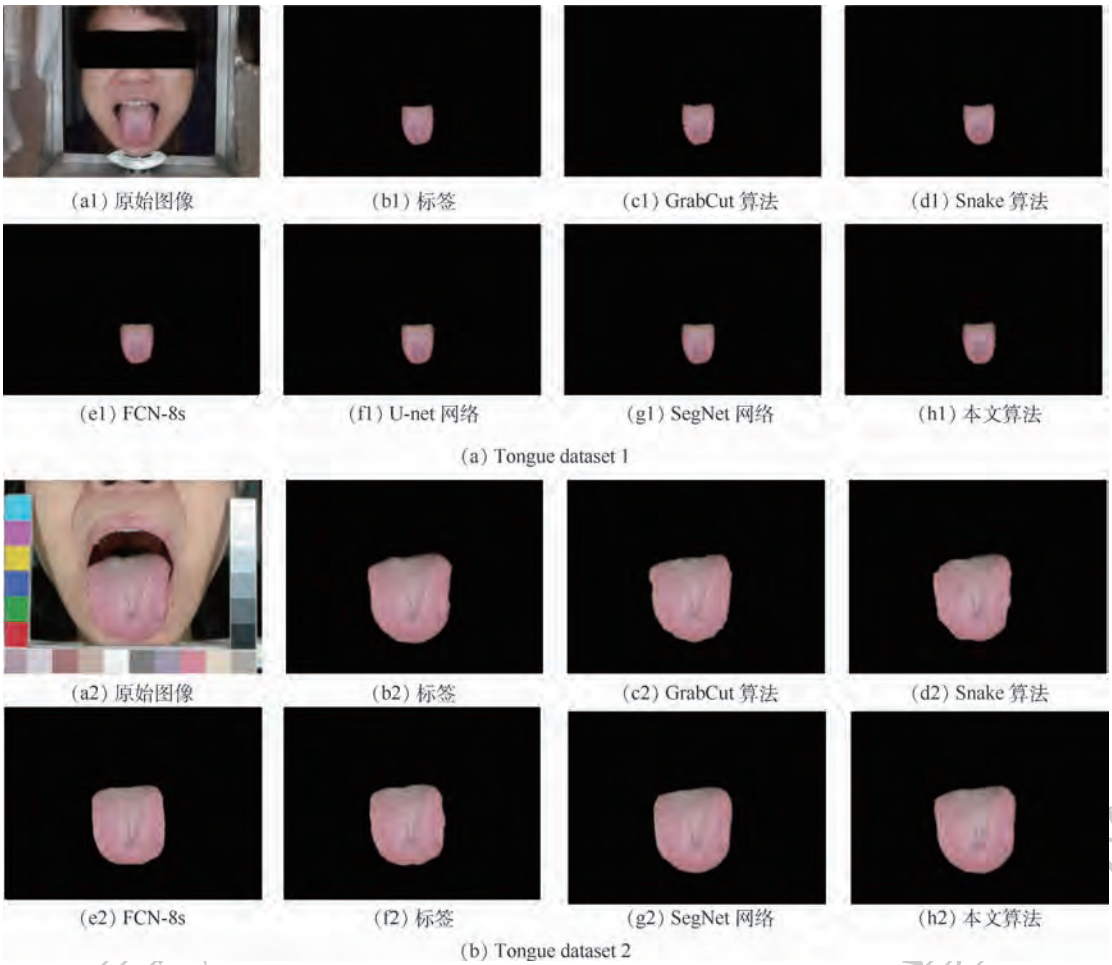


图 10 不同算法的舌图像分割结果示例

Fig. 10 Examples of tongue image segmentation results by different algorithms

3 结 论

针对传统舌图像分割算法需要预处理、人工标定候选区及分割准确率低等问题,本文提出了将 DCNN 与全连接 CRF 相结合的舌图像分割算法。

1) 使用孔卷积算法,使得网络能够在较大的特征图谱尺寸上进行特征提取,保证了边缘信息的提取。通过改变孔卷积参数,使得在不增加网络参数的条件下增大了感受野,保证了语义信息的提取。

2) 使用 ASPP 模块,使网络能够在不同感受野下对图像提取多尺度特征。

3) 使用全连接 CRF,通过计算整幅图像中像素点之间的相关性进行分割,细化分割边缘。

本文提出的舌图像分割网络能够自动提取图像的多尺度特征,对舌体区域进行分割。与传统算法相比,本文算法自动化程度较高,不需要手工标定操作。与其他分割网络相比,本文算法在精度方面有所提高。在接下来的研究中,将把本文算法与最新的 DCNN 相结合,以提高舌图像分割的效率。

致谢 感谢拍摄舌图像的志愿者,以及在制作舌图像数据集过程中对数据进行处理和拍照的同学。

参考文献 (References)

[1] SHEN L S, WANG A M, WEI B G, et al. Image analysis for tongue characterization[J]. Acta Electronica Sinica, 2001, 12(3): 317-323.

[2] 张灵,秦鉴. 基于灰度投影和阈值自动选取的舌像分割方法[J]. 中国组织工程研究与临床康复, 2010, 14(9): 1638-1641.
ZHANG L, QIN J. Tongue-image segmentation based on gray projection and threshold-adaptive method[J]. Journal of Clinical Rehabilitative Tissue Engineering Research, 2010, 14(9): 1638-1641(in Chinese).

[3] 李丹霞,韦玉科. 基于自适应阈值的舌像分割方法[J]. 计算机技术与发展, 2011, 21(9): 63-65.

LI D X, WEI Y K. Tongue image segmentation method based on adaptive thresholds[J]. Computer Technology & Development, 2011, 21(9): 63-65(in Chinese).

[4] KASS M, WITKIN A, TERZOPOULOS D. Snakes: Active, contour models[J]. International Journal of Computer Vision, 1988, 1(4): 321-331.

[5] 傅之成,李晓强,李福凤. 基于径向边缘检测和 Snake 模型的舌像分割[J]. 中国图象图形学报,2019,14(4):688-693.
FU Z C, LI X Q, LI F F. Tongue image segmentation based on snake model and radial edge detection[J]. Journal of Image & Graphics, 2009, 14(4): 688-693(in Chinese).

[6] LI Q L, XUE Y Q, WANG J Y, et al. Automated tongue segmentation algorithm based on hyperspectral image[J]. Journal of Infrared & Millimeter Waves, 2007, 26(1): 77-80.

[7] ROTHER C, KOLMOGOROV V, BLAKE A. GrabCut: Interactive foreground extraction using iterated graph cuts[J]. ACM Transactions on Graphics, 2004, 23(3): 309-314.

[8] 韦玉科,范鹏,曾贵. 改进的 GrabCut 方法在舌诊系统中的应用[J]. 传感器与微系统,2014,33(10):157-160.
WEI Y K, FAN P, ZENG G. Application of improved GrabCut method in tongue diagnosis system[J]. Transducer & Microsystem Technologies, 2014, 33(10): 157-160(in Chinese).

[9] 陈善超,符红光,王颖. 改进的一种图论分割方法在舌像分割中的应用[J]. 计算机工程与应用,2012,48(5):201-203.
CHEN S C, FU H G, WANG Y. Application of improved graph theory image segmentation algorithm in tongue image segmentation[J]. Computer Engineering & Applications, 2012, 48(5): 201-203(in Chinese).

[10] GUO J Y, YANG Y K, WU Q W, et al. Adaptive active contour model based automatic tongue image segmentation[C]// International Congress on Image and Signal Processing, Biomedical Engineering and Informatics, 2017: 1386-1390.

[11] SHI M J, LI G Z, LI F F. C^2G^2FSnake: Automatic tongue image segmentation utilizing prior knowledge[J]. Science China: Information Sciences, 2013, 56(9): 1-14.

[12] LIN B Q, XIE J W, LI C H. Deeptongue: Tongue segmentation via resnet[C]// IEEE International Conference on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE Press, 2018: 1035-1039.

[13] KRIZHEVSKY A, SUTSKEVER I, HINTON G. ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems, 2013: 1097-1105.

[14] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. (2014-09-04)[2019-01-28]. https://arxiv.org/abs/1409.1556.

[15] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 580-587.

[16] GIRSHICK R. Fast R-CNN[C]// IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 15801732.

[17] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[18] SHELHAMER E, JONATHAN L, TREVOR D. Fully convolutional networks for semantic segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(4): 640-651.

[19] NOH H, HONG S, HAN B. Learning deconvolution network for semantic segmentation[C]// IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 1520-1528.

[20] BADRINARAYANAN V, KENDALL A, CIPOLLA R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-2495.

[21] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[22] LI J, XU B C, BAN X J. A tongue image segmentation method based on enhanced HSV convolutional neural network[C]// International Conference on Cooperative Design, Visualization and Engineering. Berlin: Springer, 2017: 252-260.

[23] QU P L, ZHANG H, ZHUO L, et al. Automatic tongue image segmentation for traditional Chinese medicine using deep neural network[C]// Intelligent Computing Theories and Application. Berlin: Springer, 2017: 247-259.

[24] HOSCHNEIDER M, KRONLAND-MARTINET R, MORLET J, et al. A real-time algorithm for signal analysis with the help of the wavelet transform[C]// Wavelets: Time-Frequency Methods Phase Space, 1989: 289-297.

[25] PHILIPP K, KOLTUN V. Efficient inference in fully connected CRFs with Gaussian edge potentials[J]. Advances in Neural Information Processing Systems, 2011, 24(1): 109-117.

[26] HARIHARAN B, ARBELAEZ P, BOURDEV L, et al. Semantic contours from inverse detectors[C]// IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2011: 991-998.

[27] RONNEBERGER O, FISCHER P, BROX T. U-net: Convolutional networks for biomedical image segmentation[C]// International Conference on Medical Image Computing and Computer-Assisted Intervention. Berlin: Springer, 2015: 234-241.

作者简介:

张新峰　男,博士,副教授,硕士生导师。主要研究方向:图像处理、模式识别、机器学习。

郭宇桐　男,硕士研究生。主要研究方向:图像处理、深度学习、语义分割。

Tongue image segmentation algorithm based on deep convolutional neural network and fully conditional random fields

ZHANG Xinfeng*, GUO Yutong, CAI Yiheng, SUN Meng

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: The disadvantage of tongue image segmentation in traditional Chinese medicine are low accuracy, slow segmentation speed and manual calibration of candidate regions. To solve these problems, we propose an end-to-end tongue image segmentation algorithm. Compared with the traditional tongue segmentation algorithm, more accurate segmentation results can be obtained by the proposed method which does not need any manual operation. Firstly, the atrous convolution algorithm is used to increase the feature map of the network without increasing the parameters. Secondly, the atrous spatial pyramid pooling (ASPP) module is used to enable the network to learn the multi-scale feature of the tongue image through different receptive fields. Finally, the deep convolutional neural networks (DCNN) are combined with fully connected conditional random fields (CRF) to refine the edge of the segmented tongue image. The experimental results show that the proposed method outperforms traditional tongue image segmentation algorithm and popular DCNN with higher segmentation accuracy, and the mean intersection over union reaches 95.41%.

Keywords: deep learning; convolutional neural network (CNN); semantic segmentation; tongue images; conditional random fields (CRF)

Received: 2019-07-09; **Accepted**: 2019-08-14; **Published online**: 2019-08-27 09:40

URL: kns.cnki.net/kcms/detail/11.2625.V.20190826.1424.002.html

Foundation item: National Key R&D Program of China (2017YFC1703300)

* **Corresponding author**. E-mail: zxf@bjut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0387

基于多源图像弱监督学习的3D人体姿态估计

蔡轶珩*, 王雪艳, 胡绍斌, 刘嘉琦
(北京工业大学 信息学部, 北京 100124)

摘 要: 3D人体姿态估计是计算机视觉领域一大研究热点,针对深度图像缺乏深度标签,以及因姿态单一造成的模型泛化能力不高的问题,创新性地提出了基于多源图像弱监督学习的3D人体姿态估计方法。首先,利用多源图像融合训练的方法,提高模型的泛化能力;然后,提出弱监督学习方法解决标签不足的问题;最后,为了提高姿态估计的效果,改进了残差模块的设计。实验结果表明:改善的网络结构在训练时间下降约28%的情况下,准确率提高0.2%,并且所提方法不管是在深度图像还是彩色图像上,均达到了较好的估计结果。

关 键 词: 人体姿态估计; 沙漏网络; 弱监督; 多源图像; 深度图像
中图分类号: TP391.4
文献标识码: A **文章编号**: 1001-5965(2019)12-2375-10

基于图像的人体姿态估计是指获得给定图像中人体各部位相对位置信息的过程,可广泛用于视频监控、行为识别及人机交互等多方面领域[1-3]。

目前,使用单一的彩色图像或深度图像进行人体姿态估计均已取得了一定的成果[4-8]。相对来说,由于彩色图像更易获得,所以针对单一彩色图像的人体姿态估计的研究更为广泛[9-13],可供利用的公开数据集也更为充足,如用于2D人体姿态估计研究的LSP[14]和MPII[15],以及用于3D人体姿态估计的Human 3.6M[16]等。而对于深度图像,由于其记录的是深度相机到目标人体之间的距离信息,不包含颜色及纹理细节等信息,因此,基于深度图像的3D人体姿态估计方法,一方面不易因人体着装、肤色和光照等复杂外界环境的变化而受到影响,另一方面使用该图像在保护用户隐私方面也具有很好的优势。但由于深度相机对光照、背景等较为敏感,深度图像获取的条件较为严苛。现有的深度图像数据集一般是在实验室环境下拍摄获得的,其姿态变化有限。而关节点标签基本采用先相机标定后人工检错的方式获得[8]。由于人工检错仍存在随机性等问题,因而很少有公开的深度图像数据集可以提供充足且准确的3D关节点标签。而为获得较为准确的深度图像3D关节点标签,需要研究者准备训练样本及标签,使得研究成本增加,同时也限制了深度图像在3D姿态估计领域的研究进程。因此,对现有缺乏准确深度标签的深度图像数据集进行研究,提出可行的算法,实现对深度图像的3D人体姿态估计是值得探索与鼓励的。

为此,本文提出了一种端到端的多源图像弱监督学习方法。该方法利用多源图像融合训练的方法解决深度图像姿态单一引起的模型泛化能力不高的问题,同时使用弱监督学习技术来解决标签不足的问题,并对网络中的残差模块进行改进,提高姿态估计的准确率。

收稿日期: 2019-07-10; **录用日期**: 2019-08-23; **网络出版时间**: 2019-08-29 13:16
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190828.1803.001.html
基金项目: 国家重点研发计划(2017YFC1703302); 北京市教委科技项目(KM201710005028)
* **通信作者**. E-mail: caiyiheng@bjut.edu.cn

引用格式: 蔡轶珩, 王雪艳, 胡绍斌, 等. 基于多源图像弱监督学习的3D人体姿态估计[J]. 北京航空航天大学学报, 2019, 45(12): 2375-2384. CAI Y H, WANG X Y, HU S B, et al. Three-dimensional human pose estimation based on multi-source image weakly-supervised learning[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2375-2384 (in Chinese).

1 相关方法

在深度学习领域,基于图像的 2D 或 3D 人体姿态估计均已取得一定的成果。其中,对于单一深度图像来说,在 2D 人体姿态估计上,文献[3]采用 MatchNet[17]计算全卷积网络(Fully Convolutional Networks,FCN)预测的关节区域和模板之间相似度的方法,并结合相邻关节之间的配置关系,来达到优化关节点位置的目的。文献[18]介绍了一种基于模型的递归匹配(MRM)人体姿态的新方法,先对深度图像进行预处理以获得个性化参数,再使用模板匹配和线性拟合来估计人体骨架信息。在 3D 人体姿态估计上,文献[8]采用长短期记忆网络架构(Long Short-Term Memory,LSTM),学习局部视点不变特征,并利用自顶向下的错误反馈机制,纠正姿态位置,从而获得良好的 3D 人体姿态估计,但该方法采用强监督学习的方式完成 3D 人体姿态估计,其训练样本及标签为研究者自行准备,研究成本较高,同时也存在训练样本关节点标注不准的问题。

对于彩色图像的研究,在 2D 人体姿态估计上,文献[11]提出卷积姿态机器的方法,利用多阶段联合训练的方式,充分学习图像中的特征信息,来提高网络的姿态回归结果。文献[12]则提出了沙漏网络结构,利用多尺度特征来识别姿态,从而提高估计姿态的准确性。而对于 3D 人体姿态估计的研究方法,文献[19]提出强监督学习技术训练 3D 回归模型的方法。文献[1,13,20-23]采用了 2D 人体姿态估计结果辅助 3D 回归模型训练的方法。其中,文献[1]证明了该方法相较于直接训练 3D 回归模型,姿态估计准确率更高,而文献[21]则介绍了一种分别学习 2D 回归模型和深度回归模型的网络框架;不同于文献[1,20-23]分阶段分别训练 3D 回归模型的方法,文献[13]针对室外人体姿态数据库缺乏深度标签的问题,提出一种端到端联合训练 2D 模型和深度模型的网络结构,该方法充分利用了实验室环境下充足且准确的标签数据及室外环境下的复杂人体姿态信息,通过该弱监督学习技术,以端到端的方式,获得较好的室外图像 3D 人体姿态估计。最近,文献[24]提出一个 Graph-CNN 网络,在网络中使用 SMPL 模板网格来回归人体姿态。文献[25]提出了一种基于单目图像的 3D 人体姿态估计的全卷积网络,使用肢体方向作为一种新的 3D 表示方法。

从上述研究可以发现,研究者基本采用强监督学习技术来完成对图像的 3D 人体姿态估计,利用充足的 3D 关节点标注信息来辅助回归模型的训练。但当训练样本中缺乏标签时,上述强监督学习方法不再适用,而弱监督学习技术的优点便显露出来。基于弱监督学习的网络模型不要求训练样本提供充足的标签,即可完成对回归模型的训练,可有效解决本文深度图像缺乏深度标签的问题。因此,基于上述研究背景,受文献[13]的启发,本文提出了一个基于多源图像弱监督学习的 3D 人体姿态估计方法。该方法使用多源图像作为训练样本,利用彩色图像姿态多变且 3D 标签充足的特点,来弥补深度图像姿态单一和缺乏深度标签的问题;同时为提高姿态估计准确性,还对现有的残差模块进行改善设计,从而实现对深度图像的 3D 人体姿态估计。

2 多源图像弱监督学习方法

面对深度图像缺乏准确深度标签的问题,可利用实验室环境下获取充足的彩色图像及其准确的人体运动关节点深度信息,辅助深度图像学习到相应的人体关节点深度信息,以实现对深度图像的 3D 人体姿态估计。基于上述研究思想,本文提出了一个基于多源图像端到端弱监督的 3D 人体姿态估计框架,通过多源图像混合训练的方式,完成对缺乏标注的深度图像 3D 回归模型训练任务。

2.1 多源图像 3D 人体姿态估计框架

本文基于多源图像弱监督学习的整体架构如图 1 所示。训练样本由多源图像构成,包含带 2D 标签的深度图像和彩色图像,以及带 3D 标签的彩色图像。网络结构分为 2D 回归子网络模块和深度回归子网络模块两部分。

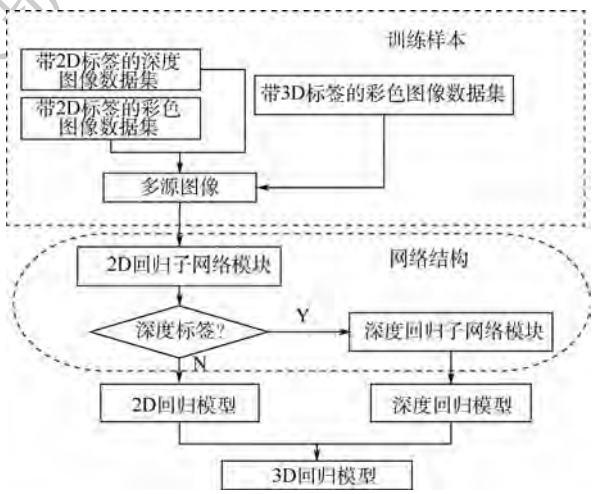


图 1 3D 人体姿态估计整体框架

Fig. 1 Overall framework of 3D human pose estimation

整个框架的具体训练流程如下:①将多源图像训练样本作为网络的输入;②利用所有带 2D 标签的训练样本训练 2D 回归子网络模块,得到 2D 回归模型;③带 3D 标签的彩色图像经过 2D 回归子网络模块输出热图特征,将其作为深度回归子网络模块的输入进行训练,得到深度回归模型;④将 2 个回归模型的结果进行连接,从而完成对 3D 回归模型的训练任务。

本文的网络结构分为 2D 回归子网络模块和深度回归子网络模块两部分,具体结构如图 2 所示。其中,2D 回归子网络模块由 2 个沙漏模块构成,通过重复使用自顶向下和自底向上的方式对 2D 关节点坐标位置进行推导,在每个沙漏模块后均使用了热图对关节点坐标进行预测,即在网络结构中引入了中继监督技术,可有效避免训练过程中,由于网络层数过深而导致的梯度消失问题,加快网络模型收敛速度;同时由于热图中包含了关节点之间的相互关系,因此,将热图预测结果作为下一个沙漏模块的输入特征继续训练,有助于提高整体网络结构的回归性能。而深度回归子网络模块则采用文献[13]网络设计,由残差模块、池化层及线性回归器构成,紧接在 2D 回归子网络模块的后面,使其可利用 2D 回归子网络模块中充分学习到的特征作为输入进行训练,同时由于 2D 回归子网络模块的输出特征中包含关节点热图结果,因而使得该模块也可充分利用热图中关节点相互关系,有助于在弱监督学习下获得更为准确的关节点深度值。

在网络测试阶段,将测试图像输入到本文网络中,2D 回归模型输出各关节点的预测热图,即 2D 关节点坐标 $\hat{Y}_{2d}$,而深度回归模型则对上述关节点的热图进行回归,用于预测出关节点的深度值 $\hat{Y}_{depth}$,对 2 个模型回归出来的结果进行连接,即可完成对测试图像的 3D 人体姿态估计。

为改善现有网络结构的关节点回归性能,本文对上述网络结构提出改进设计,使得本文方法可以在提高回归模型准确度的同时,降低网络的训练时间及存储空间。

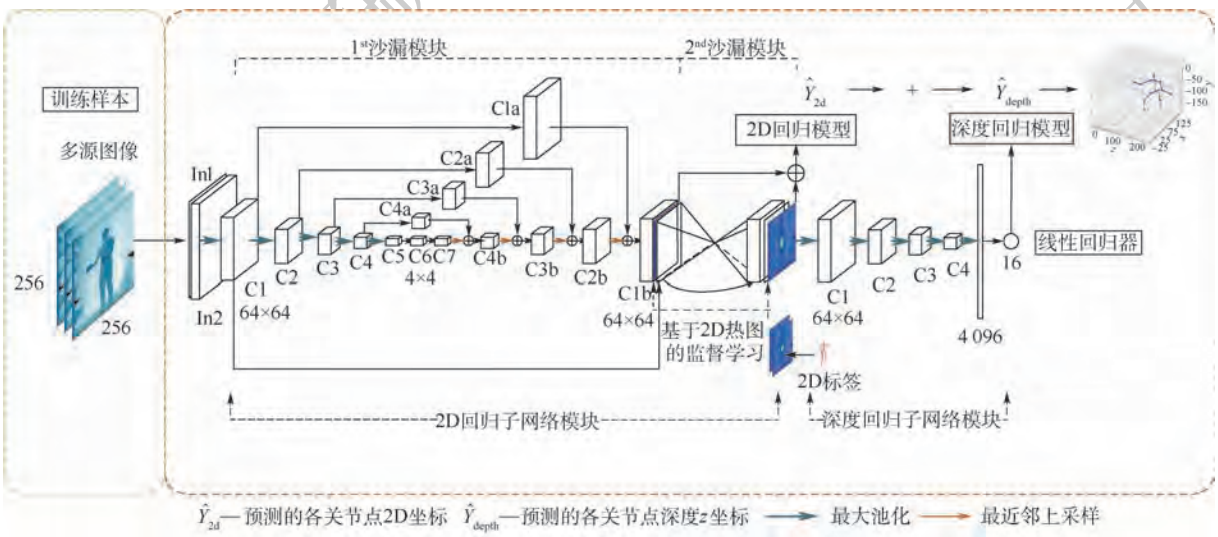


图 2　基于弱监督学习网络结构框架

Fig. 2　Network structure framework based on weakly-supervised learning

2.2　残差模块设计

一个较好的回归网络结构能在较少的训练时间内获得较优的关节点回归精度。但模型的训练时间及回归准确度与卷积网络的构成有很强的关联性。若适当地加深网络深度及特征维度,虽可获得较好的回归精度,但网络参数也大幅增加,同时也会加大模型的存储空间及训练时间;而若简单的降低网络深度及特征维度,虽可降低训练时间,但模型性能则会随之下降。因此,本文针对上述问题,为提高网络回归模型的准确度,同时降低模型训练时间,对网络结构的残差模块进行了改善。

图 2 为本文基于弱监督学习的 3D 人体姿态估计网络结构框架,其中每个矩形块(C1, …, C4, C1a, …, C4a, C1b, …, C4b)均表示的是 2 个残差模块,因而可以说本文网络结构基本是由残差模块构成的。而现有的残差模块(见图 3(a)),其输入和输出特征维度均为 256,通过交叉使用 1×1、3×3 和 1×1 的卷积进行充分的特征提取,并通过 Shortcut 连接,将卷积之后的特征和原始输入特征进行融合,使得残差模块可在提取较高层次特征的同时,又保留了原有层

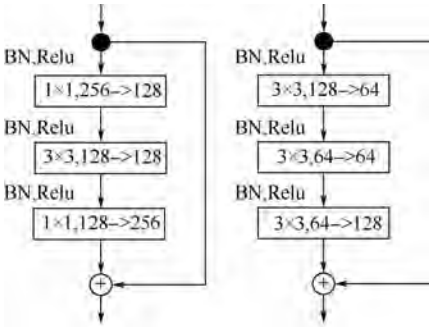


图 3 残差模块

Fig. 3 Residual module

次的信息,这一多尺度特征信息在精准人体关节点预测方面,提供了较好的帮助[12-13]。但较高的特征维度也引起了训练时间变长,因此本文降低了残差模块的输入维度,从 256 降为 128,降低特征维度后,1×1 卷积的特征重组效果则会大大降低,因而,本文将 3×3 的卷积替换了 1×1 卷积,使得网络可对输入特征进一步提取,从而弥补特征维度降低造成的性能损失,甚至提高网络的回归精度,本文改进的残差模块如图 3(b)所示。

2.3 3D 人体姿态估计

2.3.1 2D 回归子网络模块

本文利用沙漏网络可提取多尺度特征的特点,采用沙漏网络作为 2D 回归子网络模块训练 2D 回归模型,用于预测人体各关节点的位置坐标,以实现对图像的 2D 人体姿态估计。由于深度图像是在实验室环境下获取的,姿态单一且有限,因此,为提高 2D 回归模型的泛化能力,本文提出同时使用深度图像和彩色图像的混合多源图像的方式来训练 2D 回归模型。即输入数据为带 2D 标签的多源图像,输出一系列 $J(J=16)$ 的低分辨率的关节点热图。

由于深度图像记录的是目标距离相机的距离信息,不包含颜色及纹理信息,若直接将深度图像和彩色图像混合作为网络的输入,会对模型训练造成干扰。因此,需对图像做预处理。考虑到深度图像在视觉上也可看做是灰度图像,因此使用加权平均法将训练所需的彩色图像进行灰度处理,去除里面的颜色干扰信息,减少由于训练样本变化而引起的模型精度损失,提高模型的回归精度。

沙漏网络训练的输入为上述预处理后的所有带 2D 标签的混合多源图像,图像分辨率为 256×256,输出为预测到的各关节点的热图,图像分辨率为 64×64,其关节点坐标为热图中概率最高的点。2D 估计效果及其对应热图结果如图 4 所示,

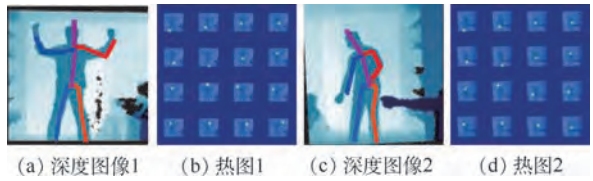


图 4 ITOP 数据集的 2D 人体姿态估计及其对应的热图结果

Fig. 4 Two-dimensional human pose estimation and corresponding heat-map results in ITOP dataset

(a)、(c)为深度图像,(b)、(d)为预测的热图结果,从左到右,从上到下依次为:右脚踝、右膝盖、右胯、左胯、左膝盖、左脚踝、臀部、胸部、脖子、头、右手腕、右手肘、右肩膀、左肩膀、左手肘、左手腕,共 16 个关节点热图,在热图中概率最高,也就是亮度最高的点即为预测的该关节点坐标位置。

本文 2D 回归模型训练的 loss 函数使用 L_2 距离[13],其公式如下:

$$L_{2d}(\hat{Y}_{HM}, Y_{2d}) = \sum_{h=1}^{H}\sum_{w=1}^{W}(\hat{Y}_{HM}^{(h,w)} - G(Y_{2d})^{(h,w)})^2 \tag{1}$$

式中:$\hat{Y}_{HM}$ 为预测的各关节点的热图;$G(Y_{2d})$ 为各关节点标签通过高斯核[12]计算后的热图。

2.3.2 深度回归子网络模块

此阶段的主要目标是获得人体各关节点的深度值,而现有的针对无深度标签的数据,一般是采用模板匹配的方式预测关节点深度值。但这种方法未考虑图像中人体信息在深度值预测的重要性。

本文在一个网络结构中构建了 2 个回归子网络,并将深度回归子网络模块接在 2D 回归子网络模块的后面,将 2D 回归子网络模块中学习到的包含语义信息及多尺度信息的输出特征作为输入继续训练,可有效利用端到端网络训练的优势,充分利用权重共享功能从而获得更好的姿态估计结果。

深度回归网络训练收敛的 loss 函数使用 L_2 距离,其公式如下:

$$L_{depth}(\hat{Y}_{depth}, Y_{depth}) = \alpha\sum_{i=1}^{J}\|\hat{Y}_{depth}^{(i)} - Y_{depth}^{(i)}\|^2 \tag{2}$$

式中:$\hat{Y}_{depth}$ 为预测的各关节点深度值;Y_{depth} 为各关节点深度值的标签;α 用于控制是否训练深度模型。

3 实验及结果分析

在本节中,为探讨本文弱监督学习姿态估计方法的预测性能,分别在深度图像数据集 ITOP[8] 和 K2HGD[3]、彩色图像数据集 MPII[15] 和 Human

3.6M[16]上进行训练及测试,并与相关姿态估计模型进行对比[13],用以评估本文方法的性能。

3.1 数据库

3.1.1 深度图像数据库

ITOP[8]是由 20 个人各做 15 个动作序列拍摄而成的,包含侧拍和顶拍 2 个视角的图像,其标签使用 Kinect 自带的 SDK 预测,虽然已通过人工检测的方式检错,但 3D 标签标定仍存在较大误差。因此本文仅使用经过前期标签检错筛查后的侧拍图像数据库进行实验,其中训练样本中仅使用了提供的 2D 关节点标签,约 11 000 张,而测试图像则使用了 3D 标签纠正后的图像数据,约 2 979 张,用于判断本文回归模型 3D 关节点的预测性能。

K2HGD[3]由 30 个人拍摄获得,共有 10 万张深度图像,提供相应的 2D 关节点标签。本文使用其中约 6 万张作为训练图像。

3.1.2 彩色图像数据库

MPII[15]是一个大型室外姿态估计数据库,提供相应的 2D 关节点标签。本文使用约 25 000 张图像进行训练。

Human 3.6M[16]由 11 个人各做 17 组动作,由 4 个角度上拍摄获得,共包含有 360 万张带 3D 标签的彩色图像,本文使用其中 30 万张图像作为训练图像,2 874 张图像作为测试图像。

3.2 评价标准

为评估回归的关节点坐标准确性,本文使用 PDJ (Percentage of Detected Joints)[3]作为评定标准,若关节预测坐标与标签之间的误差与归一化躯干长度的比值在一定阈值内,便可将其判定预测正确。使用阈值不同,检测到的关节点准确率也不同。

3.3 训练细节

本文训练平台为 Torch7[22],并基于公开代码[12-13]构建本文 2D 回归子网络模块及深度回归子网络模块,如图 2 所示。输入图像分辨率为 256×256,2D 回归子网络模块的输出为预测的人体各关节点的热图,分辨率为 64×64,其热图概率值最高的点,作为此关节点的 2D 坐标预测结果,同时深度值由深度回归模型输出获得。

为达到快速训练的目的,本文网络结构的主体由 2 个沙漏模块串联而成。在训练时,采用的学习率为 2.5×10^{-4},mini-batch 的尺寸为 6。为获得更好的 3D 回归模型准确率,本文分 2 个阶段训练 3D 回归网络,每个阶段均迭代了 28 万 batch[13]。第 1 阶段,利用混合多源图像仅训练 2D 回归模型,第 2 阶段则以端到端的方式,训练 3D 回归模型。其中,2D 回归模块的参数采用第 1 阶段的 2D 回归模型的权重进行初始化,在继续训练 2D 回归模型的同时,利用带深度标签的彩色图像更新深度回归子网络模块的权重参数,从而训练获得更好的 3D 回归模型。

由于存在训练图像中包含多个目标人体的现象,因此,本文在训练前,首先将样本进行预处理,对于每张训练及测试样本,均以人体臀部为中心进行裁剪,将目标人体放在图像的中间,其裁剪尺寸比在 1.3~1.7 之间,并归一化图像大小分辨率为 256×256,同时对图像做加权平均的灰度处理,尽量保证训练图像的一致性。为提高模型的泛化能力,本文对数据进行了扩充处理,即对样本进行左右翻转及旋转处理,旋转角度在 -6°~6° 之间随机选择。本文训练及测试样本的标签统一为头、脖子、左右肩、左右肘、左右手腕、左右胯、左右膝盖、左右脚踝、胸部及臀部共 16 个关节点。

3.4 结果对比

由于使用的网络结构及训练数据不同,本文共获得的模型如表 1 所示。其中,M-H36M、I-H36M、IK-H36M 及 IKM-H36M 模型均是在本文改善后的网络结构上,通过不同的多源图像组合方式训练获得的,用于探讨本文所提使用多源图像混合训练的方式,对 3D 回归模型准确率的影响。其中,M-H36M 训练样本同文献[13],即以 MPII 和 Human 3.6M 作为训练样本,而网络结构中的残差模块则使用了本文所提的改善设计(见图 3(b)),用于探讨本文改善后的网络结构在训练准确率及训练时间上的优越性能。

表 1 不同模型对应的训练图像

Table 1 Training images corresponding to different models

模型	训练数据			
	深度图像数据库		彩色图像数据库	
	ITOP	K2HGD	MPII	Human 3.6M
文献[13]			√	√
M-H36M(本文)			√	√
I-H36M(本文)	√			√
IK-H36M(本文)	√	√		√
IKM-H36M(本文)	√	√	√	√

3.4.1 网络性能对比

为验证本文残差模块改进方案对关节点位置回归精度的影响,分别使用不同的残差模块网络结构在相同条件下进行实验,其对比结果如表 2 所示。其中文献[13]的 131~256 表示残差模块(见图 3(a))对应的卷积依次为 1×1、3×3 和 1×1,输入和输出通道数为 256,表内其他方法数

据物理含义同上。与文献[13]结果相比,本文模型(131 ~ 128)的准确率最低,降低约 2.16% 左右;而本文模型(333 ~ 256)的准确率最高,提升约0.66% 左右,但其参数量和训练所需时间均成倍增加。这说明在保持残差模块卷积核大小不变的情况下,仅是简单地将输入输出通道数降低,其准确率会有所降低;而在保持输入输出特征维度不变的情况下,将卷积核大小放大,提高模型感受野,虽能提高模型的回归精度,但其参数量和训练所需时间也大幅增加。故本文选用 333 ~ 128 的残差模块改善方案(见图 3(b))以获得更优的回归性能。该方法可在减少模型参数的同时提高训练准确率,并且训练一个 batch 的时间与原始沙漏网络相比,下降了约 28%。因此,实验表明,本文改善残差模块后的网络结构可在有效降低训练时间的同时提高模型准确率。

本文还验证了沙漏模块的数量对回归模型准确率的影响,如表 3 所示。其中,本文模型(4 stack)代表的是将 2D 回归子网络模块中的沙漏模块增加至4 个,并基于本文改善后的残差模块结构训练获得的 3D 回归模型,与使用 2 个沙漏模块的本文模型(2 stack)相比,回归准确率提高了约 0.35% 左右,但每个 batch 的训练时间会增加约 52%,即每个 epoch 训练周期会增加近一半的训练时间。实验表明,增加网络层数可进一步提升回归模型的准确率,但其训练时间和模型参数量会大幅增加,因此,为达到快速训练的目的,将使用本文模型(2 stack)对应的网络结构进行实验。

表 2 不同模型准确率、参数量及训练时间对比

Table 2 Comparison of accuracy rate, parameter quantity and training time among different models

模型	准确率/%	参数量/10^6	训练时间/(s·batch^{-1})
本文模型(131 ~ 128)	90.10	31.00	0.16
文献[13](131 ~ 256)	92.26	116.60	0.29
本文模型(333 ~ 128)	92.48	101.10	0.21
本文模型(333 ~ 256)	92.92	390.00	0.41

表 3 不同沙漏网络个数准确率、参数量及训练时间对比

Table 3 Comparison of accuracy rate, parameter quantity and training time with different numbers of hourglass network

模型	准确率/%	参数量/10^6	训练时间/(s·batch^{-1})
文献[13](2 stack)	92.26	116.60	0.29
本文模型(2 stack)	92.48	101.10	0.21
本文模型(4 stack)	92.83	185.30	0.32

3.4.2 深度图像 3D 人体姿态估计的模型对比

为验证不同训练数据库对回归模型性能的影响,本文基于 PDJ 评判标准,使用本文基于不同训练数据获得的 3D 回归模型,分别在 ITOP 深度图像数据集的测试图像上进行人体姿态估计(该测试图像标签已经过人工纠正)。将测试结果进行对比,探讨最优的基于弱监督学习的 3D 回归模型。

使用本文不同数据集训练得到的回归模型,在 ITOP 深度图像数据集测试图像手腕和膝盖 3D 关节点的预测结果对比,如图 5 所示。可以看出,本文 IKM-H36M 模型的性能最优,IK-H36M 性能次之,说明对深度图像的 3D 人体姿态估计任务中,随着深度图像训练样本的增多,其回归模型在大部分关节点的预测精度也会逐步提高。并且从 IKM-H36M 和 IK-H36M 曲线对比可以看出,在训练样本中引入带 2D 标签的彩色图像数据 MPII,可进一步提高模型的预测准确率,验证了本文所提使用多源图像进行混合训练的方法,可有效提高模型的关节点回归精度。

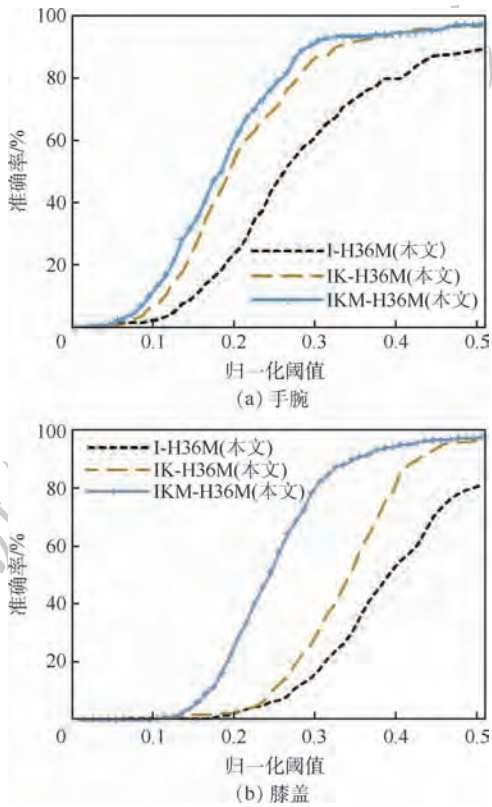


图 5 基于 PDJ 评价指标,不同训练模型在 ITOP 数据集手腕和膝盖 3D 关节点的准确率

Fig. 5 Three-dimensional articulation point accurary rate of wrist and knee using different training models based on PDJ evaluation criteria in ITOP database

3.4.3 彩色图像 3D 人体姿态估计的模型对比

本节测试了本文方法针对彩色图像的 3D 人体姿态估计效果。使用本文不同数据集训练得到

的回归模型,在 Human 3.6M 彩色图像数据集测试图像上脚踝和膝盖 3D 关节点的预测结果对比,如图 6 所示。可以看出,各回归模型的检测性能相近,说明利用多源图像混合训练的回归模型,虽在训练样本中引入了深度图像,但并不会对彩色图像上的 3D 人体姿态估计精度造成太大的影响。图6中画圈部分为各回归模型检测精度

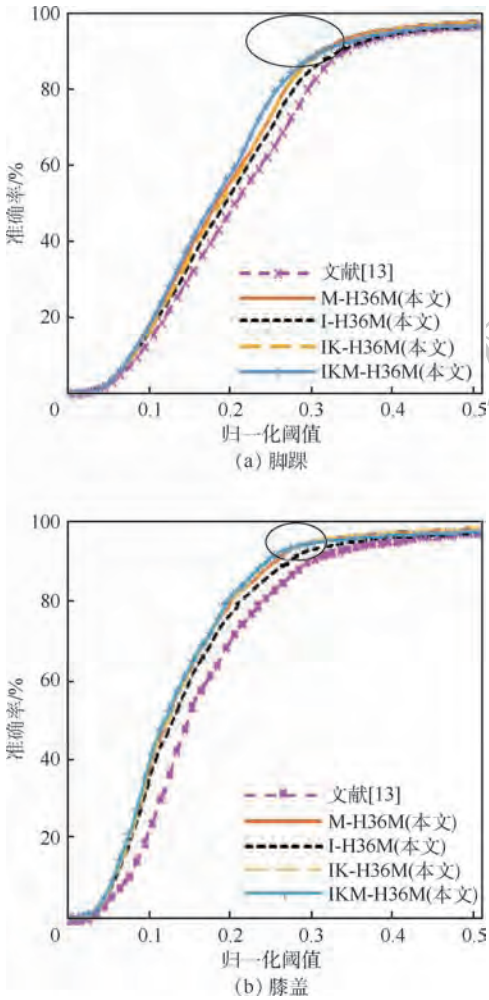


图 6　基于 PDJ 评价指标,不同训练模型在 Human 3.6M 数据集脚踝和膝盖 3D 关节点的准确率

Fig. 6　Three-dimensional articulation point accurary rate of ankle and knee using different training models based on PDJ evaluation criteria in Human 3.6M database

提升由快到慢转变的区域,其中在归一化阈值 0.25处为检测精度变化转折点,意味着在该归一化阈值之后,回归模型的检测精度即将趋于平稳,此时各回归模型已能将测试样本中绝大部分关节点正确定位。因此,为更清楚地看到各关节点在不同模型的检测差别,比较了各关节点在归一化阈值 0.25 处的准确率,如表 4 所示。可以看出,使用本文改善残差模块后的网络 M-H36M,对彩色图像的预测性能最优,相比文献[13]的预测精度提升了约 4%,而基于多源图像混合训练获得的模型,其平均检测精度由高到低,分别为 IKM-H36M > IK-H36M > I-H36M,这也说明了使用多源图像进行训练回归模型,训练的数据越多,其检测精度越高,同时这 3 个模型的平均检测精度均高于文献[13]方法,这也又一次证明了本文改善后的网络结构有助于提高姿态估计准确性。而从 M-H36M 和 IKM-36M 平均检测结果比较来看,使用多源图像训练获得的 IKM-H36M 模型,平均检测性能略低于 M-H36M 模型,检测准确度下降约 0.50%,这是因为在训练样本中,除彩色图像外,还引入了深度图像,即等于引入了干扰项,使得模型的回归性能略有下降。但从下降 0.50% 的结果上来看,使用多源图像训练的 3D 回归模型,虽然在关节点检测精度上具有轻微下降,但并不影响测试图像在各关节点的总体回归性能。

3.4.4　可视化 3D 估计结果

为更直观地看到本文模型在深度图像和彩色图像上的 3D 估计结果,本文可视化了使用 IKM-H36M 模型分别在 ITOP 和 Human 3.6M 测试图像上的姿态估计图,如图 7 和图 8 所示。其中每幅估计图中均包含测试图像、groundtruth 及本文估计结果 3 部分,(b)和(e)为 groundtruth 姿态效果,(c)和(f)为本文模型估计效果。

从图 7 和图 8 中可以看出,使用本文弱监督学习方法的 IKM-H36M 模型可对深度图像和彩色图像预测其相应的3D人体姿态,并且预测结

表 4　基于 PDJ 评价指标,不同训练模型在 Human 3.6M 测试图像上的 3D 人体姿态估计结果

Table 4　Three-dimensional pose estimation results of different regression models on Human 3.6M test images base on based on PDJ evaluation criteria

方法	准确率/%							
	脚踝	膝盖	髋	手腕	手肘	肩膀	头	平均
文献[13]	65.00	82.90	89.27	77.91	86.46	94.73	94.42	84.38
M-H36M(本文)	75.07	90.33	94.69	80.92	86.24	96.38	94.69	88.33
I-H36M(本文)	71.52	88.03	92.52	76.46	82.55	94.07	95.30	85.78
IK-H36M(本文)	79.23	91.60	94.40	77.61	81.40	92.69	93.34	87.18
IKM-H36M(本文)	74.98	91.06	94.33	78.65	85.21	95.77	95.16	87.88

果也较为接近 groundtruth 姿态。图 7 为本文针对 ITOP 深度图像数据集上进行的 3D 人体姿态估计效果图,可以看出,即使是对较为复杂的自遮挡人体侧视图,也可获得较好的 3D 人体姿态估计。由于人体下肢的自由度比上肢自由度大,使得该方法在膝盖和脚踝处的深度预测结果不如上肢预测效果理想,但本文方法也对无深度标签的深度图像实现 3D 人体姿态估计提供了可能。同时从图 8 结果图中可看出,本文模型对彩色图像同样可实现较为理想的 3D 人体姿态估计。

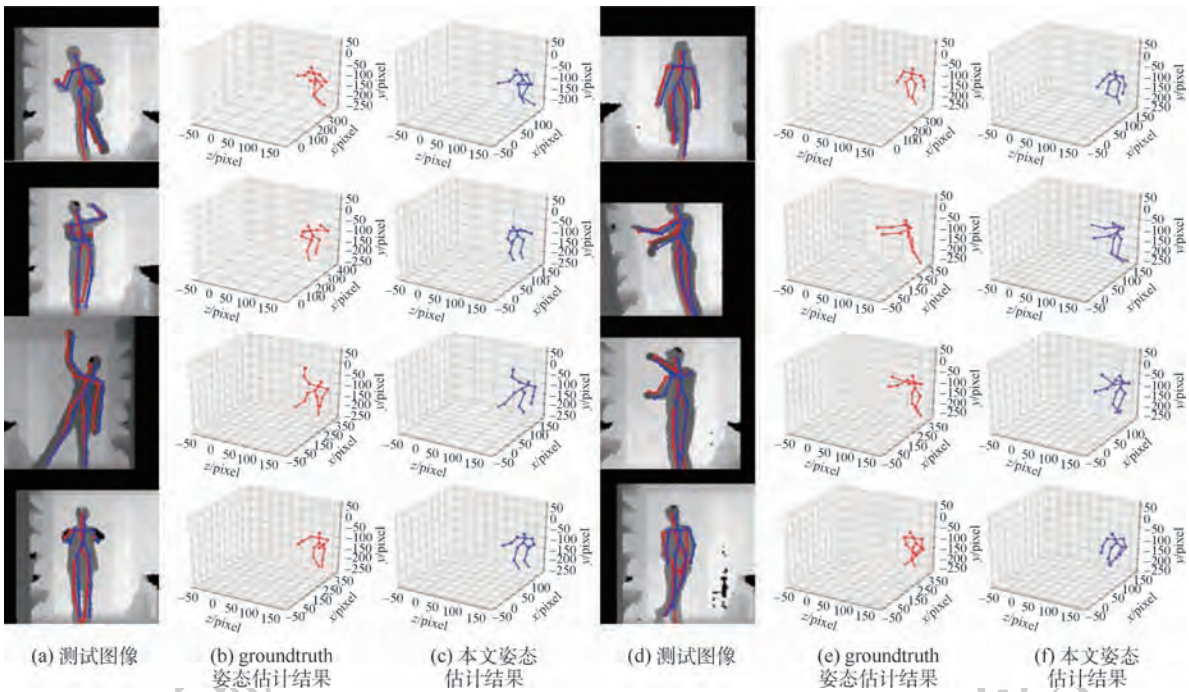


图 7 ITOP 数据集上的 3D 人体姿态估计

Fig. 7 Three-dimensional human pose estimation on ITOP dataset

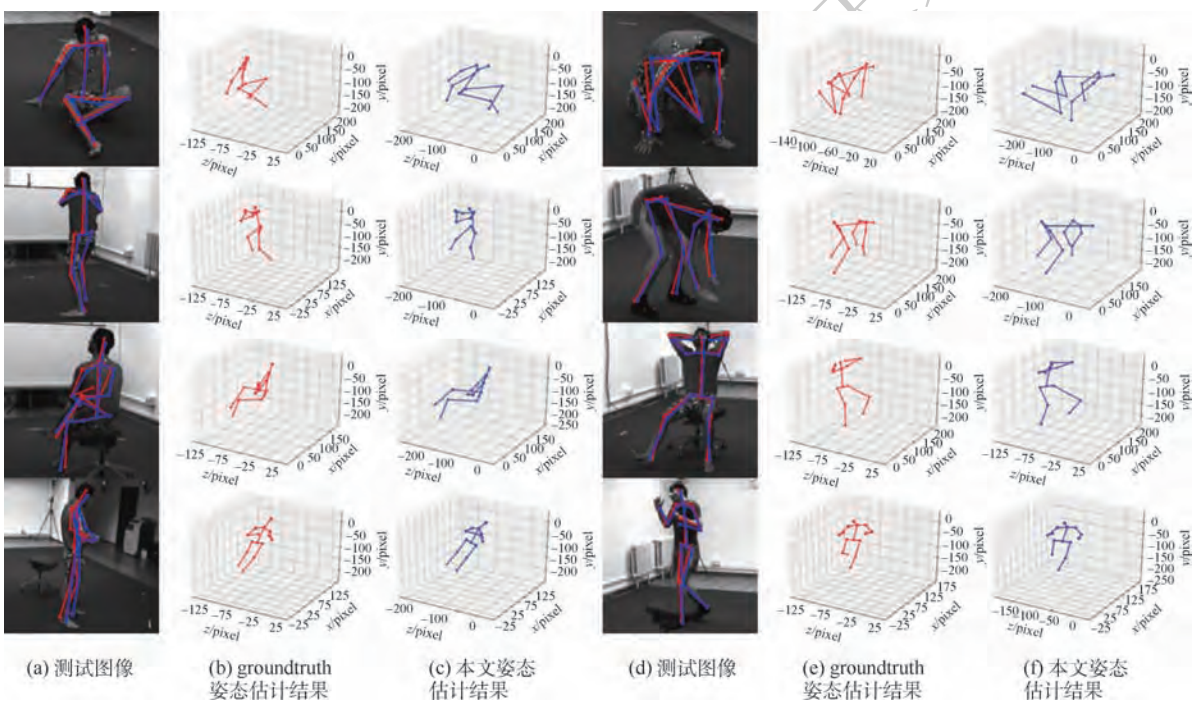


图 8 Human 3.6M 数据集上的 3D 人体姿态估计

Fig. 8 Three-dimensional human pose estimation on Human 3.6M dataset

4　结　论

本文针对缺乏深度标签的深度图像训练样本进行研究,提出了一种基于多源图像弱监督学习的 3D 人体姿态估计方法,以实现对深度图像的 3D 人体姿态估计任务。同时为改善网络的估计性能,本文对网络结构中的残差模块进行了改善设计。

1)针对深度图像训练样本中 3D 标注不足的问题,使用弱监督学习技术来完成 3D 回归模型训练任务。

2)针对深度图像姿态单一造成的模型泛化能力不高的问题,提出一种多源图像融合训练技术。该方法主要利用彩色图像姿态多变的特点,在网络训练阶段引入较为充分的人体姿态信息,提高模型的回归性能。

3)为提高姿态估计结果,基于提升回归模型准确率的基本思想,对残差模块的构成提出改善设计,并且通过实验结果证明该设计方案可在降低训练时间及模型存储空间基础上提高对图像的 3D 人体姿态估计准确度。

实验验证了在训练回归模型的网络结构中,一个合适的残差模块对提高回归模型准确率、降低参数量及训练时间等均有重要影响,因此接下来,本文将对如何更好地改善残差模块进行研究。

参考文献(References)

[1] PARK S, HWANG J, KWAK N. 3D human pose estimation using convolutional neural networks with 2D pose information [C]// European Conference on Computer Vision. Berlin: Springer, 2016: 156-169.

[2] YANG W, OUYANG W, LI H, et al. End-to-end learning of deformable mixture of parts and deep convolutional neural networks for human pose estimation[C]// IEEE Computer Society Conference on Computer Vision and Patter Recognition. Piscataway, NJ: IEEE Press, 2016: 3073-3082.

[3] ZE W K, FU Z S, HUI C, et al. Human pose estimation from depth images via inference embedded multi-task learning [C]// Proceedings of the 2016 ACM on Multimedia Conference. New York: ACM, 2016: 1227-1236.

[4] SHEN W, DENG K, BAI X, et al. Exemplar-based human action pose correction[J]. IEEE Transactions on Cybernetics, 2014, 44(7): 1053-1066.

[5] GULER R A, KOKKINOS L, NEVEROVA N, et al. DensePose: Dense human pose estimation in the wild[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 7297-7306.

[6] RHODIN H, SALZMANN M, FUA P. Unsupervised geometry-aware representation for 3D human pose estimation[C]// European Conference on Computer Vision. Berlin: Springer, 2018: 765-782.

[7] OMRAN M, LASSNER C, PONS-MOLL G, et al. Neural body fitting: Unifying deep learning and model based human pose and shape estimation[C]// International Conference on 3D Vision. Piscataway, NJ: IEEE Press, 2018: 484-494.

[8] HAQUE A, PENG B, LUO Z, et al. Towards viewpoint invariant 3D human pose estimation[C]// European Conference on Computer Vision. Berlin: Springer, 2016: 160-177.

[9] TOSHEV A, SZEGEDY C. DeepPose: Human pose estimation via deep neural networks[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 1653-1660.

[10] CAO Z, SIMON T, WEI S E, et al. Realtime multi-person 2D pose estimation using part affinity fields[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 7291-7299.

[11] WEI S E, RAMAKRISHNA V, KANADE T, et al. Convolutional pose machines[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 4724-4732.

[12] NEWELL A, YANG K, DENG J, et al. Stacked hourglass networks for human pose estimation[C]// European Conference on Computer Vision. Berlin: Springer, 2016: 483-499.

[13] YI Z X, XING H Q, XIAO S, et al. Towards 3D pose estimation in the wild: A weakly-supervised approach[C]// IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 398-407.

[14] SAM J, MARK E. Clustered pose and nonlinear appearance models for human pose estimation[C]// Proceedings of the 21st British Machine Vision Conference, 2010: 12.1-12.11.

[15] ANDRILUKA M, PISHCHULIN L, GEHLER P, et al. 2D human pose estimation: New benchmark and state of the art analysis[C]// IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2014: 3686-3693.

[16] CTALIN I, DRAGOS P, VLAD O, et al. Human 3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2014, 36(7): 1325-1339.

[17] HAN X F, LEUNG T, JIA Y Q, et al. MatchNet: Unifying feature and metric learning for patch-based matching[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 3279-3286.

[18] XU G H, LI M, CHEN L T, et al. Human pose estimation method based on single depth image[J]. IEEE Transactions on Computer Vision, 2018, 12(6): 919-924.

[19] SI J L, ANTONI B. 3D human pose estimation from monocular images with deep convolutional neural network[C]// Asian Conference on Computer Vision. Berlin: Springer, 2014: 332-347.

[20] GHEZELGHIEH M F, KASTURI R, SARKAR S. Learning camera viewpoint using CNN to improve 3D body pose estimation [C]// International Conference on 3D Vision. Piscataway, NJ: IEEE Press, 2016: 685-693.

[21] CHEN C H, RAMANAN D. 3D human pose estimation = 2D pose estimation + matching[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 5759-5767.

[22] POPA A I, ZANFIR M, SMINCHISESCU C. Deep multitask architecture for integrated 2D and 3D human sensing[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 4714-4723.

[23] COLLOBERT R, KAVUKCUOGLU K, FARABET C. Torch7: A Matlab-like environment for machine learning[C]// Conference and Workshop on Neural Information Processing Systems, 2011: 1-6.

[24] NIKOS K, GEORGIOS P, KOSTAS D. Convolutional mesh regression for single-image human shape reconstruction[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2019: 4510-4519.

[25] CHENXU L, XIAO C, ALAN Y. OriNet: A fully convolutional network for 3D human pose estimation[C]// British Machine Vision Conference, 2018: 321-333.

作者简介：

蔡轶珩 女，博士，副教授。主要研究方向：图像处理与识别、视觉感知信息处理、颜色科学。

王雪艳 女，硕士研究生。主要研究方向：图像与视频处理。

胡绍斌 男，硕士研究生。主要研究方向：图像处理。

刘嘉琦 男，硕士研究生。主要研究方向：图像与视频处理与分析。

Three-dimensional human pose estimation based on multi-source image weakly-supervised learning

CAI Yiheng*, WANG Xueyan, HU Shaobin, LIU Jiaqi

(Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China)

Abstract: Three-dimensional human pose estimation is a hot research topic in the field of computer vision. Aimed at the lack of labels in depth images and the low generalization ability of models caused by single human pose, this paper innovatively proposes a method of 3D human pose estimation based on multi-source image weakly-supervised learning. This method mainly includes the following points. First, multi-source image fusion training method is used to improve the generalization ability of the model. Second, weakly-supervised learning approach is proposed to solve the problem of label insufficiency. Third, in order to improve the attitude estimation results, this paper improve the design of the residual module. The experimental results show that the regression accuracy from our improved network increases by 0.2%, and meanwhile the training time reduces by 28% compared with the original network. In a word, the proposed method obtains excellent estimation results with both depth images and color images.

Keywords: human pose estimation; hourglass networks; weakly-supervised; multi-source image; depth image

Received: 2019-07-10; **Accepted**: 2019-08-23; **Published online**: 2019-08-29 13:16

URL: kns.cnki.net/kcms/detail/11.2625.V.20190828.1803.001.html

Foundation items: National Key R & D Project of China (2017YFC1703302); Science and Technology Plan Projects of Beijing Municipal Education Commission of China (KM201710005028)

* **Corresponding author**. E-mail: caiyiheng@bjut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0364

QoE 驱动的 VR 视频自适应采集与传输

黎洁[1,*]，冯燃生[1]，杨阳朝[2]，孙伟[3]，李奇越[3]

（1. 合肥工业大学 计算机与信息学院，合肥 230000； 2. 中国电子科技集团公司电子科学研究院，北京 100084；
3. 合肥工业大学 电气与自动化工程学院，合肥 230000）

摘 要：在虚拟现实(VR)视频流媒体传输中，如何在带宽受限的条件下提高用户的质量体验(QoE)是一项巨大的挑战。为了更好地提高资源利用率和用户的QoE，提出了一个面向多用户的QoE驱动上下行链路联合的VR视频流媒体自适应采集与传输系统。与传统的VR视频无线传输系统不同的是，所提系统考虑了上行传输部分。其中，视频服务器根据上行链路和下行链路的带宽信息、用户的实时视角信息，以速率自适应为基础进行码率选择和资源分配。定义了QoE驱动的码率选择和资源分配问题，以最大化整个系统所有用户的QoE值。提出了联合KKT条件和分支定界法的速率自适应选择算法。实验结果表明：所提系统可以有效提高所有用户的QoE值，与上行链路平均分配资源算法相比，系统QoE值提高了14.27%，同时与传统的VR视频速率自适应算法相比，系统QoE值提高了23.47%。

关 键 词：虚拟现实(VR)；质量体验(QoE)；资源分配；流媒体传输；KKT条件；分支定界法

中图分类号：V221+.3；TB553

文献标识码：A **文章编号**：1001-5965(2019)12-2385-08

近年来，随着交互式应用需求的兴起，虚拟现实(Virtual Reality，VR)技术越来越受到用户的广泛欢迎，全景视频流媒体传输就是其中最重要的应用之一。根据资料显示[1]，全球几大主要的流媒体服务运营商(Facebook、YouTube)已开始提供全景视频流媒体服务[2]。为了给用户提供更好的质量体验(Quality of Experience，QoE)，大多数VR视频的分辨率已经达到了4K(4 096×2 160)甚至更高[3]。然而，由于网络带宽的限制，直接传输如此高分辨率的视频不是一个可行的方案。因此，对于研究者来说在带宽受限的条件下，进一步提高用户的QoE是一项巨大的挑战[4]。

当用户观看VR视频时，由于显示设备视场(Field of View，FoV)的限制，在任一时刻用户往往只能观看到整个视频的某一部分。如果服务器传输整个视频的话，大部分的带宽资源会被不可避免地浪费掉。因此，基于DASH(Dynamic Adaptive Streaming over HTTP)[5]的自适应分块传输方式是进行VR视频流媒体传输最为提倡的一种方式。在分块传输方式中，一个完整的VR视频被分割成许多视频块，每个视频块被编码成不同的质量等级，服务器根据网络带宽等因素为每个视频块自适应选择一种最优的质量等级并传输到客户端[6-8]。

在自适应分块传输方式中，有2个最重要的方案：视角自适应和速率自适应。前者是预测用户视角变化的核心，后者是服务器进行资源分配

收稿日期：2019-07-08；**录用日期**：2019-08-23；**网络出版时间**：2019-08-27 13:27
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190827.1041.004.html
基金项目：国家自然科学基金(51877060)；中央高校基本科研业务费专项资金(JZ2019HGTB0089，JZ2018HGTB0253，PA2019GDQT0006)

*通信作者. E-mail：lijie@hfut.edu.cn

引用格式：黎洁，冯燃生，杨阳朝，等. QoE驱动的VR视频自适应采集与传输[J]. 北京航空航天大学学报，2019，45(12)：2385-2392. LI J，FENG R S，YANG Y Z，et al. QoE driven adaptation for VR video capturing and transmission[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2385-2392(in Chinese).

的核心。在此基础上，出现了很多相关的研究[9-12]。Qian 等[9]提出了一种视角自适应的360°视频传输系统，该系统利用机器学习算法进行视角预测。Corbillon 等[10]研究了基于通用理论模型的最优全景视频流媒体传输系统，结果显示所提出的最优系统与一般传输系统相比能够节省 45% 的带宽。文献[11]设计了一种速率自适应算法，该算法可以最大化预定义的 360°视频的 QoE 值。此外，也有结合视角自适应和速率自适应的研究。Xie 等[12]提出了基于概率模型的 360° 视频传输系统 360ProbDASH，该系统通过结合速率和视角自适应解决基于 QoE 的传输优化问题。

然而，上述 VR 视频流媒体传输系统存在以下关键性限制：①只考虑了单个用户，却忽略了多用户的情况。②仅针对下行链路（视频服务器到终端用户）传输，忽略了上行链路（VR 视频采集设备到视频服务器）传输。事实上，VR 视频直播是另一种重要的应用，在学术界和工业界都受到越来越多的关注[13-14]。在 VR 视频直播中，采集设备用于记录 VR 视频，并将视频实时传输到服务器。多用户可以使用相关显示设备请求和观看 VR 视频。显然，在带宽有限的约束下，应仔细研究上下行链路视频速率自适应传输算法。此外上行视频和下行视频是具有相关性的，单独优化设计会降低系统整体性能，这使得在该 VR 视频直播应用中速率自适应算法变得更加困难。因此，本文设计了一种面向多用户的上下行链路联合全景视频流媒体采集和传输系统。该系统不仅加入了上行和下行链路传输，还同时考虑了多用户情况下的速率自适应。此外，在所设计的系统基础上，定义了一个 QoE 驱动下的多用户资源分配问题。

1 上下行链路联合 VR 视频流媒体采集和传输

1.1 系统模型

图 1 为系统应用场景。系统上行端 VR 采集设备一共有 C 个相机，下行端共有 N 个用户。首先，每个相机拍摄出相同码率的高清视频，每个视频都被编码成不同的质量等级，获得不同码率的视频，由服务器为 C 个视频各选择一种质量等级进行上行传输。然后，服务器将这 C 个视频缝合成一个完整的 VR 视频，并进行分块和压缩处理。最后，处理过的 VR 视频会通过下行链路传输到每个用户。此外，假设下行链路中还存在从用户

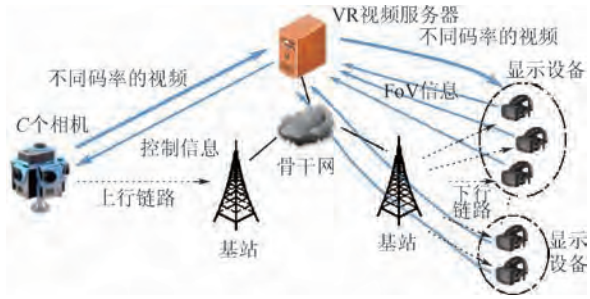


图 1 系统应用场景

Fig. 1 System application scenario

到服务器的端到端反馈信道，此反馈信道可以将每个用户的带宽信息及实时的视角信息传输到服务器，更好地帮助服务器进行视频速率自适应选择。

本文设计系统的结构如图 2 所示。系统主要由 4 部分组成：C 个相机、服务器、客户端和 N 个用户。

QoE 驱动上行模块和下行模块一起完成优化传输的工作，即合理进行视频码率自适应选择去最大化所有用户的 QoE。具体来说，服务器的工作流程为：首先，上行处理模块为 C 个相机各选择一种合适的码率等级，当这些视频上传到服务器后，经过缝合、映射等处理生成了一个 VR 视频。然后，下行处理模块把完整的全景视频分成多块，并且将每块视频编码成不同的质量等级。此时，在反馈信道中特征数据（用户实时 FoV 信息、下行链路带宽信息）的帮助下，下行处理模块可以确定被传输的块并为每块视频选择最合适的质量等级。最后，这些块都将通过下行链路进行传输，在传输的同时，下行处理模块还将为每块视频选择一种合适的调制与编码策略（Module and Coding Scheme，MCS）。在客户端中，当所有视频块通过下行链路传输到客户端，它们经过解码、映射、渲染等处理，最终通过显示设备呈现给每个用户。此外，客户端的另一个功能是把用户的实时 FoV 信息和下行链路带宽信息通过反馈信道传输给服务器。

1.2 问题形成

假设系统中在上行端有 C 个相机（每个相机用 c 表示），下行端有 N 个用户（每个用户用 n 表示）。在各个方向上均匀安装的 C 个相机能够形成一个 VR 视频采集系统，对每个相机 c 来说，其能够拍摄产生 D'种不同码率等级（每种码率等级为 d'）的视频，用 $\lambda_{c,d'}^{\mathrm{UL}}$ 表示码率等级为 d' 的相机 c 产生视频的码率。假设上行链路总带宽为

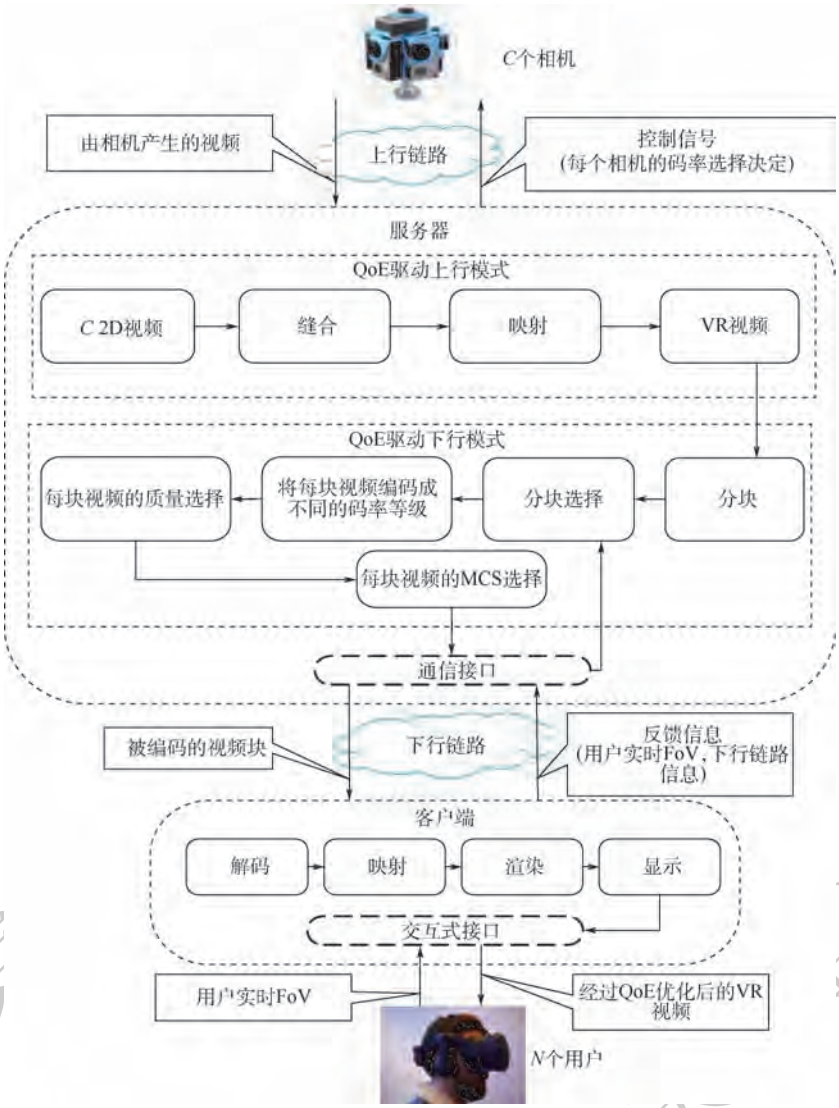


图 2　系统结构

Fig. 2　System structure

$\mathrm{BW}^{\mathrm{UL}}$。当 C 个视频上传到服务器,服务器会对其进行缝合以生成一个 VR 视频,再将其分成 T 块(每块用 t 表示),然后每块被编码成 D 种质量等级(d 表示每种质量等级)。用 $\lambda_{t,d}^{\mathrm{DL}}$ 表示质量等级为 d 的视频块 t 的码率。显然,下行链路中每块视频的质量等级必定小于上行链路中每个视频的质量等级。在下行链路中,假设一共有 Y^{DL} 个资源块(RB),同时下行链路可以提供 M 种 MCS 方案(每种方案用 m 表示)。用 R_m^{DL} 表示在 MCS 方案为 m 时,单个资源块所能提供的比特率。

对于用户 n 来说,T_{FoV}^n 表示其 FoV 所覆盖的视频块。因此,用户 n 的 QoE 值可以定义为

$$\mathrm{QoE}_n = \log_2\left(\frac{\sum_{t=1}^{T_{\mathrm{FoV}}^n}\sum_{d=1}^{D}\sum_{m=1}^{M}\lambda_{t,d}^{\mathrm{DL}}\chi_{t,d,m}^{\mathrm{DL}}}{\lambda_{t,D}^{\mathrm{DL}}}\right) \tag{1}$$

式中:$\chi_{t,d,m}^{\mathrm{DL}}$ 为一个二进制变量,只有当质量等级为 d 的块 t 以 MCS 方案 m 传输时值才为 1,否则为 0;$\lambda_{t,D}^{\mathrm{DL}}$ 为块 t 以最高的质量等级 D 传输时的码率。

通过 QoE 表达式,本文的 QoE 驱动上下行链路联合 VR 流媒体自适应采集和传输的问题可以描述为:确定上行链路中每个相机所拍摄的视频的码率等级,为下行链路中全景视频的视频块选择一种质量等级和一种 MCS 方案,来最大化整个系统中的所有用户的 QoE 值。可以用问题 P1 来表示。

问题 P1:

$$\max \sum_{n=1}^{N}\mathrm{QoE}_n = \sum_{n=1}^{N}\log_2\left(\frac{\sum_{t=1}^{T_{\mathrm{FoV}}^n}\sum_{d=1}^{D}\sum_{m=1}^{M}\lambda_{t,d}^{\mathrm{DL}}\chi_{t,d,m}^{\mathrm{DL}}}{\lambda_{t,D}^{\mathrm{DL}}}\right) \tag{2}$$

s. t.

$$\sum_{d'=1}^{D'}\chi_{c,d'}^{\mathrm{UL}} = 1 \qquad \forall c \tag{3}$$

$$\sum_{c=1}^{C}\sum_{d'=1}^{D'}\chi_{c,d'}^{\mathrm{UL}}\lambda_{c,d'}^{\mathrm{UL}}\leqslant \mathrm{BW}^{\mathrm{UL}} \tag{4}$$

$$\sum_{m=1}^{M}\chi_{t,d,m}^{\mathrm{DL}}=1 \qquad \forall t,d \tag{5}$$

$$\sum_{d=1}^{D}\chi_{t,d,m}^{\mathrm{DL}}=1 \qquad \forall t,m \tag{6}$$

$$\sum_{t=1}^{T_{\mathrm{FoV}}^{n}}\sum_{d=1}^{D}\sum_{m=1}^{M}\chi_{t,d,m}^{\mathrm{DL}}\left[\frac{\lambda_{t,d}^{\mathrm{DL}}}{R_{m}^{\mathrm{DL}}}\right]\leqslant Y^{\mathrm{DL}} \tag{7}$$

$$\sum_{d=1}^{D}\sum_{m=1}^{M}\chi_{t,d,m}^{\mathrm{DL}}\lambda_{t,d}^{\mathrm{DL}}\leqslant\frac{1}{T}\sum_{d'=1}^{D'}\chi_{c,d'}^{\mathrm{UL}}\lambda_{c,d'}^{\mathrm{UL}} \qquad \forall c,t \tag{8}$$

问题 P1 的解变量是 2 个二进制变量 $\chi_{c,d'}^{\mathrm{UL}}$ 和 $\chi_{t,d,m}^{\mathrm{DL}}$。$\chi_{c,d'}^{\mathrm{UL}}$ 与 $\chi_{t,d,m}^{\mathrm{DL}}$ 类似，都是二进制变量，只有当相机 c 产生的视频以码率等级 d' 上传时值才为 1，否则为 0。

约束条件(3)和约束条件(4)是上行部分的约束。约束条件(3)表示任何一个相机 c 只能选择一种码率等级，约束条件(4)表示传输的 C 个视频的总码率不应当超出整个上行链路的总带宽。

约束条件(5)～约束条件(7)是下行部分的约束。约束条件(5)和约束条件(6)分别表示任意视频块只能选择一种 MCS 方案、一种质量等级。约束条件(7)确保了传输的所有视频块的总码率不超过整个下行链路所有资源块所能提供的比特率。

约束条件(8)是整个系统上下行链路联合的约束，表示下行链路中任意视频块的码率不大于上行链路中任意相机产生的视频的码率。

2 基于联合 KKT 条件和分支定界法的速率自适应选择算法

显然，问题 P1 是一个非线性整数规划(NIP)问题。通过分析该问题的性质，若将问题 P1 的求解变量松弛为连续变量后，其目标函数和约束条件均为便于求解的凸函数。因此，先将问题 P1 进行连续松弛，将非线性整数规划问题变成非线性规划问题，用 KKT 条件[15]求出该非线性规划问题的解；再采用分支定界法求解出原问题 P1 的 0-1 变量解。

2.1 松弛后非线性规划问题的 KKT 条件

先对问题 P1 的 2 组解 $\chi_{c,d'}^{\mathrm{UL}}$ 和 $\chi_{t,d,m}^{\mathrm{DL}}$ 进行松弛，该问题变成了非线性规划问题，可用 KKT 条件求解。

该问题的拉格朗日函数为

$$\begin{aligned}L(\lambda_{c,d'}^{\mathrm{UL}},\chi_{t,d,m}^{\mathrm{DL}},\lambda,\mu)=&-\sum_{n=1}^{N}\mathrm{QoE}_{n}+\lambda_{1}h(\chi_{c,d'}^{\mathrm{DL}})+\\&\lambda_{2}h_{1}(\chi_{t,d,m}^{\mathrm{DL}})+\lambda_{3}h_{2}(\chi_{t,d,m}^{\mathrm{DL}})+\mu_{1}g(\chi_{c,d'}^{\mathrm{DL}})+\\&\mu_{2}g(\chi_{t,d,m}^{\mathrm{DL}})+\mu_{3}g(\chi_{c,d'}^{\mathrm{DL}},\chi_{t,d,m}^{\mathrm{DL}})\end{aligned} \tag{9}$$

式中：

$$h(\chi_{c,d'}^{\mathrm{DL}})=\sum_{d'=1}^{D'}\chi_{c,d'}^{\mathrm{UL}}-1 \tag{10}$$

$$h_{1}(\chi_{t,d,m}^{\mathrm{DL}})=\sum_{m=1}^{M}\chi_{t,d,m}^{\mathrm{DL}}-1 \tag{11}$$

$$h_{2}(\chi_{t,d,m}^{\mathrm{DL}})=\sum_{d=1}^{D}\chi_{t,d,m}^{\mathrm{DL}}-1 \tag{12}$$

$$g(\chi_{c,d'}^{\mathrm{DL}})=\sum_{c=1}^{C}\sum_{d'=1}^{D'}\chi_{c,d'}^{\mathrm{UL}}\lambda_{c,d'}^{\mathrm{UL}}-\mathrm{BW}^{\mathrm{UL}} \tag{13}$$

$$g(\chi_{t,d,m}^{\mathrm{DL}})=\sum_{t=1}^{T_{\mathrm{FoV}}^{n}}\sum_{d=1}^{D}\sum_{m=1}^{M}\chi_{t,d,m}^{\mathrm{DL}}\left[\frac{\lambda_{t,d}^{\mathrm{DL}}}{R_{m}^{\mathrm{DL}}}\right]-Y^{\mathrm{DL}} \tag{14}$$

$$g(\chi_{c,d'}^{\mathrm{DL}},\chi_{t,d,m}^{\mathrm{DL}})=\sum_{d=1}^{D}\sum_{m=1}^{M}\chi_{t,d,m}^{\mathrm{DL}}\lambda_{t,d}^{\mathrm{DL}}-\frac{1}{T}\sum_{d'=1}^{D'}\chi_{c,d'}^{\mathrm{UL}}\lambda_{c,d'}^{\mathrm{UL}} \tag{15}$$

因此，可以得到相关的 KKT 条件为

$$\frac{\partial L(\lambda_{c,d'}^{\mathrm{UL}},\chi_{t,d,m}^{\mathrm{DL}},\lambda,\mu)}{\partial\lambda_{c,d'}^{\mathrm{UL}}}=\lambda_{1}\frac{\partial h(\chi_{c,d'}^{\mathrm{DL}})}{\partial\lambda_{c,d'}^{\mathrm{UL}}}+\mu_{1}+\mu_{3}\frac{\partial g(\chi_{c,d'}^{\mathrm{DL}},\chi_{t,d,m}^{\mathrm{DL}})}{\partial\lambda_{c,d'}^{\mathrm{UL}}}=0 \tag{16}$$

$$\begin{aligned}\frac{\partial L(\lambda_{c,d'}^{\mathrm{UL}},\chi_{t,d,m}^{\mathrm{DL}},\lambda,\mu)}{\partial\chi_{t,d,m}^{\mathrm{DL}}}=&-\sum_{n=1}^{N}\frac{\partial\mathrm{QoE}_{n}}{\partial\chi_{t,d,m}^{\mathrm{DL}}}+\\&\lambda_{2}\frac{\partial h_{1}(\chi_{t,d,m}^{\mathrm{DL}})}{\partial\chi_{t,d,m}^{\mathrm{DL}}}+\lambda_{3}\frac{\partial h_{2}(\chi_{t,d,m}^{\mathrm{DL}})}{\partial\chi_{t,d,m}^{\mathrm{DL}}}+\\&\mu_{2}\frac{\partial g(\chi_{t,d,m}^{\mathrm{DL}})}{\partial\chi_{t,d,m}^{\mathrm{DL}}}+\mu_{3}\frac{\partial g(\chi_{c,d'}^{\mathrm{DL}},\chi_{t,d,m}^{\mathrm{DL}})}{\partial\chi_{t,d,m}^{\mathrm{DL}}}=0\end{aligned} \tag{17}$$

$$g(\chi_{c,d'}^{\mathrm{DL}})\leqslant 0,g(\chi_{t,d,m}^{\mathrm{DL}})\leqslant 0,g(\chi_{c,d'}^{\mathrm{DL}},\chi_{t,d,m}^{\mathrm{DL}})\leqslant 0 \tag{18}$$

$$h(\chi_{c,d'}^{\mathrm{DL}})=0,h_{1}(\chi_{t,d,m}^{\mathrm{DL}})=0,h_{2}(\chi_{t,d,m}^{\mathrm{DL}})=0 \tag{19}$$

$$\lambda_{1},\lambda_{2},\lambda_{3}\neq 0,\mu_{1},\mu_{2},\mu_{3}\geqslant 0 \tag{20}$$

$$\mu_{1}g(\chi_{c,d'}^{\mathrm{DL}})=0,\mu_{2}g(\chi_{t,d,m}^{\mathrm{DL}})=0,\mu_{3}g(\chi_{c,d'}^{\mathrm{DL}},\chi_{t,d,m}^{\mathrm{DL}})=0 \tag{21}$$

通过联合 KKT 条件相关的公式(式(16)～式(21))进行求解，可以得出经过松弛的非线性问题的最优解，记为 χ_{relax}，接下来使用分支定界法求出原始问题的 0-1 变量解。

2.2 基于分支定界的上下行链路 VR 视频速率自适应选择算法

起始输入为 χ_{relax} 和 Z_{relax}，Z_{relax} 表示 χ_{relax} 对应的最优目标函数值。输出结果为所求的 0-1 变量解 $\chi_{0\text{-}1}$ 和相应的最优目标函数值 $Z_{0\text{-}1}$。

算法 1 上下行链路 VR 视频速率自适应选择算法。

输入：用 KKT 条件求解出的对应松弛问题的解 χ_{relax} 和松弛问题的最优目标函数值 Z_{relax}。

输出：问题 P1 的符合 0-1 约束条件的最优解 $\chi_{0\text{-}1}$ 和问题 P1 的最优 0-1 解对应的最优目标函数值 $Z_{0\text{-}1}$。

初始化：$k=0, L=0, U=Z_{\text{relax}}$。

1. 从 χ_{relax} 中任意选则一个不符合 0-1 约束条件的解 χ_j，在（0，1）范围内随机产生一个随机数 ε。

2. IF $0 \leqslant \chi_j < \varepsilon$，THEN 将约束条件 $\chi_j = 0$ 加入到问题 P1 中，形成子问题Ⅰ。

3. ELSE，将约束条件 $\chi_j = 1$ 加入到问题 P1 中，形成子问题Ⅱ。

4. END IF

5. 继续求出子问题Ⅰ和子问题Ⅱ的松弛问题解，记为 χ_k，并将对应最优目标函数值记为 Z_k。

6. $U = \max\{Z_k\}, \chi_k \in [0,1]$。

7. $L = \max\{Z_k\}, \chi_k \in \{0,1\}$。

8. IF $Z_k < L$，THEN 剪掉这个分支。

9. ELSE IF $Z_k > L$，THEN 回到分支步骤并重复。

10. ELSE $Z_k = L$，THEN 问题 P1 的最优解已经找到，$Z_{0\text{-}1} = Z_k, \chi_{0\text{-}1} = \chi_k$。

11. END IF

3　仿真实验与结果

3.1　仿真实验

为了验证本文算法的有效性，选择了 6 个用 Insta 360 Pro2 全景相机所拍摄的视频作为上行链路的原始视频，每个视频的分辨率都为 960 × 720，长度为 35 s，帧率为 30 fps（fps 为帧/s）。图 3 为所采用的 6 个原始视频的缩略图。采用 HEVC（High Efficiency Video Coding）编码器对每个原始视频进行压缩，分别以 17，20，23，26 这 4 种量化参数（Quantization Parameter，QP）压缩，为每个视频设置 4 种码率等级：{1.5，2，2.5，3} Mbit/s。使用 VR 视频合成软件 AVP（Kolor Autopano Video Pro）完成全景视频的合成工作。把 VR 视频分成 4 × 4 = 16 块，每块覆盖视角为 90° × 45°。为每块下行视频设置 6 种不同码率等级以供选择：{0.2，0.4，0.6，0.8，1，1.2} Mbit/s。图 4 为使用 6 个原始视频所合成的 VR 视频缩略图。实验中，用户端使用 VR 头戴式显示设备 HTC Vive，使用户进行观看以此来获得不同用户的 FoV。具体选择了 100 个用户和 48 个的 VR 视频[16]，用户观看不同的 VR 视频，通过所获得的 100 个用户 FoV 数据，经过计算分析得出有 60% 的用户 FoV 大约为 120° × 90°，这说明大多数用户在观看实验所用全景视频时，视角都会在这个范围内。因此，将 FoV 设置为 120° × 90°。在下行传输过程中，为每块视频提供 4 种 MCS 方案。具体见表 1。在仿真实验中，使用所设计的算法平均求解一次最优解需要 1 min。

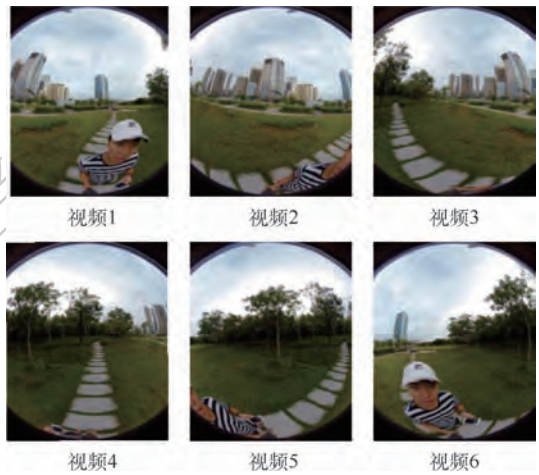


图 3　6 个原始视频的缩略图

Fig. 3　Thumbnail of six original videos

图 4　采用分块后的全景视频

Fig. 4　Tiled panoramic video

表 1　MCS 方案

Table 1　MCS scheme

等级	调制方式	每个资源块的数据率/(Kbit · s^{-1})
1	QPSK 1/2	4.8
2	QPSK 3/4	7.2
3	QAM16 1/2	9.6
4	QAM16 3/4	21.6

3.2　仿真结果

为了证明本文算法的性能，采用 3 种算法来对比。第 1 种算法是在上行部分中采用平均分配资源，即把上行链路中的网络带宽资源平均分配到不同的相机中。第 2 种算法是下行速率自适应算法，该算法与传统的 VR 视频传输算法相同，都是基于 DASH 的自适应分块传输算法，只考虑从服务器到用户的下行部分自适应传输。第 3 种是

下行视角自适应算法,该算法通过结合用户视角预测和网络带宽预测,来实时地进行从服务器到用户的下行部分自适应传输。

分析不同上行链路带宽条件下所有用户的 QoE 值,性能如图 5 所示。能够清楚地发现,当上行链路带宽相同时,本文算法能够达到更高的 QoE 值,说明在上行部分有着更好的性能。此外,下行速率自适应算法在不同上行带宽条件下,在 4 种算法之中性能较差。

对比下行链路中不同资源块数量对系统整体 QoE 的影响,如图 6 所示。可以发现,在 4 种算法中,本文算法在不同下行链路资源块数量条件下仍有较高的 QoE 值,性能要远高于下行速率自适应算法和上行平均分配资源算法,略高于下行视角自适应算法。同时可以发现,下行视角自适应算法的性能要优于下行速率自适应算法,这也体现出了视角预测与网络带宽预测在 VR 视频自适应传输中的重要性。下行视角自适应算法的性能仍要略低于本文算法,这也说明了上下行传输一体化的优良性。

表 2 给出了 4 种算法在相同上下行链路带宽、用户数目时的系统 QoE 值和比较情况。可以发现,本文算法的 QoE 值比上行平均分配资源算法提高了14.27%,比下行视角自适应算法提高了11.58%,比下行速率自适应算法提高了 23.47%。

图 7 为用户数目与系统 QoE 值的关系。可以发现,系统 QoE 值随总用户数的增加而一直增加,但是增加的速度会逐渐下降,甚至会降低。这

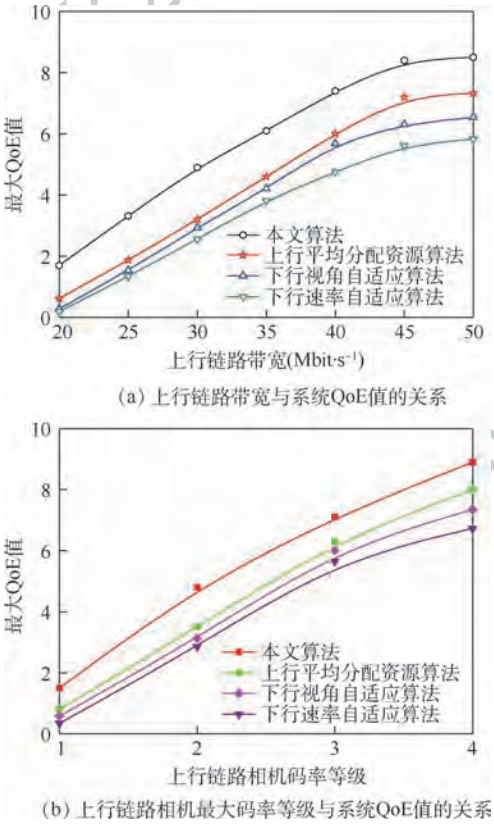


图 5 上行链路与系统 QoE 值的关系

Fig. 5 Relationship between uplink and system QoE value

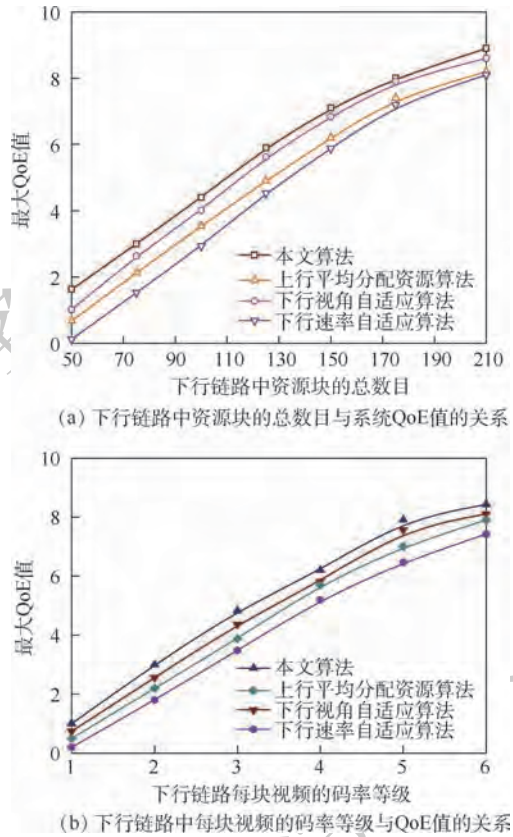


图 6 下行链路与系统 QoE 值的关系

Fig. 6 Relationship between downlink and system QoE value

表 2 不同算法 QoE 值比较

Table 2 QoE value comparison among different algorithms

算法	QoE 值
本文算法	5.828
上行平均分配资源算法	5.1
下行视角自适应算法	5.223
下行速率自适应算法	4.72

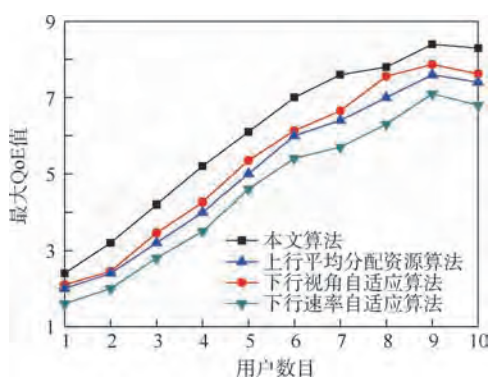


图 7 用户数目与系统 QoE 值的关系

Fig. 7 Relationship between user number and system QoE value

是因为当用户数目增加到一定程度时,系统为了确保每个用户都能正常地观看视频,牺牲了每个用户所得到的码率等级,即下降了每块视频的码率等级。

4　结　论

本文提出了一种基于 QoE 的多用户自适应传输系统,通过对仿真结果的分析可以得出如下结论:

1)提出的 VR 视频传输系统与传统的 360°全景视频传输系统相比,不仅考虑了下行传输也考虑了上行传输部分。

2)提出的基于联合 KKT 条件和分支定界法的速率自适应选择算法可以解决所提出的传输系统中的资源分配问题。

3)提出的速率自适应选择算法与上行链路平均分配资源算法相比系统 QoE 值提高了 14.27%,与传统的 VR 视频速率自适应算法相比系统 QoE 值提高了 23.47%。

参考文献(References)

[1] Cisco. Cisio visual networking index:Forecast and methodology 2016—2021[R]. San Jose:Cisco,2016.

[2] ZHAO T, LIU Q,CHEN C W. QoE in video transmission:A user experience-driven strategy[J]. IEEE Communications Surveys Tutorials, 2017, 19(1):285-302.

[3] LI J, FENG R S,LIU Z,et al. Modeling QoE of virtual reality video transmission over wireless networks[C]// Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM). Piscataway,NJ:IEEE Press,2018:1-7.

[4] LIU C,KAN N,ZOU J,et al. Server-side rate adaptation for multi-user 360-degree video streaming[C]// Proceedings of the 25th IEEE International Conference on Image Processing (ICIP). Piscataway,NJ:IEEE,2018:3264-3268.

[5] SODAGAR I. The MPEG-DASH standard for multimedia streaming over the internet[J]. IEEE Multimedia, 2011, 18(4):62-67.

[6] GUO C J,YING C,LIU Z. Optimal multicast of tiled 360 VR video[J]. IEEE Wireless Communications Letters, 2019, 8(1):145-148.

[7] LIU Y,XU M,LI C,et al. A novel rate control scheme for panoramic video coding[C]// Proceedings of the 2017 IEEE International Conference on Multimedia and Expo(ICME). Piscataway,NJ:IEEE Press,2017:691-696.

[8] XU Z,BAN Y, ZHANG K,et al. Tile-based QoE-driven http/2 streaming system for 360 video[C]// Proceedings of the 2018 IEEE International Conference on Multimedia Expo Workshops (ICMEW). Piscataway,NJ:IEEE Press,2018:1-4.

[9] QIAN F,HAN B,XIAO Q,et al. Flare:Practical viewport-adaptive 360-degree video streaming for mobile devices[C]// Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom). New York: ACM, 2018:99-114.

[10] CORBILLON X, DEVLIC A, SIMON G, et al. Optimal set of 360-degree videos for viewport-adaptive streaming[C]// Proceedings of the 25th ACM International Conference on Multimedia. New York:ACM,2018:943-951.

[11] GHOSH A,VANEET A,FENG Q. A rate adaptation algorithm for tile-based 360-degree video streaming[EB/OL]. (2017-04-26)[2019-09-21]. https://arxiv.org/abs/1704.08215.

[12] XIE L,XU Z,BAN Y,et al. 360ProbDASH:Improving QoE of 360 video streaming using tile-based HTTP adaptive streaming [C]// Proceedings of the 25th ACM International Conference on Multimedia. New York:ACM,2017:315-323.

[13] HUNG Y,WANG C,HWANG R. Optimizing social welfare of live video streaming services in mobile edge computing[J/OL]. IEEE Transactions on Mobile Computing(2019-02-26)[2019-09-21]. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8653413.

[14] ZHANG L,SUN L,WANG W,et al. Unlocking the door to mobile social VR:Architecture, experiments and challenges[J]. IEEE Network,2018,32(1):160-165.

[15] KIM H J,SON Y S,KIM J T. KKT-conditions based resource allocation algorithm for DASH streaming service over LTE [C]//2018 IEEE International Conference on Consumer Electronics(ICCE). Piscataway,NJ:IEEE Press,2018:1-3.

[16] XU M,LI C,WANG Z L,et al. Visual quality assessment of panoramic video[EB/OL]. (2019-01-15)[2019-07-01]. https://arxiv.org/abs/1709.06342v1.

作者简介:

黎洁　女,博士,副研究员。主要研究方向:视频流媒体传输、虚拟现实。

冯燃生　男,硕士研究生。主要研究方向:虚拟现实流媒体传输。

杨阳朝　男,博士。主要研究方向:无线通信。

孙伟　男,博士,副教授。主要研究方向:无线通信。

李奇越　男,博士,副研究员。主要研究方向:无线通信。

QoE driven adaptation for VR video capturing and transmission

LI Jie[1,*], FENG Ransheng[1], YANG Yangzhao[2], SUN Wei[3], LI Qiyue[3]

(1. School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230000, China;
2. Academy of Electronics and Information Technology, China Electronics Technology Group Corporation, Beijing 100084, China;
3. School of Electrical Engineering and Automation, Hefei University of Technology, Hefei 230000, China)

Abstract: In virtual reality (VR) video streaming media transmission, how to further improve user's quality of experience (QoE) under bandwidth-constrained conditions is a huge challenge. In order to improve resource utilization rate and user QoE, a multi-user QoE-driven uplink and downlink joint VR video streaming media adaptive acquisition and transmission system is proposed, which is different from the traditional VR video wireless transmission system. The proposed system considers the uplink transmission part. The video server selects code rate and allocates resources based on the rate adaptation based on the bandwidth information of the uplink channel and the downlink channel and the real-time view information of the user. In addition, we define the problem of QoE-driven rate selection and resource allocation to maximize the QoE value of all users across the system. Finally, we propose an optimal adaptive rate selection algorithm combining the KKT condition and the branch and bound method. The experimental results show that the system can effectively improve the QoE value of the total system users, improve the system performance by 14.27% based on the average uplink allocation, and improve the performance of the VR video rate adaptive algorithm by 23.47%.

Keywords: virtual reality (VR); quality of experience (QoE); resource allocation; streaming media transmission; KKT condition; branch and bound method

Received: 2019-07-08; **Accepted**: 2019-08-23; **Published online**: 2019-08-27 13:27

URL: kns.cnki.net/kcms/detail/11.2625.V.20190827.1041.004.html

Foundation items: National Natural Science Foundation of China (51877060); the Fundamental Research Funds for the Central Universities (JZ2019HGTB0089, JZ2018HGTB0253, PA2019GDQT0006)

* **Corresponding author**. E-mail: lijie@hfut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0377

基于水下机器人的海产品智能检测与自主抓取系统

徐凤强，董鹏，王辉兵，付先平*

（大连海事大学 信息科学技术学院，大连 116026）

摘 要：针对水下机器人实现自主抓取过程中缺乏引导系统的问题，提出了一种依托水下机器人的海产品智能检测与自主抓取系统，用来解决水下目标的智能检测问题，并引导水下机器人进行海产品的自主抓取。将卷积神经网络检测算法应用到水下场景，利用水下图像数据集训练特定的网络模型 DSOD 检测海产品。建立短基线定位系统定位水下作业的机器人。通过分析相机成像坐标系与定位系统坐标系之间的关系，提出了一种计算海产品实际位置的坐标转换方法，计算海产品的实际位置。设计了一种基于反馈机制的多信号分析方法，引导机器人在水下移动并抓捕海产品。为了验证所提系统的有效性，搭建了一款水下抓捕机器人，并成功将所提算法应用到机器人，在真实海洋环境中进行海产品的自主抓取实验。

关 键 词：水下抓捕机器人；智能检测；水下目标定位；短基线定位系统；自主抓取
中图分类号：V221+.3；TB553
文献标识码：A **文章编号**：1001-5965(2019)12-2393-10

利用水下机器人智能检测和自主抓取海产品成为当前海产养殖业的迫切需要，这是建立现代化海洋牧场的重要途径。目前，海参、扇贝等海产品的捕捞主要依靠潜水员和拖网船 2 种方式。潜水员不能在水下持续作业，并且受水下压强影响，常年从事水下捕捞工作的渔民容易得潜水病。大面积养殖的贝类主要依靠拖网船捕捞，但是这种拖网捕捞方式容易破坏水下生态环境。而水下敏捷机器人能够克服这些问题，因此，利用水下机器人进行海产品的捕捞成为当下研究的热点。

目前，水下敏捷机器人的研究大多集中在水下观测和水下探测等方面，而水下抓捕机器人的研究成果却很少。仅有的水下抓捕机器人也仍然停留在依靠人工操控进行抓取作业的层面，而自主抓取机器人更加难以实现。造成这种研究困境的原因为：①复杂的水下环境对机器人部件的防水性和气密性等要求极高，市场缺乏成熟的适用于水下机器人的各种配件；②缺乏服务于水下机器人的智能检测与自主抓取系统。

因此，本文提出了一种智能检测与自主抓取系统，用来与机器人协同作业，引导机器人在脱离人工操控的情况下，实现对海产品的智能检测及自主抓取操作。该系统主要解决水下目标的检测问题和机器人的水下运动问题，这是实现机器人自主抓取目标必须解决的 2 个关键问题。

针对水下目标检测问题，将目前比较流行的深度学习算法引入到水下识别场景中，来提高水下目标检测效果。由于光在水下的折射及散射作用，导致相机拍摄的图像呈现模糊、偏色、光线偏暗等现象。传统的图像检测方法对于水下图像的处理效果并不理想。鉴于深度学习在各个领域的应用都取得了非常好的效果，本文将目前效果比

收稿日期：2019-07-09；**录用日期**：2019-08-20；**网络出版时间**：2019-08-26 15:22
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190826.1509.003.html
基金项目：国家自然科学基金（61272368，61370142）；中央高校基本科研业务费专项资金（3132016352）；交通运输部应用基础研究项目（2015329225300）；大连市科技创新项目（2018J12GX037）

* **通信作者**. E-mail：fxp@dlmu.edu.cn

引用格式：*徐凤强，董鹏，王辉兵，等. 基于水下机器人的海产品智能检测与自主抓取系统[J]. 北京航空航天大学学报，2019，45(12)：2393-2402. XU F Q，DONG P，WANG H B，et al. Intelligent detection and autonomous capture system of seafood based on underwater robot[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2393-2402 (in Chinese).*

较好的 DSOD(Deeply Supervised Object Detector)[1]应用到水下目标检测任务,利用标注的水下图像数据集训练网络模型,检测海参、扇贝、海胆等水下目标。

对于机器人的水下运动问题,本文提出了基于反馈机制的多信号分析方法,通过获取海产品位置信息、机器人定位信息、机器人深度等信号进行分析,得出机器人下一步的运动方案并反馈给机器人,从而引导其在水下寻找海产品并抓取目标。为了定位机器人在水下的位置,搭建了短基线定位系统[2]。利用声波信号在接收器和定位器之间的传播,计算机器人在定位系统中的位置。通过分析相机成像坐标系与定位系统坐标系之间的关系,提出了一种坐标转换方法,将海产品的检测位置坐标转换到定位系统坐标系,进而控制机器人向目标位置移动。机器人的深度信息在机器人下潜运动中起到重要作用。

为了验证本文系统的有效性,搭建了一个水下抓捕机器人,将本文算法应用到机器人上,引导其在水下完成海产品的智能检测及自主抓取工作。

本文的主要贡献如下:①将 DSOD 检测算法应用到水下场景,解决水下目标检测问题;②提出了一种坐标转换方法,可以将检测目标的位置坐标转换到定位系统坐标系;③设计了一种基于反馈机制的多信号分析方法,引导机器人自动检测目标并完成自主抓取操作。

1 相关工作

1.1 基于深度学习的目标检测

目标检测是计算机视觉领域的一个研究热点,其不仅需要区分目标的种类,还要精确地定位到目标的包围框。与传统的目标检测算法相比,基于深度学习的目标检测算法显著提高了目标检测的准确率。依据是否需要提取候选区域,可以将基于卷积神经网络(CNN)的目标检测算法分为 2 类:基于候选区域的目标检测算法和基于回归的目标检测算法[1]。

基于候选区域的目标检测算法先从整幅图像提取候选区域,再利用候选区域进行分类和回归,得到检测结果。Girshick 等[3]提出了候选区域与卷积神经网络相结合的方法 R-CNN,检测效果提升明显,但是存在大量的重复计算。因此,Girshick[4]提出了一个更加高效的深度卷积网络 Fast R-CNN,提高了检测效率。Ren 等[5]进一步提出了 RPN(Region Proposal Network),实现卷积特征的共享,降低网络训练的资源消耗。在以上算法基础上改进的网络,如 Mask R-CNN[6]、R-FCN[7]等也取得了很好的检测效果。

基于回归的目标检测算法摒弃了候选区域提取操作,是一种端到端的目标检测算法,如 YOLO[8]、SSD[9]、DSSD[10]、DSOD[1]等。这类算法把目标检测看成一个回归问题,直接利用神经网络从整幅图上检测并定位目标,检测速度较快,但是检测效果稍逊一筹。

当前主流目标检测算法主要集中解决陆地场景问题[11],本文将基于卷积神经网络的目标检测算法引入水下场景进行水下目标检测。综合考虑算法的检测速度和准确率,选择 DSOD 作为检测网络。

1.2 水下机器人定位

目前,比较成熟的定位技术主要基于无线电波和声波传输实现。无线电波在水下传输时,由于衰减迅速,导致定位不准确。而水声定位技术可以克服这一缺点,实现水下稳定传输。近年来,基于水声定位的水下机器人定位技术引起了广泛的研究。根据声学基线之间的距离,可以把声学定位系统分为 3 类:长基线定位(LBL,基线长度 100 ~ 6 000 m,甚至有时超过 6 000 m)、短基线定位(SBL,基线长度 20 ~ 50 m)和超短基线定位(USBL,基线长度小于 10 cm)[2]。杨放琼等[12]利用传统的长基线定位系统,实现了深海采矿 ROV 的精确定位。佛罗里达大西洋大学海洋工程系和海军研究所在 AUV 上装备了短基线定位系统,用来在侦察任务中进行导航[13]。Mandić 等[14]描述了利用超短基线定位与声呐图像测量追踪水下目标的技术发展。由于长基线定位系统复杂且昂贵,超短基线定位系统校准困难,本文采用短基线定位搭建机器人的水下定位系统。

1.3 水下目标定位

常用的水下目标定位方法有 3 种:基于单目的目标定位、基于双目或者多目的目标定位及基于多传感器融合的目标定位[15-18]。Marani 等[15]通过曲线拟合得到球体目标的圆形轮廓,结合已知的球形半径信息就可以进行目标的三维定位。Zannatha 等[16]利用 Canny 边缘检测算子提取目标的边缘,利用相似三角形原理完成了机器人的目标定位。Lee 等[17]为 ROV 的 2 个机械手开发了双目定位视觉系统,采用 2 个摄像机对目标进行三维定位,协助 2 台机械手协同工作。Zhang 等[18]研究了利用激光传感器与摄像机相结合的三维定位方法,通过相机的焦距等参数,完成水下球体目标的三维定位。本文采用多传感器融合的目标定位方式,利用声呐获取到机器人距离地面

的参数,配合相机获取目标物三维坐标。

2　基于卷积神经网络的水下目标检测

水下机器人的目标检测任务描述为:水下机器人潜入水中航行,并通过搭载的水下相机实时捕捉视频图像;利用检测算法对生成的图像进行实时检测,得到图像中目标的类别及其在图像中的包围框坐标信息;检测结果反馈给机器人。

由于光在水下的折射和散射作用,导致水下成像存在偏色、模糊等问题,这些问题直接影响水下图像特征提取的丰富程度,导致目标检测精度降低。而卷积神经网络在图像特征提取方面优势明显,因此,本文将卷积神经网络引入到水下场景来解决水下目标检测问题。通过对比当下主流的目标检测算法,本文采用检测速度相对较快且检测效果最好的 DSOD 检测算法进行水下目标检测。

DSOD 是一个端到端的多尺度检测模型,其不需要提取候选框,也不需要预训练过程,能够从头开始训练模型,并达到很好的检测效果。DSOD 的网络结构主要分为 2 部分:用于特征提取的主干网络和多尺度预测网络。

主干网络是 DenseNet 网络的一个变体,其由 1 个主干网络、4 个密集块、2 个转换层和 2 个转换池化层组成。主干网络包括 3 个 3×3 卷积层和 1 个 2×2 最大池化层,这种结构可以减少原始输入图像的信息丢失。密集块中所有前面的层都会连接到当前层,其优势是后面的层可以接受前面网络层的监督信息。每个转换层都包含一个池化操作,用于对特征图进行下采样。而转换池化层可以在不降低特征图分辨率的情况下增加密集块的数量。

预测网络利用精细的密集结构来融合多尺度的预测响应。输入图像经过 DSOD 网络可以提取出 6 个不同尺度的特征图做预测。从第 2 个特征图到第 6 个特征图,每一个特征图有一半的特征是从上一个特征图学到,而剩下的一半特征是对连续的高分辨率特征图进行下采样得到。这种预测结构可以用较少的参数产生更精确的结果,甚至对图像中的小目标也能产生很好的检测效果。

DSOD 网络结构及模型训练过程如图 1 所示。先利用水下图像数据集从头开始训练 DSOD 模型,经过反复迭代训练,可以得到一个效果比较好的网络模型,该模型能够产生稳定的参数,满足目标检测的性能要求。训练好的网络模型集成到水下机器人的检测算法中,利用机器人实时检测水下目标,并产生目标类别信息及包围框位置坐标,这些检测信息将进一步影响机器人的水下运动情况。

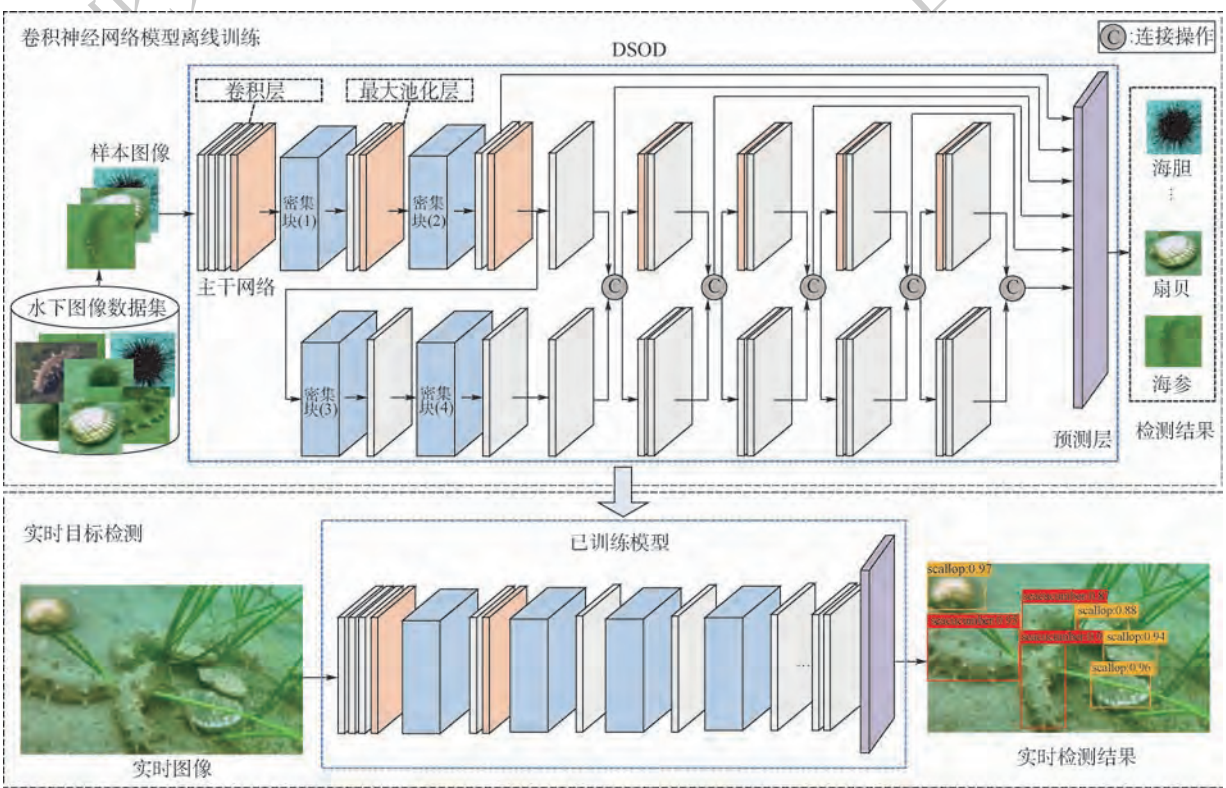


图 1　卷积神经网络模型的训练和水下目标实时检测

Fig. 1　CNN model training and real-time marine object detection

3 水下目标定位方法

水下目标定位是一个非常复杂的工作。首先,利用检测算法通过相机检测目标;然后,利用短基线定位系统定位机器人;最后,将目标坐标从相机成像坐标系转换到机器人世界坐标系,再从机器人世界坐标系转换到短基线定位系统的坐标系。通过分析不同坐标系之间的关系,本文提出了一种坐标转换方法来计算目标物在定位系统中的实际位置坐标。

3.1 水下定位系统的原理及搭建

短基线定位系统是一种经典的水下定位技术,其采集各个阵元接收到发射器发射信号的时间延迟,通过数学方程获取发射器相对于基阵坐标系的三维坐标。

如图 2 所示,4 个阵元分别位于 A、B、C、O,目标 P 与各阵元之间的距离为

$$R_i^2 = (x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2 \tag{1}$$

式中:R_i 为各阵元与发射器之间的距离;(x_i, y_i, z_i) 为短基线阵元 i 的位置坐标;(x, y, z) 为搭载发射器的水下机器人的坐标。

目标 P 与各阵元之间通过水声传播的距离表达式为

$$R_i = c(t_i - t - \Delta t) \tag{2}$$

式中:c 为声音在水中的传播速度;t 为短基线定位系统地面站发送询问信号的时刻;t_i 为短基线定位系统各阵元接收到信号的时刻;Δt 为短基线定位系统内的延时时间。

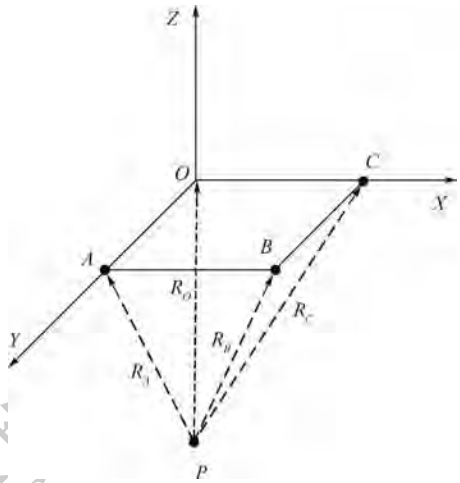


图 2 水声定位几何原理

Fig. 2 Geometric principle of underwater sound positioning

联立式(1)和式(2)得到

$$c^2(t_i - t - \Delta t)^2 = (x_i - x)^2 + (y_i - y)^2 + (z_i - z)^2 \tag{3}$$

将目标 P 与阵元坐标代入式(3)可以得到一个方程组,该方程组有 4 个未知量,因此,利用 4 个方程可以得到位置 (x, y, z) 和时延 Δt。

本文搭建的短基线定位系统如图 3 所示。图中:z_0 为机器人相对于地面的高度。将发射器集成到水下机器人上,作为发出信号的信标。在水面附近,放置 4 个短基线阵元,短基线阵元作为接收器监听发射器发出的信号。定位系统通过计算发射器发出信号到每个短基线阵元接收到信号的时间差来计算携带发射器的机器人的位置。

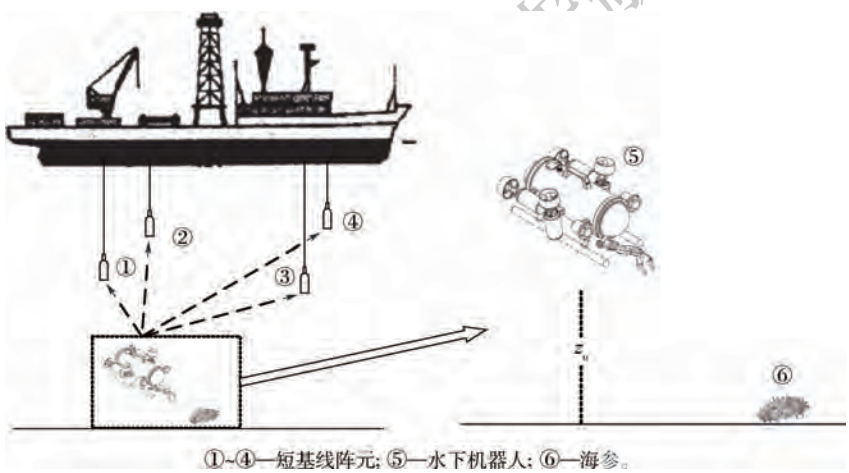


图 3 机器人与目标在定位系统中的位置关系

Fig. 3 Relative location of robot and target in positioning system

3.2 坐标转换方法

水下定位系统搭建成功后,通过分析相机成像坐标系与机器人世界坐标系之间的关系,以及机器人世界坐标系与定位系统坐标系之间的关

系,可以将目标坐标从相机成像坐标系经由机器人世界坐标系转换到定位系统坐标系,具体计算过程如下:

1) 利用 DSOD 检测算法检测水下目标在相机成像坐标系中的坐标(u,v)。

2) 通过声呐获取机器人距离地面的高度z_o。

3) 通过相机内参矩阵及高度z_o将目标从相机成像坐标系转换为机器人世界坐标系,进而计算出物体在机器人世界坐标系中的实际位置。

相机的内参矩阵定义为

$$z_o\begin{bmatrix}u\\v\\1\end{bmatrix}=\begin{bmatrix}\frac{f}{dx}&0&u_0\\0&\frac{f}{dy}&v_0\\0&0&1\end{bmatrix}\begin{bmatrix}x_o\\y_o\\z_o\end{bmatrix}\tag{4}$$

式中:f为相机的焦距;dx、dy为像元尺寸;(u_0,v_0)为图像中心像素坐标;(x_o,y_o,z_o)为目标在机器人世界坐标系的实际位置。由此计算出(x_o,y_o,z_o),即

$$\begin{bmatrix}x_o\\y_o\\z_o\end{bmatrix}=\begin{bmatrix}\frac{f}{dx}&0&u_0\\0&\frac{f}{dy}&v_0\\0&0&1\end{bmatrix}^{-1}z_o\begin{bmatrix}u\\v\\1\end{bmatrix}\tag{5}$$

4) 通过短基线定位系统获取机器人在定位系统坐标系中的坐标$F(x,y,z)$。

5) 根据机器人世界坐标系与定位系统坐标系之间的关系,将目标坐标转换到定位系统坐标系。如图 4 所示,XOY表示短基线定位系统坐标系,MFN表示机器人世界坐标系,α为 2 个坐标系之间的夹角。F为机器人在短基线定位系统坐标系中的坐标,T为目标在机器人世界坐标系中的坐标。

根据 2 个坐标系之间的角度关系,可以得出坐标转换表达式为

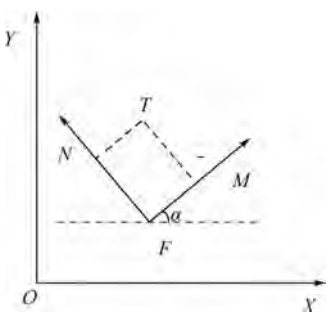


图 4　机器人世界坐标系和定位系统坐标系的关系

Fig. 4　Relation between robot world coordinate system and positioning system coordinate system

$$\begin{bmatrix}x_f\\y_f\\z_f\end{bmatrix}=\begin{bmatrix}\cos\alpha&-\sin\alpha&0&x\\\sin\alpha&\cos\alpha&0&y\\0&0&1&z\end{bmatrix}\begin{bmatrix}x_o\\y_o\\z_o\\1\end{bmatrix}\tag{6}$$

式中:(x,y,z)为机器人的坐标;(x_f,y_f,z_f)为目标物体T在定位系统中的坐标。

4　基于反馈机制的多信号分析方法

水下机器人的自主运动是实现自主抓取的前提和关键,其目的在于检测并接近水下目标,最终进行抓取。机器人需要综合考虑当前所处环境,做出合理的运动选择。水下环境的不确定性和复杂性给机器人自主运动问题带来了巨大挑战。

机器人利用水下定位系统探测目标的运动轨迹,如图 5 所示。利用水下定位系统规划机器人的运动区域,运动区域的边界附近代表警示区域。机器人从出发区开始,寻找水下目标。机器人的水平运动有 3 种策略:左转、右转和直行,不同的轨迹颜色表示从起点出发的不同运动策略。当机器人进入警示区域时,需要调整运动路线,远离警示区域并继续移动,直到检测到目标为止。当机器人检测到目标时,需要控制机器人继续靠近目标方向,直到到达机械手可操作的行程范围,利用机械手抓取水下目标。

根据以上分析,本文提出了一种基于反馈机制的多信号分析方法,如图 6 所示。该方法通过收集机器人的定位信息、深度信息,以及利用目标检测结果信息,对机器人的运动状态进行综合

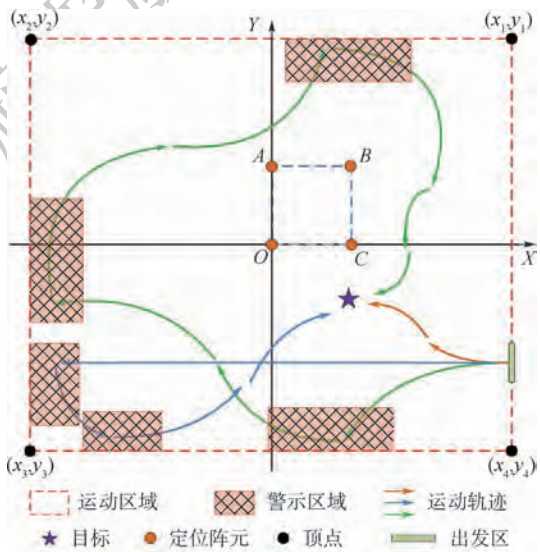


图 5　水下机器人在限制区域内的运动轨迹

Fig. 5　Trajectory of underwater robot moving within constrained area

分析,得出最终的移动方案,并将最终决策信号反馈给水下机器人,引导机器人继续运动。

多信号分析方法的决策流程如图 7 所示。首先,通过短基线定位系统获得机器人的位置信息,并判断机器人是否驶入运动区域的限制区域。如果机器人已经进入限制区域,则控制机器人向左前方运动,直到远离限制区域。然后,获取目标检测算法对当前机器人视觉区域的目标检测结果,如果当前视觉区域没有目标,则控制机器人向正前方运动;否则,根据检测结果的目标分布情况,判断机器人应该向正前方、左前方或右前方运动。根据目标定位算法可以进一步计算出目标在水下定位系统中的位置,从而精确机器人的移动方向。最后,利用机器人搭载的声呐向海底发射脉冲信号,根据响应时间计算距离海底的深度信息,判断机器人能否继续下潜。如果可以继续下潜,则机器人除了水平方向的运动之外还应该伴随下潜运动,即机器人应该向正前下、左前下或右前下的方向运动。多信号分析方法的决策结果会及时反馈给机器人并控制机器人的下一步运动,而机器人完成一次运动操作后会继续利用多信号分析方法获取下一次移动方案。

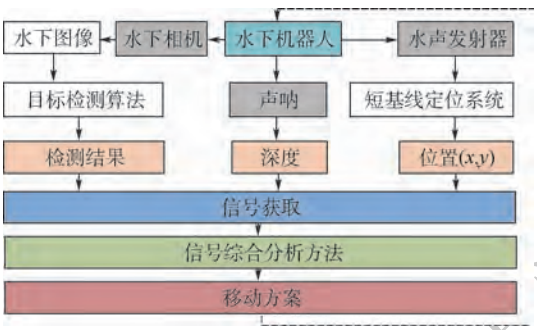


图 6 基于反馈机制的多信号分析方法

Fig. 6 Multi-signal analysis method based on feedback mechanism

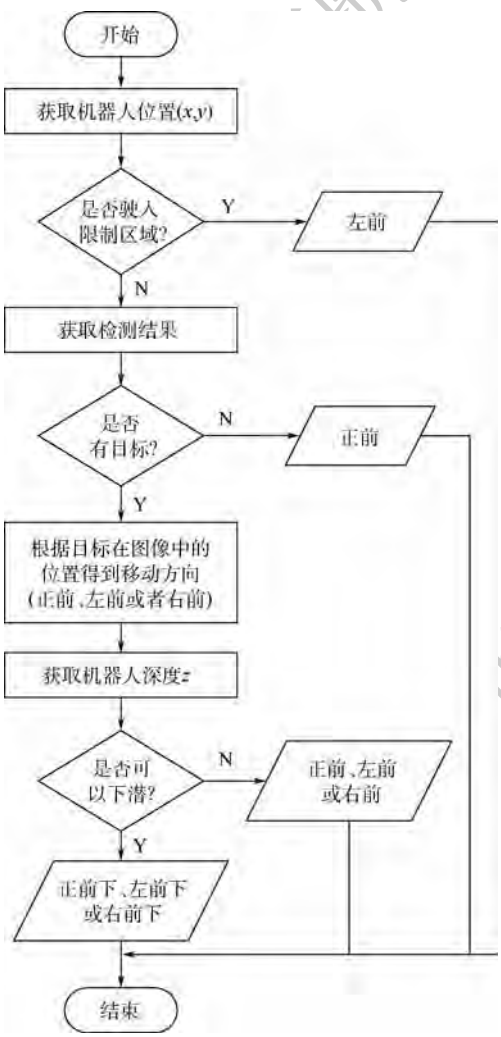


图 7 多信号分析方法流程

Fig. 7 Flowchart of multi-signal analysis method

目标检测结果除了用来计算目标在水下定位系统中的位置坐标之外,本文还定义了一种通信协议,用来帮助机器人理解目标在当前视野中的分布情况。如图 8 所示,将图像分为 9 个区域。根据检测结果,可以分别统计每个区域内各个类别的目标数目,并将统计结果按照顺序拼接成检测信号发送给机器人,该检测信号可以帮助机器人分析朝哪个方向运动更接近目标。

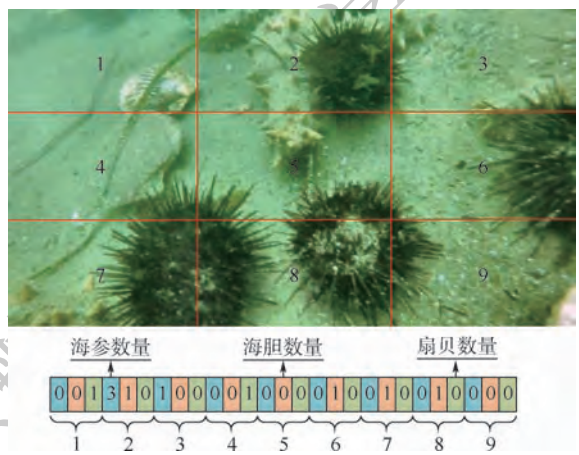


图 8 检测模块和机器人之间的通信协议

Fig. 8 Communication protocol between detection module and robot

5 实验与结果

为了验证本文算法,设计和开发了一款水下抓取机器人,将算法应用到机器人上,实现机器人的智能检测和自主抓取。

5.1 水下机器人设计

水下敏捷机器人的自主抓取研发需要解决许多水下运行环境所带来的问题。如图9所示,本

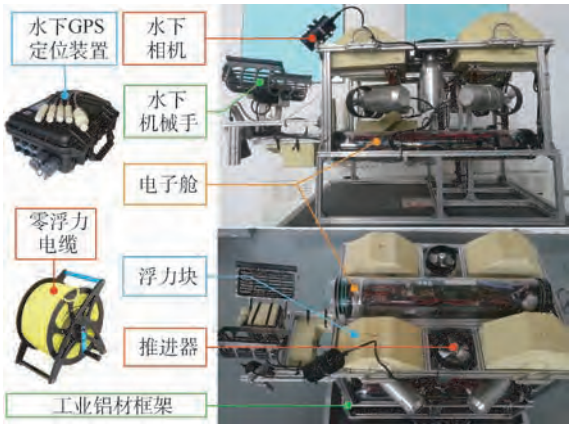


图 9　水下机器人组成结构

Fig. 9　Underwater robot composition structure

文设计的机器人是一款 6 推进器的 ROV,其搭载了 3 个亚克力舱室作为电子部件的防水处理,1 个作为主舱保护树莓派、Pixhawk 控制板等核心电子器件,2 个作为电池仓放置电压供电模块。由于海水腐蚀性很强,选择具有防腐蚀性的工业铝搭建机器人主体框架。

Pixhawk 控制板主要用来收集并处理机器人各部件发出的信号,树莓派是机器人的通信中心。6 个推进器协调工作实现机器人的上浮、下潜、左右滑动及转向控制。由于水下能见度低,为了提高图像采集的质量,选用 4 K 高清水下相机,并搭载亮度可调的水下光源作为补充。

水下机器人在水下作业需要克服重力影响,本文机器人通过亚克力舱室及浮力块提供的浮力来维持自身平衡,与水面供电及信号通信的电缆都选用零浮力线缆。

水下定位系统的 4 个接收器搭建在水面,发射器搭载到机器人上,发射器与接收器之间通过发射脉冲信号通信。声呐安装在机器人的底部中心位置,竖直向下发射声波信号。

为了抓取海产品,设计了一款二自由度的机械手,搭载在机器人的前方。机械手一组完整的抓取动作如图 10 所示。

本文设计的机器人经过反复水下实验,已经能够在真实海洋环境中实现海产品的自主抓取。

5.2　海产品检测

对于基于卷积神经网络的目标检测任务,数据集的选取非常重要。Imagenet、COCO、Pascal VOC 等公共数据集,体积庞大,虽然种类丰富,但是扇贝、海参、海胆等海产品的图片数量并不充足。因此,笔者标注了水下图像数据集用于训练检测网络,数据集包含海参、海胆、扇贝 3 类海产品共 29 500 张图片。

本文目标检测算法在 caffe 框架上实现,实验环境使用 Nvidia Geforce GTX 1080 TI GPU、Cudnn v5.1 及 Intel Core i7-6700k@4.00 GHz。初始学习率设置为 0.1,每迭代 20 000 次学习率降低 10%,迭代 100 000 次停止训练。经过多次训练,本文目标检测算法可以在水下图像数据集上获得 81.2% 的准确率。

目前,比较主流的检测算法在水下图像数据集上的检测效果对比如表 1 所示。SSD300[9]、YoLov3[19]、DSOD[1] 的检测速度都能够达到 10 帧/s 以上,其中 YoLov3[19] 检测速度最快,DSOD[1] 的检测效果最好。

对于现阶段的水下自主抓取机器人,10 帧/s 以上的处理速度已经能够满足实际作业要求,为了追求更高的目标检测准确率,选择 DSOD 作

图 10　机器人完整的抓取动作

Fig. 10　Robot's complete capture action

表 1　水下图像数据集检测结果对比

Table 1　Comparison of detection results on underwater image dataset

检测算法	网络模型	速度/(帧·s^{-1})	准确率/%
Faster R-CNN[5]	VGGNet	7	79.6
SSD300[9]	VGGNet	47	77.9
YoLov3[19]	Darknet-53	71	78.4
DSOD[1]	DS/64-192-48-1	17	81.2

为本文系统的检测算法。DSOD 目标检测的部分效果如图 11 所示,图中不同颜色的标签代表不同类别,每个标签显示了检测算法判定该目标所属的类别及属于该类别的概率。

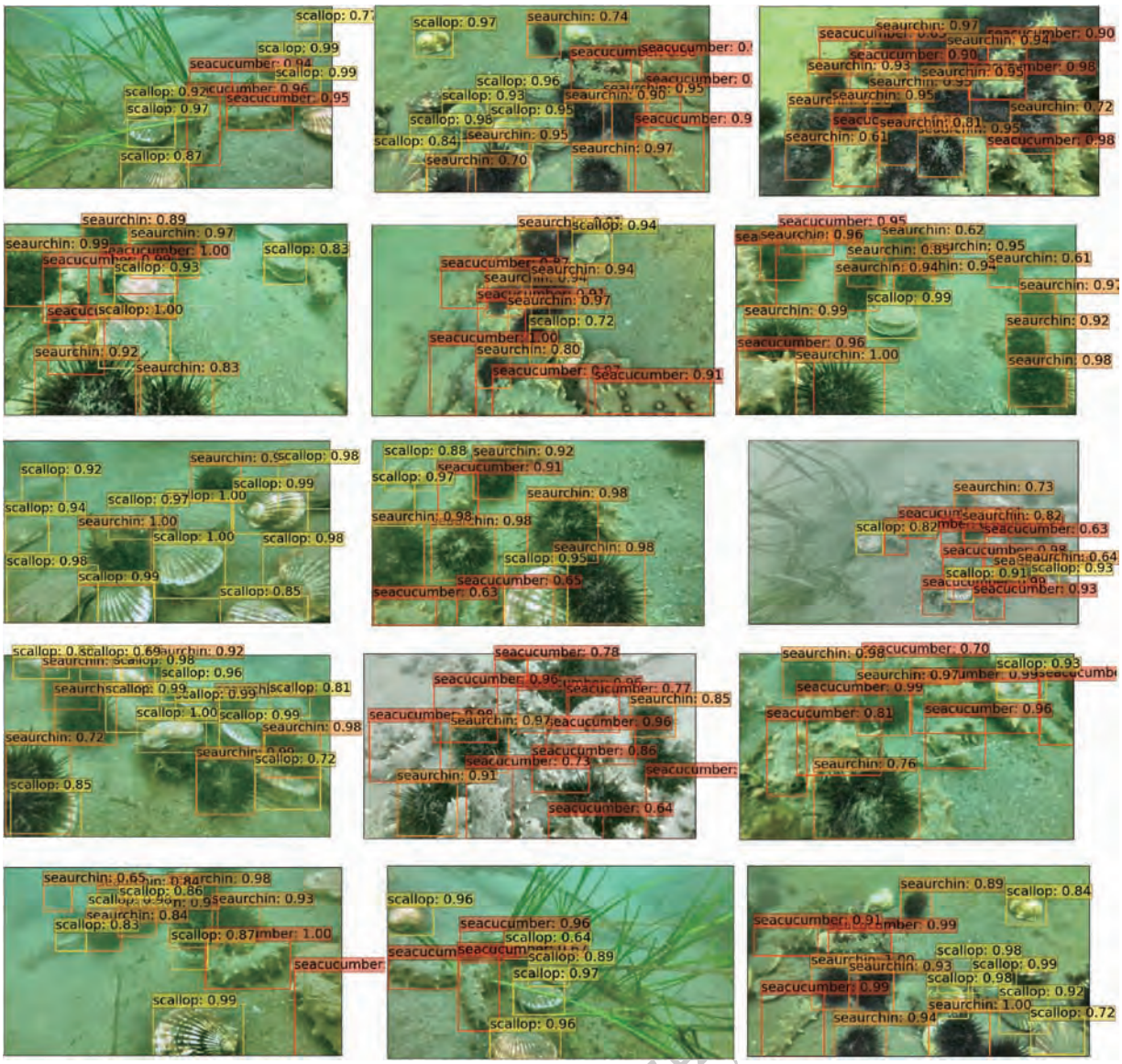


图 11 水下目标检测结果

Fig. 11 Detection results of marine object

5.3 系统性能分析

本文设计的海产品智能检测与自主抓取系统经过深入实际调研,考虑了机器人实际作业的场景,能够在真实海洋环境中实时检测目标,并引导机器人进行海产品的自主抓取作业。但是在实际抓取效率方面,搭载该系统的水下机器人尚未达到人工捕捞的效率。原因主要有:①水下环境异常复杂,机器人在实际作业时容易受到暗流、洋流、礁石等因素的干扰;②系统及机器人还存在进一步改进和完善的空间。

6 结 论

本文提出一种智能检测和自主抓取系统,用来协助机器人解决水下目标的智能检测问题并引导机器人自主抓取海产品。

1)针对水下目标检测问题,将 DSOD 算法应用到水下场景,利用标注的水下图像数据集训练检测模型来检测海参、海胆、扇贝等海产品。

2)为了计算海产品的实际位置,搭建了短基线定位系统来定位水下机器人,并通过分析相机成像坐标系与水下定位系统坐标系之间的关系,提出一种计算海产品实际位置的坐标变换方法。

3)针对水下机器人的自主运动问题,设计了一种基于反馈机制的多信号分析方法,用来引导机器人在水下移动并抓捕海产品。为了验证本文算法的有效性,搭建了一款水下抓捕机器人,并成功将算法应用到机器人进行海产品的智能检测和自主抓取。

参考文献（References）

[1] SHEN Z,LIU Z,LI J,et al. DSOD:Learning deeply supervised object detectors from scratch[C]// IEEE International Conference on Computer Vision(ICCV). Piscataway,NJ:IEEE Press,2017:1937-1945.

[2] VICKERY K. Acoustic positioning systems. A practical overview of current systems[C]// Proceedings of the 1998 Workshop on Autonomous Underwater Vehicles. Piscataway,NJ:IEEE Press,1998:5-17.

[3] GIRSHICK R,DONAHUE J,DARRELL T,et al. Rich feature hierarchies for accurate object detection and semantic segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE Press,2014:580-587.

[4] GIRSHICK R. Fast R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway,NJ:IEEE Press,2015:1440-1448.

[5] REN S,HE K,GIRSHICK R,et al. Faster R-CNN:Towards real-time object detection with region proposal networks[C]// Proceedings of the 29th International Conference on Neural Information Processing Systems. Cambridge:MIT Press,2015:91-99.

[6] HE K,GKIOXARI G,DOLLÁR P,et al. Mask R-CNN[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway,NJ:IEEE Press,2017:2961-2969.

[7] DAI J,LI Y,HE K,et al. R-FCN:Object detection via region-based fully convolutional networks[C]// Proceedings of the 30th International Conference on Neural Information Processing Systems. Cambridge:MIT Press,2016:379-387.

[8] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:Unified,realtime object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE Press,2016:779-788.

[9] LIU W,ANGUELOV D,ERHAN D,et al. SSD:Single shot multibox detector[C]// European Conference on Computer Vision. Berlin:Springer,2016:21-37.

[10] FU C Y,LIU W,RANGA A,et al. DSSD:Deconvolutional single shot detector[EB/OL]. (2017-01-23)[2019-07-01]. https://arxiv.org/abs/1701.06659.

[11] SUN X,SHI J,LIU L,et al. Transferring deep knowledge for object recognition in low-quality underwater videos[J]. Neurocomputing,2018,275:897-908.

[12] 杨放琼,谭青,彭高明. 基于长基线系统深海采矿 ROV 精确定位[J]. 海洋工程,2006(3):95-99.
YANG F Q,TAN Q,PENG G M. Accurate positioning for ROV in deep-sea mining based on LBL acoustic system[J]. The Ocean Engineering,2006(3):95-99(in Chinese).

[13] SMITH S M,KRONEN D. Experimental results of an inexpensive short baseline acoustic positioning system for AUV navigation[C]// Proceedings of IEEE Oceans'97. Piscataway,NJ:IEEE Press,1997,1:714-720.

[14] MANDIĆ F,RENDULIC I,MISKOVIC N,et al. Underwater object tracking using sonar and USBL measurements[J]. Journal of Sensors,2016,2016:1-10.

[15] MARANI G,CHOI S K,YUH J. Experimental study on autonomous manipulation for underwater intervention vehicles[C]// The Seventeenth International Offshore and Polar Engineering Conference,2007:1088-1094.

[16] ZANNATHA J M I,LIMON R C,SANCHEZ A D G,et al. Monocular visual self-localization for humanoid soccer robots[C]//21st International Conference on Electrical Communications and Computers. Piscataway,NJ:IEEE Press,2011:100-107.

[17] LEE P M,JEON B H,HONG S W,et al. System design of an ROV with manipulators and adaptive control of it[C]// Proceedings of the 2000 International Symposium on Underwater Technology. Piscataway,NJ:IEEE Press,2000:431-436.

[18] ZHANG M,ZHANG L,LI Y. A three-dimensional locating method for underwater robot based on sensors fusion[C]// 2009 International Conference on Mechatronics and Automation. Piscataway,NJ:IEEE Press,2009:1207-1212.

[19] REDMON J,FARHADI A. YoLov3:An incremental improvement[EB/OL]. (2018-04-08)[2019-07-02]. https://arxiv.org/abs/1804.02767.

作者简介:

徐凤强　男,博士研究生。主要研究方向:数字图像处理、水下目标检测、计算机视觉。

董鹏　男,硕士研究生。主要研究方向:数字图像处理、计算机视觉。

王辉兵　男,博士。主要研究方向:计算机视觉、机器学习。

付先平　男,博士,教授,博士生导师。主要研究方向:多媒体技术、驾驶行为、模式识别、数字图像处理和视频图像低码率编码。

Intelligent detection and autonomous capture system of seafood based on underwater robot

XU Fengqiang, DONG Peng, WANG Huibing, FU Xianping*

(Information Science and Technology College, Dalian Maritime University, Dalian 116026, China)

Abstract: Currently, underwater robot faces the tough challenges of lacking intelligent detection and autonomous capture system to guide. Therefore, autonomous capture is hard to be achieved. Toward this end, this paper proposes an intelligent detection and autonomous capture system to achieve intelligent detection of marine target and guide the underwater robot to autonomously capture seafood. First, we employ convolutional neural network to perform object detection task in underwater scene and train the DSOD with underwater dataset to accurately detect marine objects. What's more, the short baseline positioning system is built to locate the underwater robot. To calculate the position of the object relative to robot, this paper proposes a coordinate transforming method to transform the target's location from camera coordinates system to underwater positioning coordinates. Furthermore, this paper designs a multi-signal analysis method based on feedback mechanism to command the robot to move ahead to the seafood until grasping them. To verify the effectiveness of the system, we develop an underwater picking robot and successfully apply the proposed methods to the robot to autonomously detect and capture the marine object.

Keywords: underwater picking robot; intelligent detection; marine target positioning; short baseline positioning system; autonomous capture

Received: 2019-07-09; **Accepted**: 2019-08-20; **Published online**: 2019-08-26 15:22

URL: kns.cnki.net/kcms/detail/11.2625.V.20190826.1509.003.html

Foundation items: National Natural Science Foundation of China (61272368, 61370142); the Fundamental Research Funds for the Central Universities (3132016352); the Fundamental Research of Ministry of Transport of P. R. China (2015329225300); Dalian Science and Technology Innovation (2018J12GX037)

* **Corresponding author**. E-mail: fxp@dlmu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0369

基于图像纹理的自适应水印算法

黄樱[1], 牛保宁[1,*], 关虎[2], 张树武[2]
(1. 太原理工大学 信息与计算机学院, 太原 030024; 2. 中国科学院自动化研究所 数字内容技术与服务研究中心, 北京 100190)

摘 要: 图像水印技术是一种在图像中嵌入被称为水印的版权标记,以证明图像版权归属的技术。利用图像纹理粗糙区域易于隐藏水印的优势,提出了一种基于图像纹理的自适应水印算法。首先,设计了一种能够真实反映图像纹理丰富程度的纹理度量方法,引入全局纹理值和局部纹理值的概念来综合分析图像纹理的分布情况;其次,利用滑动窗口和窗口内区域的局部纹理值,精确地获取图像的纹理粗糙区域,将水印嵌入在纹理粗糙区域中,保证嵌入水印图像的视觉质量;然后,通过多元回归分析,得到水印嵌入参数与纹理粗糙区域的全局纹理值和局部纹理值的函数关系,根据区域的纹理值自适应地调整水印的嵌入参数,最大限度地保证水印的不可见性,增强水印的鲁棒性;最后,通过在多个不重叠的纹理粗糙区域中嵌入相同的水印,进一步提高水印提取的准确率。在100幅自然场景图像上进行模拟实验,从不可见性、自适应性和鲁棒性三个方面证实了所提算法相比已有自适应水印算法的优越性。

关 键 词: 图像水印; 图像纹理; 不可见性; 鲁棒性; 自适应性
中图分类号: TP301.6
文献标识码: A **文章编号**: 1001-5965(2019)12-2403-12

数字媒体内容的使用和传播在给人们带来方便的同时,也引发了严重的网络版权问题。图像水印技术为数字图像版权的查证提供了技术支撑。它将一个代表版权的数字标记不可察觉地嵌入到图像中(该标记被称为水印),当发生版权纠纷时,可以通过提取水印来确定图像的所有权。

一个好的水印算法必须使嵌入的水印兼备良好的不可见性和鲁棒性。不可见性是指水印嵌入到图像后不能影响其质量和使用;鲁棒性是指即使图像被各种攻击(噪声、压缩、缩放、剪切等)所改变,水印也能被正确地提取。水印的不可见性和鲁棒性由水印的嵌入强度决定。增大水印的嵌入强度有利于提高水印的鲁棒性,但同时会降低水印的不可见性;反之亦然。因此,鲁棒性与不可见性之间存在矛盾:增加鲁棒性通常以牺牲不可见性为代价,两者相互制约。

一幅图像通常包含纹理平滑区域(单位空间上灰度变化平缓)和纹理粗糙区域(单位空间上灰度变化剧烈)。观察嵌入水印后的图像可以发现,嵌入在纹理平滑区域中的水印相比嵌入在纹理粗糙区域中的水印更容易让人眼察觉。这说明纹理平滑区域和纹理粗糙区域对水印的承受力是不同的。然而目前的水印技术[1-15]没有利用这一事实,通常它们为避免影响图像的视觉质量(往往取决于纹理平滑区域的视觉质量),只能减小水印的嵌入强度,同时也降低了水印的鲁棒性。

在水印嵌入时通常利用一个嵌入参数来控制水印的嵌入强度,嵌入参数的设定需要同时兼顾鲁棒性和不可见性。大多数自适应水印技术[1-4]利用图像的某些特性(均值[2]、方差[3-4]、熵[1]等)相应调整嵌入参数,它们虽然可以平衡不同图像中水印的不可见性之间的差异,但由于这些特性

收稿日期: 2019-07-09; **录用日期**: 2019-08-20; **网络出版时间**: 2019-08-26 15:08
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190826.1052.001.html
基金项目: 国家重点研发计划(2017YFB1401000); 山西省重点研发计划(国际合作)项目(201603D421015)
* **通信作者**. E-mail: niubaoning@tyut.edu.cn

引用格式: 黄樱, 牛保宁, 关虎, 等. 基于图像纹理的自适应水印算法[J]. 北京航空航天大学学报, 2019, 45(12): 2403-2414.
HUANG Y, NIU B N, GUAN H, et al. Image texture based adaptive watermarking algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2403-2414 (in Chinese).

与水印的不可见性没有直接关联,在提高水印鲁棒性的同时无法保证所有图像中嵌入的水印拥有一致的不可见性。现有的自适应水印算法[5-6]将嵌入参数与峰值信噪比(PSNR,不可见性的评估指标之一)相关联,可使所有含水印图像获得基本一致的 PSNR,然而,PSNR 是基于误差敏感的图像质量评价,存在评价结果与人的主观感觉不一致的情况。

为解决这些问题,本文从图像频率(表征图像灰度变化剧烈程度的指标)角度分析,利用图像纹理粗糙区域易于隐藏水印的优势,提出了一种基于图像纹理的自适应水印算法,将其命名为 TBAQT(Texture-based Adaptive Quantization Threshold)。

本文创新点如下:

1)提出了一种纹理粗糙度——纹理值的度量方法,用该方法计算得到的纹理值能够真实地反映图像的纹理情况;并引入全局纹理值和局部纹理值的概念,分别用于表征图像(区域)整体和局部纹理的丰富程度。

2)利用滑动窗口及窗口内区域的局部纹理值,通过适当地设置滑动窗口的尺寸和移动步长,精确地获取图像的纹理粗糙区域;将水印嵌入在图像的纹理粗糙区域中,以保证嵌入水印后的图像质量;通过同时在多个纹理粗糙区域中嵌入相同的水印,提高提取水印的准确率。

3)运用多元回归分析得到水印嵌入参数关于纹理区域的全局纹理值和局部纹理值的函数关系,通过该函数,可根据嵌入区域的纹理值自适应地调整相应区域的水印嵌入参数,在保证每幅图像中嵌入的水印都能获得基本一致的不可见性的前提下,最优化每个嵌入区域的嵌入参数,最大限度地提高水印的鲁棒性。

结合算法的整体设计,本文算法能够有效抵抗各种图像处理攻击及几何攻击。

1 相关工作

在图像水印技术中,水印的嵌入强度通常由一个嵌入参数来控制,设置合适的嵌入参数可以平衡水印的鲁棒性和不可见性。非自适应水印技术[12-13]在为每幅图像嵌入水印时都使用一个凭经验得到的固定的嵌入参数,无法保证每幅图像中嵌入的水印都具有良好的不可见性。因为水印的不可见性还与载体图像本身相关,导致不同图像中水印的不可见性存在差异。相对地,自适应水印技术[1-6]利用一个自适应调参模型,根据载体图像的相关信息自适应地调整水印的嵌入参数。相比非自适应水印技术,自适应水印技术可以减小不同图像中嵌入水印之间不可见性的差异,并在此基础上提高水印的鲁棒性。

如图 1 所示,一个完整的自适应水印技术主要包含三部分内容:嵌入区域的确定、自适应调参模型、水印嵌入和提取算法。

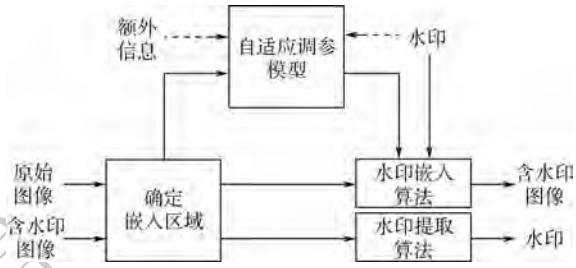


图 1 自适应水印技术框架

Fig. 1 Framework of adaptive watermarking technology

嵌入区域表示水印在图像中的嵌入位置。水印技术既可将一幅完整的图像作为水印的嵌入区域(全局水印技术[12-15]),也可仅在图像多个特定的区域中嵌入水印(局部水印技术[7-9])。由于全局水印技术在执行过程中利用了图像的全部信息,这就决定了其难以完全抵抗剪切攻击,更无法避开图像中的纹理平滑区域。局部水印技术通常利用特征点定位嵌入区域,它们将从图像中提取的某些具有平移、旋转或尺度不变性的特征点作为参考点,在其定位的非重叠的区域中嵌入水印。Tang 和 Hang[8] 使用 Mexican-Hat 小波尺度交互方法提取特征点,将水印嵌入到以特征点为中心的规范化圆形区域中,但该特征点对图像修改比较敏感。Nasir 等[7] 采用基于终止(end-stopped)小波的特征提取方法来提取重要的几何不变特征点,然后通过归一化特征点定位的水印嵌入区域来实现旋转不变性。Ye 等[9] 利用 SIFT 特征点的旋转和尺度不变性,将水印嵌入到 SIFT 特征点定位的圆形区域中。然而特征区域的定位及同步与特征点的稳定性紧密相关,图像被攻击前后的特征区域难以保证完全一致,这会影响水印提取的准确率;而且特征点并不会刻意捕获图像中的纹理粗糙区域,因此也难以避免特征区域中包含纹理平滑的部分。

如式(1)所示,自适应调参模型的实质是一个函数,自变量是一个或者多个与嵌入区域相关的参考值$(x_1,x_2,\cdots,x_n)$,因变量为嵌入参数(p)。

$$p = f(x_1,x_2,\cdots,x_n) \tag{1}$$

在嵌入水印时,通过该函数关系,可使每个嵌入区域的嵌入参数都能够跟随这些参考值自适应地改变。现有大多数自适应水印技术利用图像的

某些特性(均值[2]、方差[3-4]、熵[1]等)调整嵌入参数。Guan 等[2]使用待嵌入水印的系数的平均值来调整水印嵌入强度。Mohrekesh 等[3]利用图像块和邻域之间的局部方差自适应地改变每个块的嵌入参数。Yadav 和 Singh[4]利用具有较高信息熵的图像块的标准差来调节相应强度因子。Fazlali 等[1]提出嵌入参数由近似图像中图像块的边缘和熵决定。虽然这些自适应水印技术可以减小不同图像中嵌入水印的不可见性之间的差异,但是影响图像中水印不可见性的图像特征是多样的,它们难以考虑得面面俱到,因此,仍然无法保证每幅图像中嵌入水印的不可见性。近年来,Huang 等[5-6]提出一种可根据指定的 PSNR 自适应地调整水印的嵌入参数的模型,能够使嵌入水印的图像获得基本一致的 PSNR,然而 PSNR 是基于对应像素点误差的质量评价,对图像中的不同区域不具有区分性,评价结果可能与人的主观感觉不一致。

在水印算法中,空间域水印算法[10-11]通过直接修改载体图像的像素来嵌入水印;变换域水印算法[2,5-6,12-15]首先执行域变换,再通过修改变换域的系数来嵌入水印。变换域水印算法相比空间域水印算法具有更好的不可见性和鲁棒性。Parah 和 Loan 等[12-13]利用离散余弦变换(DCT)系数块之间的关联性,通过调整相邻块中低频系数的数值来嵌入水印,这些方法具有很高的水印容量,但它们都只进行了一级 DCT,因此对攻击的抵抗能力有限。Huynh-The 等[14-15]提出一种信道选择机制,将水印嵌入到 DWT 的最优信道,并在提取时利用大津法(OTSU)确定分割阈值,虽然该算法能够很好地平衡水印的不可见性和鲁棒性,但其对旋转和剪切攻击敏感。Guan 和 Huang 等[2,5-6]将扩频和量化方案的优势分别与两级 DCT 域中一些系数的稳定性相结合,使得水印能够抵抗常见的图像处理攻击和几何失真。近年来,深度学习备受关注,尤其是生成模型[16-20]的发展,为水印算法的研究提供了新的思路。但由于模型训练需要较大的时间开销,以及如何设计合适的训练策略等问题,此类方法还有待进一步探索。

为克服现有嵌入区域确定方法及自适应调参模型存在的缺陷,并结合 Huang 等[6]提出的差值量化水印算法的优势,本文提出了基于图像纹理的自适应水印算法,能够充分利用图像纹理粗糙区域易于隐藏水印的优势,在保证每幅图像中嵌入水印的不可见性的同时,最大限度地提高水印的鲁棒性。

2　图像的纹理粗糙区域

图像纹理在形式上存在广泛性和多样性[21],因此目前没有一个统一的定义,通常研究者会根据不同的应用提出相应的概念与定义。根据本文的实际需求,纹理粗糙区域/图像和纹理平滑区域/图像的定义如下,判定规则见 2.2 节。

定义 1　单位空间内灰度变化剧烈的区域/图像定义为纹理粗糙区域/图像。

定义 2　单位空间内灰度变化平缓的区域/图像定义为纹理平滑区域/图像。

图像的频率[22]是表征图像灰度变化剧烈程度的指标。图像可以分解成不同的频率成分,其中,低频成分表示灰度连续渐变的一块区域,描述图像的概貌和轮廓;高频成分是指图像灰度变化剧烈的地方,主要描述图像的细节和边缘。噪声是在图像上出现的一些随机的、离散的、孤立的像素点,因为它与附近的像素点灰度值明显不一样,导致灰度出现较大的变化,所以对应高频成分。根据以上定义可知,纹理粗糙区域应该包含更多的高频成分,而纹理平滑区域则包含更多的低频成分。

图 2 为采用文献[6]提出的水印算法嵌入水印后的图像,放大后可见嵌入在纹理平滑区域中的水印,但无法察觉嵌入在纹理粗糙区域中的水印(SSIM 为结构相似度,是不可见性的评价指标)。这是因为人眼对低频数据的敏感度高于高频数据[22],所以改变低频成分对图像视觉上的影响远大于高频成分。在图像中嵌入水印,相当于在图像上叠加了与之无关的噪声,若将水印嵌入在图像的纹理平滑区域,可能会使嵌入水印的像素点的灰度值明显异于邻域像素点,致使水印变得可见;相对地,纹理粗糙区域本身就属于灰度急剧变化的区域,因此能够使嵌入的水印不易被察觉。

在保证图像视觉效果的前提下,尽可能地提高水印的鲁棒性,是对水印算法的基本要求。

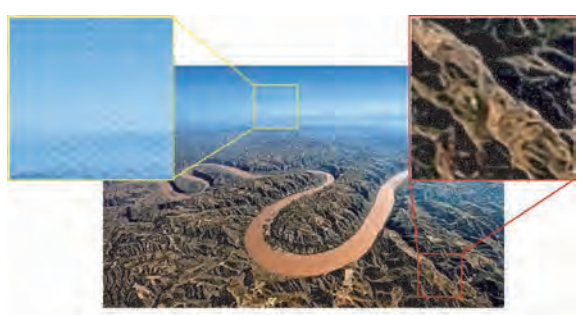


图 2　嵌入水印的图像

Fig. 2　Watermarked image

综上所述,相比纹理平滑区域,在相同嵌入强度下,水印嵌入在纹理粗糙区域具有更好的不可见性;相对地,在相同不可见性条件下,纹理粗糙区域可以容忍更大的嵌入强度。因此,本文考虑仅在纹理粗糙区域嵌入水印来同时提高水印的鲁棒性和不可见性,并且还能通过在多个纹理粗糙区域嵌入相同的水印来进一步提高鲁棒性。

2.1 图像纹理的度量

根据以上所述可知,图像纹理与图像的频率相关,这促使笔者联想到 DCT。DCT 的特点就是将图像中零散分布的频率成分变成有序分布。二维 DCT 正变换核为

$$g(x,y,u,v)=\frac{2}{\sqrt{MN}}C(u)C(v)\cos\left(\frac{(2x+1)u\pi}{2M}\right)\cdot\cos\left(\frac{(2y+1)v\pi}{2N}\right) \tag{2}$$

式中:x 和 y 分别为图像像素的横坐标和纵坐标;u 和 v 分别为 DCT 系数的横坐标和纵坐标;$x,u=0,1,\cdots,M-1$;$y,v=0,1,\cdots,N-1$;M 和 N 分别为图像的长和宽。

$$C(u)=\begin{cases}\sqrt{1/N} & u=0\\ \sqrt{2/N} & u\neq 0\end{cases}$$

$$C(v)=\begin{cases}\sqrt{1/M} & v=0\\ \sqrt{2/M} & v\neq 0\end{cases}$$

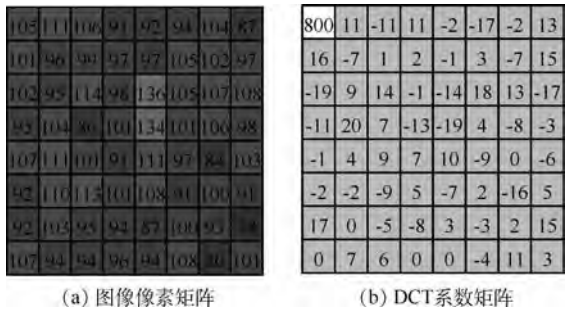
二维 DCT 定义如下:

$$\boldsymbol{F}(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}\boldsymbol{f}(x,y)g(x,y,u,v) \tag{3}$$

式中:$\boldsymbol{f}(x,y)$ 为 $M\times N$ 维的图像像素矩阵;$\boldsymbol{F}(u,v)$ 表示与图像像素矩阵对应的 DCT 系数矩阵。

根据二维 DCT 正变换核可知,随着 u、v 值的增大,余弦函数的频率越大,因此,图像的 DCT 系数矩阵的左上角代表低频系数,右下角表示高频系数。其中,$F(0,0)$ 与余弦函数无关($\cos 0=1$),它是图像抽样信号的均值,被称为 DCT 的直流系数(DC 系数),其他 DCT 系数都是由余弦函数参与得到,所以被称为交流系数(AC 系数)。DC 系数反映图像的灰度级别,与灰度值的分布无关;AC 系数与图像频率相关,可以用来反映图像的灰度变化程度,因此适合描述图像的纹理。

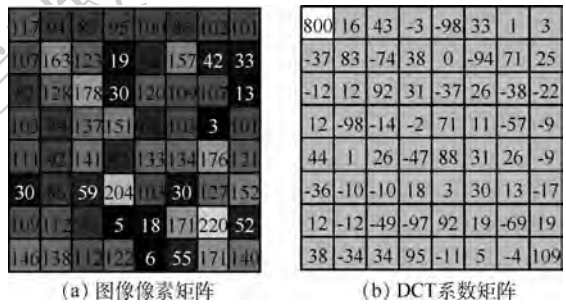
随机产生 2 组均值为 100、方差分别为 10 和 50、取值范围为 0~255 的随机整数,组成 2 个 8×8 大小的矩阵,分别用以模拟纹理平滑图像和纹理粗糙图像,如图 3(a)和图 4(a)所示,它们对应的 DCT 系数矩阵(仅保留整数)如图 3(b)和图 4(b)所示。尽管 2 幅图像 DCT 系数矩阵不同位置上的AC系数的绝对值有大有小,但从总体上来看,纹理粗糙图像的 AC 系数的绝对值普遍大于纹理平滑图像。这是因为纹理平滑图像的灰度值变化缓慢,图像信号在 2 组正交函数的投影值出现了大量正负相抵消的情景,从而导致 AC 系数在数值上较小。相反,纹理粗糙图像的灰度值变化剧烈,因此,发生正负抵消的情况要相对较少,得到的 AC 系数的数值也相对较大。



(a) 图像像素矩阵 (b) DCT系数矩阵

图 3 纹理平滑图像像素及 DCT 系数矩阵

Fig. 3 Pixel and DCT coefficient matrix of smooth image



(a) 图像像素矩阵 (b) DCT系数矩阵

图 4 纹理粗糙图像像素及 DCT 系数矩阵

Fig. 4 Pixel and DCT coefficient matrix of textured image

由此推断,纹理粗糙图像的 AC 系数绝对值之和必然会大于纹理平滑图像的 AC 系数绝对值之和,并且纹理越丰富,即图像的灰度值变化越剧烈,AC 系数绝对值之和会越大。因此,利用图像 DCT 的 AC 系数的绝对值之和能够很好地描述图像的纹理情况。

定义 3 将图像纹理粗糙程度的度量定义为图像的纹理值(T_v),用于反映图像纹理的丰富程度。

计算式为

$$T_v=\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}|\mathbf{AC}(i,j)|/(MN-1) \tag{4}$$

式中:$i\neq 0$ 或 $j\neq 0$;$\mathbf{AC}(i,j)$ 为位于第 i 行第 j 列的 AC 系数。

如图 5 所示,纹理值能够很好地反映图像的纹理粗糙程度,并且与人的直观视觉感受一致。

2.2 纹理粗糙区域的判定

为了获取图像的纹理粗糙区域,最原始的方法是直接求取某块区域的纹理值,若纹理值大于

阈值(T_t),则认为该区域是纹理粗糙区域。

定义 4　将一幅完整图像或一块完整区域的纹理值定义为该图像或者区域的全局纹理值(T_g)。

但这样做存在一个问题,如图 6 所示,如果某一区域大部分是纹理比较丰富区域,但仍然存在小块平滑的区域(见图 6(a)),它的全局纹理值也会比较大。虽然图 6(b)的全局纹理值小于图 6(a)的全局纹理值,但图 6(b)并不包含纹理平滑部分,因此更适合嵌入水印。

为了避免选取包含平滑部分的纹理粗糙区域,对区域进行平均划分,求取区域中每个分块的纹理值。

定义 5　将所有分块中的最小纹理值定义为该图像或区域的局部纹理值(T_l)。

在对区域进行划分时,可根据实际需求设置划分比例(P),分块数量越多,对区域的纹理要求越精细,这样虽然可以获取到纹理更为丰富的区域,但能够满足阈值条件的区域可能会越少,因此需要进行适当权衡。

图 7 给出了随机选取的 100 个区域的全局纹理值和局部纹理值,因为局部纹理值刻画纹理更为细致,所以对于一个区域而言,局部纹理值通常比全局纹理值小,因此,纹理粗糙区域的判定规则如下。

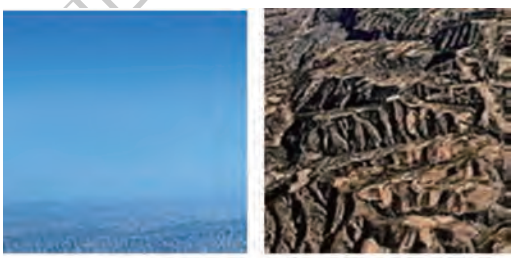


图 5　图像的纹理值

Fig. 5　Texture value of images

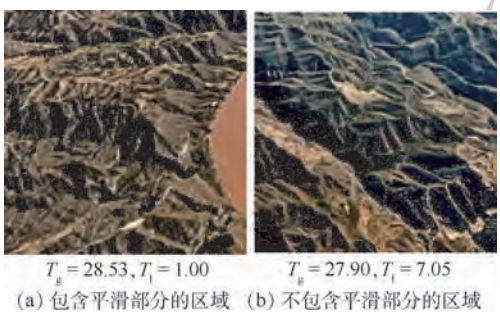


图 6　区域的全局纹理值和局部纹理值

Fig. 6　Global texture value and local texture value of a region

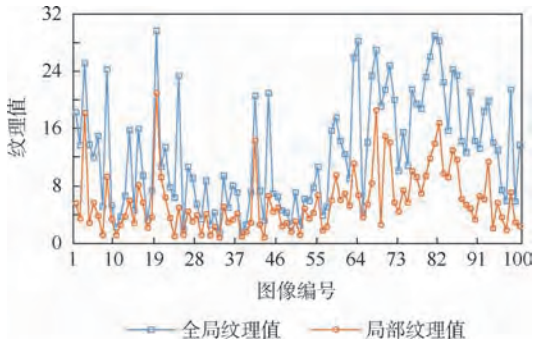


图 7　100 个区域的局部纹理值和全局纹理值

Fig. 7　Local texture values and global texture values of 100 regions

判定规则　若图像中某个区域的局部纹理值大于阈值,该区域可被判定为纹理粗糙区域。

如图 6 所示,利用局部纹理值不但能够反映区域的纹理情况,还能避免纹理粗糙区域中出现小块平滑区域。

2.3　水印嵌入区域的定位

利用滑动窗口按照一定步长遍历原始图像,计算每个窗口的全局纹理值(T_g)和局部纹理值(T_l);根据 2.2 节的判定规则判定每个窗口区域是否为纹理粗糙区域,保存纹理粗糙区域的位置信息与纹理值;运算结束后窗口向右或向下移动一个步长的位移,直到完成对整幅图像的处理。

本文要求滑动窗口的尺寸和移动步长应与原始图像尺寸成比例,以保证图像在缩放、剪切、翻转或者旋转 90°的整数倍后获取的纹理粗糙区域与原始图像的纹理粗糙区域保持一致。设置滑动窗口的大小与原始图像的比例为 W,移动步长与原始图像的比例为 S,可将 W 和 S 作为密钥,以增加算法的安全性。

并不是所有纹理粗糙区域都用于嵌入水印。如果 2 个纹理粗糙区域发生重叠,嵌入在其中的水印信息会相互干扰,从而影响到水印提取的正确性。因此,将水印的嵌入区域定义如下。

定义 6　图像中用于嵌入水印的非重叠的纹理粗糙区域被定义为水印的嵌入区域。

获取水印嵌入区域的方法如下:将所有纹理粗糙区域按照局部纹理值从大到小的顺序排序,逐个遍历纹理粗糙区域,计算当前访问纹理粗糙区域中心点与其他纹理粗糙区域中心点的坐标差,若横坐标差和纵坐标差分别都小于纹理粗糙区域的长和宽,则判定 2 个纹理粗糙区域发生重叠,删除其中局部纹理值较小的纹理粗糙区域,得到最终 n 个非重叠的纹理粗糙区域。

如果图像中没有一个通过滑动窗口遍历到的

区域能够被判定为纹理粗糙区域,则选择前 3 个具有较大局部纹理值且不重叠的窗口区域作为水印的嵌入区域。

3 自适应嵌入参数

在纹理粗糙区域中嵌入和提取水印时,可采用已有的水印嵌入和提取算法。本文采用了文献[6]中提出的差值量化水印算法(DQAQT)。

在水印嵌入算法中,水印的嵌入强度由一个嵌入参数控制,使用自适应调参模型可根据载体图像的相关信息自适应地调整水印的嵌入参数,能够减小不同图像中嵌入水印之间不可见性的差异。若要保证每幅图像中嵌入水印的不可见性,意味着需要使每幅嵌入水印的图像与其原始图像之间具有一致良好的水印不可见性指标,因为水印的不可见性是通过反映水印嵌入前后图像的变化情况的一些评价指标来衡量的。DQAQT 中采用了基于 PSNR 的自适应调参模型,能够根据期望的 PSNR 自适应地调整水印的嵌入参数,使得每幅嵌入水印的图像与原始图像之间的 PSNR 都能保持基本一致。那么,一致的 PSNR 能否保证每幅含水印图像都拥有一致的视觉质量呢?

3.1 不可见性的评价指标

自然图像的像素间存在着很强的相关性[23],这些相关性在视觉场景中携带着关于物体结构的重要信息。人类视觉系统(HVS)主要从可视区域内获取结构信息,所以可以通过探测结构信息是否改变来近似感知图像的失真情况。结构相似度(SSIM)就是衡量数字图像、视频主观感受质量的一种方法。PSNR 是基于像素点之间误差的图像质量评价,它并未考虑到人眼的视觉特性,而 SSIM 源于 HVS 感知图像的原理,评价结果与主观感受更接近。使用相同的水印算法在图 8 所示的纹理粗糙区域中嵌入与图 2 等量的水印,虽然能够得到基本相同的 PSNR,但 SSIM 差距很大,对比图 2 和图 8 可以证实,PSNR 与人的实际感受是不一致的,而 SSIM 更符合人眼视觉感知。

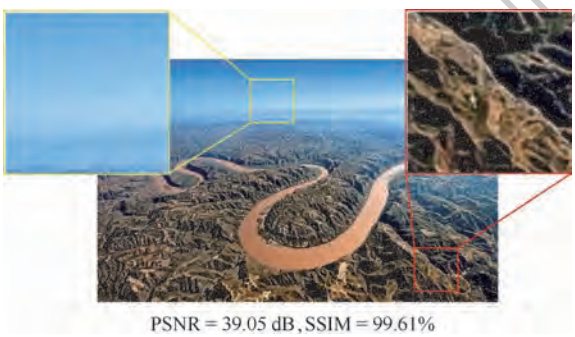


图 8 纹理粗糙区域嵌入水印

Fig. 8 Watermark embedded in textured region

DQAQT 在进行水印嵌入时需要利用到图像的每一个像素,对于图像整体而言,水印的嵌入强度是一致的,因此,它更适合利用 PSNR 作为不可见性的评价指标。本文算法是在图像中的部分区域嵌入水印,嵌入区域以外的像素不会因嵌入水印而发生变化,并且对于不同区域水印的嵌入强度可以不同。由于人眼对图像不同区域变化的承受力不同,SSIM 的评价结果更符合人眼感知,因此,SSIM 更适合用作为本文算法不可见性的评价指标。

综上所述,DQAQT 中采用的基于 PSNR 的自适应调参模型并不适合用于调整本文算法的嵌入参数。为使所有图像中嵌入的水印都获得一致良好的不可见性,希望能够设计一种新的自适应调参模型来调节本文算法的嵌入参数,在提高水印鲁棒性的同时使得嵌入水印的图像拥有基本一致的 SSIM。

3.2 自适应调参模型

从图 8 可以看出,纹理粗糙区域嵌入的水印具有更好的不可见性(SSIM);相对地,图像纹理越丰富,人眼能够忍受像素的变化量越大,即在相同不可见性条件下,纹理值越大的区域可设置更大的嵌入参数。所以希望能够利用嵌入区域的纹理值自适应地调整水印的嵌入参数,使得在所有嵌入区域获得一致 SSIM 的前提下,最大限度地增加每个嵌入区域的水印嵌入强度,从而提高水印算法的鲁棒性。

局部纹理值反映了一个区域的最弱纹理情况,因此在 2.2 节中通过利用区域的局部纹理值来判定一个区域是否为纹理粗糙区域。然而,由于局部纹理值仅为区域中某一分块的纹理值,无法反映区域整体的灰度分布及对比度,因此,需要同时考虑区域的局部纹理值和全局纹理值。

本文选取了 100 个纹理值不同的区域,调整对应的嵌入参数,使得它们在嵌入水印后的 SSIM 相同。为得到嵌入参数与纹理值的函数关系,将局部纹理值(T_l)、全局纹理值(T_g)、局部纹理值的二次方(T_l^2)及全局纹理值的二次方(T_g^2)作为自变量,嵌入参数(P_e)作为因变量,利用多元线性回归的方法拟合数据点。表 1 给出了各种自变

表 1　多元线性回归结果

Table 1　Multiple linear regression results

自变量	自变量系数					
	模型 1	模型 2	模型 3	模型 4	模型 5	模型 6
T_l	7.852***	11.40***	0	0	3.686***	4.654***
T_l^2	0	−0.204***	0	0	0	−0.053 0
T_g	0	0	4.371***	3.087***	2.732***	1.948***
T_g^2	0	0	0	0.044 3*	0	0.025 1
常数	28.48***	18.06***	18.67***	25.22***	17.87***	19.33***
R^2/%	89	91	94.7	95.2	100	100

注:*** 表示 $p<0.01$, ** 表示 $p<0.05$, * 表示 $p<0.1$,p 值表示模型中自变量系数的标准差,反映该系数在模型中表现是否显著,*** 表示十分显著。

量组合情况下得到的关系式对数据点的拟合情况。

以模型 6 为例,它所代表的关系式为:$P_e = -0.053T_l^2 + 0.025\,1T_g^2 + 4.654\,T_l + 1.948T_g + 19.33$。$R^2$ 代表有多少数据点能够被相应模型解释,如对于模型 1,有 89% 的数据点可由局部纹理值单独解释。

如表 1 所示,模型 1 ~ 模型 3、模型 5 的系数均十分显著,而模型 5 能够拟合更多的数据点。计算每幅图像通过这些模型估计的嵌入参数与实际设置的嵌入参数之间的误差值,如图 9 所示。可以看到,模型 5 的误差值的波动幅度最小,由此也反映模型 5 拟合度是最高的。

根据表 1 结果可看到,模型 5 和模型 6 都能够很好地拟合所有的数据点,但对于模型 6,局部纹理值和全局纹理值的二次项表现并不显著,属于过拟合的情况,而模型 5 中的 3 个估计值均十分显著,因此,嵌入参数与纹理值之间的关系可由模型 5 表示,关系式为

$$P_e = 3.686T_l + 2.732T_g + 17.87 \tag{5}$$

该模型是在给定嵌入水印后的 SSIM 条件下生成的,可使嵌入水印后的图像获得基本一致的 SSIM。通过该模型,可利用嵌入区域的纹理值自适应地调整对应的嵌入参数,最优化每幅图像中每个嵌入区域的嵌入参数,最大限度地保证了水印的不可见性和鲁棒性。

确定水印嵌入区域后,在每个嵌入区域中利用差值量化水印算法[6]进行水印的嵌入和提取,并通过模型 5 自适应地调整每个嵌入区域的水印嵌入参数,从而提出一种基于图像纹理的自适应水印算法,将其命名为 TBAQT。

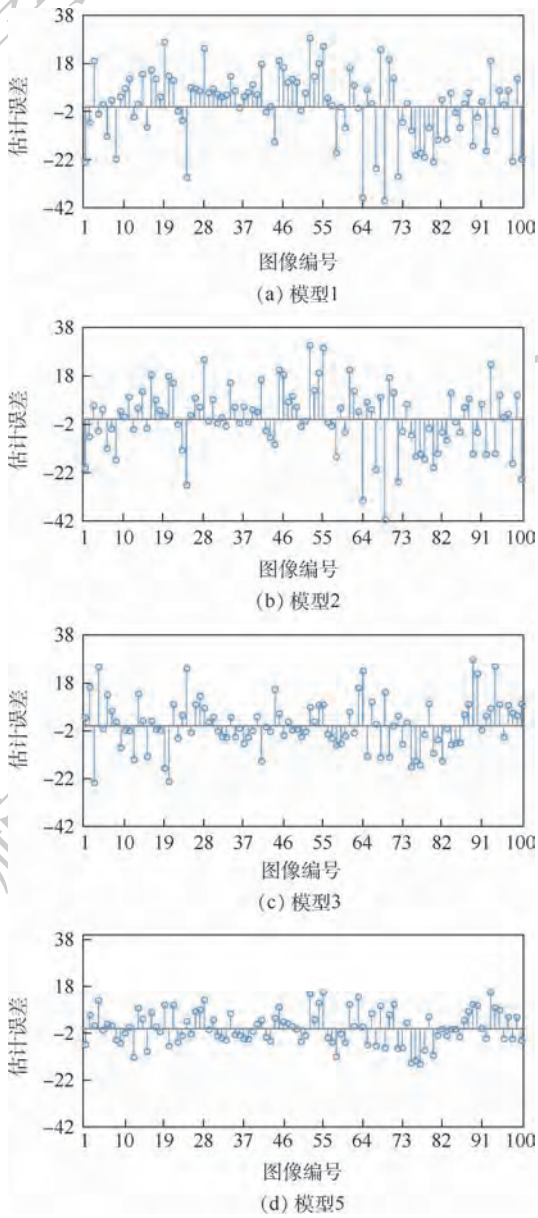


图 9　模型拟合误差

Fig. 9　Fitting error of models

4 实验结果与分析

4.1 实验设置

从华盛顿大学的图像数据库(http://www.cs.washington.edu/research/imagedatabase/)中随机选择 100 幅不同尺寸的图像组成实验的图像数据集,这些图像都是一般的自然场景图像,能够更好地体现算法的实用性。通过分析 100 幅图像所产生结果的平均值来对算法的各项指标进行评估。待嵌入这些图像中的水印为相同的 128 位 0/1 序列。

利用原始图像与含水印图像之间的 SSIM 来评估水印的不可见性,该评价指标更符合人眼的视觉感知;利用原始水印与提取水印之间的比特错误率(BER)来评估水印的鲁棒性。一个好的水印算法应具有较高的 SSIM 和较低的 BER。

在本文算法中,将滑动窗口的大小与原始图像的比例 W 设置为 1/4,滑动窗口的移动步长与原始图像的比例 S 设置为 1/16,此时滑动窗口尺寸仅为原始图像的 1/16,并且一幅图像包含 256 个窗口,可尽量避免窗口中包含一小块纹理平滑区域;计算窗口内区域局部纹理值时的划分块尺寸与滑动窗口尺寸的比例 P 设置为 1/16,纹理阈值 T_t 设置为 3.0;以保证纹理粗糙区域的纹理丰富程度。

在平衡不可见性和鲁棒性上,自适应水印算法 TDSS[2] 和 DQAQT[6] 表现突出,它们与本文算法(TBAQT)嵌入了相同的水印容量(128 bit),采用了相同的两级 DCT 变换,并选择了相同的 DCT 系数用于嵌入水印。除此之外,DQAQT 与 TBAQT 还采用相同的差值量化方法来嵌入和提取水印,而 TDSS 中采用的是扩频嵌入和提取方法。TDSS、DQAQT 与 TBAQT 的主要区别在于水印的嵌入区域及采用的自适应调参模型不同。TDSS 和 DQAQT 将一幅完整的图像作为水印的嵌入区域,TBAQT 仅在图像的纹理粗糙区域中嵌入水印。TDSS、DQAQT 与 TBAQT 分别采用了基于均值、PSNR 与图像纹理的自适应调参模型。因此,将 TDSS 和 DQAQT 与 TBAQT 进行整体性能的比较,能够公正地体现 TBAQT 的优势。

4.2 性能评估

4.2.1 不可见性比较

实验通过在相同水印嵌入强度的情况下对比原始图像与含水印图像之间的 SSIM 值来评估水印的不可见性。由于 TDSS 与 DQAQT、TBAQT 采用的嵌入方法不同,无法设置相同的嵌入强度,因此仅将 DQAQT 与 TBAQT 进行不可见性的比较。DQAQT 与 TBAQT 均使用了相应的自适应调参模型调整水印嵌入参数,由于嵌入参数控制着水印的嵌入强度,嵌入参数不同,水印的嵌入强度则不同。为了能够有效对比水印的不可见性,实验将 DQAQT 与 TBAQT 的嵌入参数设置为 100,以保证它们水印的嵌入强度相同。利用 TBAQT 与 DQAQT 分别对数据集中的图像进行水印嵌入,最终得到每幅图像的 SSIM 值如图 10 所示。

在相同水印嵌入强度的情况下,图像 SSIM 值之间的差异较大,这是因为每幅图像的纹理情况都是不同的,因此嵌入水印后图像的视觉效果必定有差异。能够很明显地看到,TBAQT 得到的每幅图像的 SSIM 值均远远优于 DQAQT,这足以证明 TBAQT 可使水印具有更好的不可见性。出现这样的结果原因很显然,DQAQT 是在整幅图像上进行水印嵌入,虽然水印的嵌入强度会均摊到每个像素上,但嵌入在纹理平滑区域的水印相比在纹理粗糙区域会让视觉更敏感,从而拉低了整幅图像的 SSIM 值;而 TBAQT 仅在纹理粗糙区域嵌入水印,嵌入区域在图像中占比很小,虽然单位像素嵌入强度较 DQAQT 大,但由于纹理粗糙区域对水印具有很好的隐蔽性,不会对图像质量产生明显的影响,因此使得整幅图像能够获得较高的 SSIM 值。这也证明了第 2 节所述,在相同水印嵌入强度的情况下,纹理粗糙区域中嵌入的水印会具有更好的不可见性。

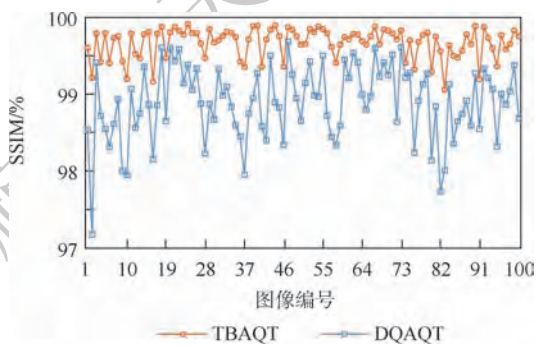


图 10 相同水印嵌入强度情况下 100 幅图像的 SSIM 值

Fig. 10 SSIM of 100 images with the same watermark embedding strength

4.2.2 自适应性比较

TDSS、DQAQT 与 TBAQT 均为自适应水印算法,TDSS 使用待嵌入水印的变换域系数的均值自适应地调整嵌入参数;DQAQT 通过给定 PSNR 自适应地调整嵌入参数;TBAQT 利用嵌入区域的纹理值自适应地调整嵌入参数。实验将 TDSS 的嵌入参数初始值设置为 62,DQAQT 的 PSNR 阈值设

置为 46 dB,使 TDSS、DQAQT 和 TBAQT 得到的 100 幅含水印图像与它们的原始图像之间的平均 SSIM 值基本相同(平均 SSIM 值分别为99.193%、99.195% 和 99.202%,SSIM 值大于 99% 已能够反映非常好的视觉质量),以对比采用这 3 种自适应水印算法得到的每幅图像的 SSIM 值的波动情况,以及在相同不可见性条件下对比 3 个算法的鲁棒性。

图 11 给出了 TDSS、DQAQT 和 TBAQT 得到的每幅图像的 SSIM 值。可以看到,由 TDSS 获得的图像的 SSIM 曲线波动剧烈,并且其中许多幅图像的 SSIM 值远远低于平均值,意味着嵌入的水印可能被人眼察觉。这也说明,待嵌入水印的变换域系数的均值并不能准确地反映图像的视觉特性。通过 DQAQT 获得的含水印图像与原始图像之间的 SSIM 值的波动幅度明显大于 TBAQT。这是因为虽然 DQAQT 能够让每幅图像获得基本一致的 PSNR,但 PSNR 并未考虑到人眼的视觉特性,因此导致了与 SSIM 值(更符合人眼主观感受)不一致的情况。而 TBAQT 的自适应调参模型是在给定的 SSIM 值的条件下进行的,自然能够获得符合期望的视觉效果。

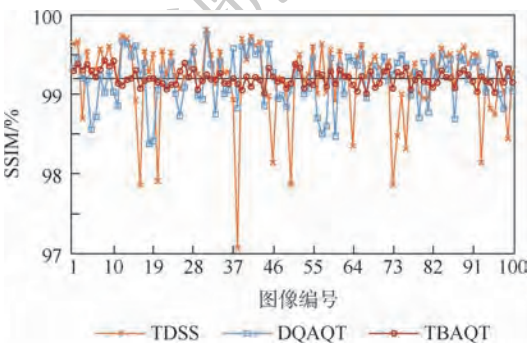


图 11　不同算法得到的 100 幅图像的 SSIM 值

Fig. 11　SSIM of 100 images obtained by different algorithms

4.2.3　鲁棒性比较

本节主要比较 TDSS、DQAQT 与 TBAQT 从经过各种攻击的含水印图像中恢复隐藏水印的能力,其中不但包括各种图像处理攻击,还包括多种几何攻击。实验分别利用 TDSS、DQAQT 与 TBAQT 首先给图像数据库中的 100 幅原始图像嵌入水印,然后对嵌入水印的图像进行各种攻击,最后从攻击后的含水印图像中提取水印。计算提取的水印与原始水印之间的 BER,用以衡量通过 TDSS、DQAQT 与 TBAQT 嵌入到图像中的水印的鲁棒性。

1) 图像处理攻击

为了评估 3 个算法的鲁棒性,对它们生成的含水印图像应用各种常见的图像处理攻击,主要攻击类型和攻击参数设置及模拟实验的结果如表 2所示。表中:黑体数据为最小值。

表 2　图像处理攻击后不同算法提取水印的 BER 值

Table 2　BERs of watermark extracted by different algorithms under image processing attacks

攻击类型	参数设置	BER/%		
		TDSS	DQAQT	TBAQT
JPEG 压缩	30%	4.59	0.62	**0.58**
	50%	2.89	**0.08**	**0.08**
	70%	1.60	**0**	**0**
高斯噪声	0.04	7.45	3.58	**0.06**
	0.03	5.88	2.22	**0.02**
	0.02	4.28	0.68	**0**
椒盐噪声	0.048	3.24	0.27	**0**
	0.04	2.91	0.15	**0**
	0.035	2.83	0.08	**0**
均值滤波	3×3	2.56	**0.01**	5.75
中值滤波	3×3	3.37	**0.03**	3.18
直方图均衡	—	0.90	0.14	**0**
亮度改变	0.6	1.36	**0**	**0**
	1.6	2.21	0.12	**0**
	2	3.03	0.71	**0.01**

从实验结果可以看出:

(1) 除均值滤波攻击外,在任何情况下 TDSS 提取水印的 BER 都远高于 DQAQT 和 TBAQT,这一方面是因为 TDSS 采用的扩频方法存在载体信号干扰(HSI),这是扩频方法固有的缺陷,会影响水印提取的准确率;另一方面是因为 TDSS 中采用的自适应调参模型仅根据均值调整水印的嵌入参数,对嵌入强度的调节能力有限,无法最优化水印的鲁棒性。

(2) DQAQT 与 TBAQT 抵抗 JPEG 压缩的能力不相上下,DQAQT 提取水印的 BER 略微高于 TBAQT。因为两者均将水印嵌入在 DCT 的中低频系数中,而 JPEG 主要是通过删除图像 DCT 的高频系数来达到压缩的目的,因此对水印的影响较小。

(3) 面对噪声叠加、直方图均衡及亮度改变等图像处理攻击,TBAQT 很好地展现了它的优势,实验结果非常优秀,其提取水印的 BER 明显低于 DQAQT。这归功于 TBAQT 利用图像纹理来调整水印的嵌入参数,嵌入区域的纹理越丰富,水印的嵌入强度越大,因此能够有效阻挡各种噪声的干扰及像素值的改变;并且 TBAQT 在多个纹理粗糙区域嵌入了相同的水印,进一步提高了算法的鲁棒性。

(4) DQAQT 在抵抗低通滤波(均值滤波和中值滤波)方面表现良好,相比之下 TBAQT 略有逊

色。这是由于滤波模板在处理 TBAQT 嵌入区域的边缘时,会引入嵌入区域以外的边缘附近的像素,导致嵌入区域边缘像素受到影响,从而容易使嵌入在这些位置的水印发生提取错误。

2)几何攻击

对含水印图像应用各种几何攻击来进一步对比 3 个算法的鲁棒性。主要攻击类型和参数设置及模拟实验的结果如表 3 所示。

表 3 几何攻击后不同算法提取水印的 BER 值

Table 3 BERs of watermark extracted by different algorithms under geometric attacks

攻击类型	参数设置	BER/%		
		TDSS	DQAQT	TBAQT
翻转	水平/垂直	1.56	**0**	**0**
旋转	90°整数倍	1.56	**0**	**0**
剪切	25%	2.12	0.04	**0**
	32%	5.55	1.38	**0**
	50%	9.63	4.84	**0**
遮挡	0.5	12.58	1.20	**0**
	0.6	16.10	3.60	**0**
	0.7	19.71	8.66	**0.01**
放缩	0.6	1.77	**0**	**0**
	0.8	1.52	**0**	**0**
	2.0	1.56	**0**	**0**
缩放	0.6	2.23	**0**	2.16
	0.8	1.60	**0**	**0**
	2.0	1.60	**0**	**0**
纵横比调整	0.8×1.4	1.62	**0**	**0**
	1.8×0.7	1.63	**0**	**0**

从实验结果可以看出:

(1)在所有情况下,DQAQT 与 TBAQT 嵌入水印的鲁棒性都优于 TDSS。原因在分析图像处理攻击实验结果的(1)中已进行解释。

(2)针对旋转、翻转、缩放、放缩及纵横比调整等几何攻击,TBAQT 和 DQAQT 的表现势均力敌,提取水印的 BER 基本都为 0。这一方面得益于采用的差值量化方法能够有效抵抗这些攻击;对于 TBAQT 而言,更重要的是其纹理粗糙区域大小及滑动窗口移动方式的适当设置,使得获取的纹理粗糙区域在攻击前后仍然保持一致。TBAQT 唯独对于缩放倍数低于 0.6 的图像,提取水印的能力不如 DQAQT,这是因为嵌入区域本身就比图像小得多,当图像缩放到很小时,纹理粗糙区域的尺寸相对更小,导致部分水印信息丢失而无法正确提取。

(3)剪切和遮挡攻击都是裁剪掉图像的部分像素,只是裁剪的位置有所不同。遮挡攻击裁剪中心区域相比剪切攻击裁剪左上角区域对图像影响更严重,因为一般重要信息都倾向于放置在图像的中心位置。TBAQT 能够完美抵抗这 2 种攻击,而 DQAQT 在裁剪范围超过 25% 后 BER 显著上升。这是由于 DQAQT 将水印嵌入在完整的图像中,部分图像信息丢失必然导致嵌入在其中的水印信息无法提取;而 TBAQT 在多个纹理粗糙区域中嵌入相同的水印,嵌入区域在图像中占比较小且没有重叠,即便部分区域被裁剪,仍然可能存在不受影响的嵌入区域,因此使得 TBAQT 能够有效抵抗剪切和遮挡攻击,这也是其的优势所在。

5 结 论

1)TBAQT 可使嵌入水印的所有图像获得一致良好的不可见性。在实验中,100 幅嵌入水印的图像与它们的原始图像之间的 SSIM 均高于 99%,且不同图像之间的 SSIM 差异不超过 0.5%。

2)TBAQT 对于绝大多数图像处理攻击和几何攻击均表现出优异的鲁棒性。在实验中,从被攻击后的嵌入水印图像中提取的水印与原始水印之间的 BER 基本接近于 0。

3)TBAQT 在嵌入水印前首先需要定位水印的嵌入区域,因此算法的运行时间相比直接在图像中嵌入水印的 TDSS 和 DQAQT 高,后续将对如何提高算法的运行效率展开研究。生成模型[16-20]是近期受关注的热点,在未来的研究工作中也将借鉴生成模型的相关方法,对水印算法进行深入研究。

参考文献(References)

[1] FAZLALI H R, SAMAVI S, KARIMI N, et al. Adaptive blind image watermarking using edge pixel concentration[J]. Multimedia Tools and Applications, 2017, 76(2): 3105-3120.

[2] GUAN H, ZENG Z, LIU J, et al. A novel robust digital image watermarking algorithm based on two-level DCT[C]// Proceedings of the International Conference on Information Science, Electronics and Electrical Engineering. Piscataway, NJ: IEEE Press, 2014: 1804-1809.

[3] MOHREKESH M, AZIZI S, SHIRANI S, et al. Hierarchical watermarking framework based on analysis of local complexity variations[J]. Multimedia Tools and Applications, 2018, 77(23): 30865-30890.

[4] YADAV N, SINGH K. Robust image-adaptive watermarking using an adjustable dynamic strength factor[J]. Signal, Image and Video Processing, 2015, 9(7): 1531-1542.

[5] HUANG Y, GUAN H, NIU B, et al. A spread-spectrum watermarking scheme with adaptive embedding strength and PSNR guarantee[C]// Proceedings of the International Conference on Anti-Counterfeiting, Security, and Identification, 2018: 82-87.

[6] HUANG Y, NIU B, GUAN H, et al. Enhancing image water-

marking with adaptive embedding parameter and PSNR guarantee[J]. IEEE Transactions on Multimedia, 2019, 21(10): 2447-2460.

[7] NASIR I, KHELIFI F, JIANG J, et al. Robust image watermarking via geometrically invariant feature points and image normalisation[J]. Image Processing IET, 2012, 6(4): 354-363.

[8] TANG C W, HANG H M. A feature-based robust digital image watermarking scheme[J]. IEEE Transactions on Signal Processing, 2003, 51(4): 950-959.

[9] YE X, CHEN X, MENG D, et al. A SIFT-based DWT-SVD blind watermark method against geometrical attacks[C]// Proceedings on the International Conference on Image & Signal Processing. Piscataway, NJ: IEEE Press, 2015: 14841650.

[10] SU Q, CHEN B. Robust color image watermarking technique in the spatial domain[J]. Soft Computing, 2018, 22(1): 91-106.

[11] ZONG T, XIANG Y, NATGUNANATHAN I, et al. Robust histogram shape-based method for image watermarking[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2015, 25(5): 717-729.

[12] LOAN N A, HURRAH N N, PARAH S A, et al. Secure and robust digital image watermarking using coefficient differencing and chaotic encryption[J]. IEEE Access, 2018, 6: 19876-19897.

[13] PARAH S A, SHEIKH J A, LOAN N A, et al. Robust and blind watermarking technique in DCT domain using inter-block coefficient differencing[J]. Digital Signal Processing, 2016, 53: 11-24.

[14] HUYNH-THE T, BANOS O, LEE S, et al. Improving digital image watermarking by means of optimal channel selection[J]. Expert Systems with Applications, 2016, 62(1): 177-189.

[15] HUYNH-THE T, HUA C, TU N A, et al. Selective bit embedding scheme for robust blind color image watermarking[J]. Information Sciences, 2018, 426: 1-18.

[16] PENG Y, QI J. Show and tell in the loop: Cross-modal circular correlation learning[J]. IEEE Transactions on Multimedia, 2019, 21(6): 1538-1550.

[17] PENG Y, ZHU W, ZHAO Y, et al. Cross-media analysis and reasoning: Advances and directions[J]. Frontiers of Information Technology & Electronic Engineering, 2017, 18(1): 44-57.

[18] QI J, PENG Y. Cross-modal bidirectional translation via reinforcement learning[C]// Proceedings of the International Joint Conference on Artificial Intelligence, 2018: 2630-2636.

[19] YUAN M, PENG Y. Text-to-image synthesis via symmetrical distillation networks[C]// Proceedings of the 26th ACM Multimedia Conference. New York: ACM, 2018: 1407-1415.

[20] 彭宇新,綦金玮,黄鑫. 多媒体内容理解的研究现状与展望[J]. 计算机研究与发展,2019,56(1):183-208.
PENG Y X, QI J W, HUANG X. Current research status and prospects on multimedia content understanding[J]. Journal of Computer Research and Development, 2019, 56(1): 183-208(in Chinese).

[21] 刘丽,匡纲要. 图像纹理特征提取方法综述[J]. 中国图象图形学报,2009,14(3):622-635.
LIU L, KUANG G Y. Overview of image texture feature extraction methods[J]. Journal of Image and Graphics, 2009, 14(3): 622-635(in Chinese).

[22] GONZALEZ R C, WOODS R E. Digital image processing, third edition[J]. Journal of Biomedical Optics, 2009, 14(2): 029901.

[23] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: From error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

作者简介:

黄樱　女,博士研究生。主要研究方向:图像水印。

牛保宁　男,博士,教授,博士生导师。主要研究方向:媒体大数据管理与分析、数据库系统性能管理、区块链数据管理。

关虎　男,博士,副研究员。主要研究方向:数字媒体内容版权保护与跟踪、数字媒体内容分析与处理、数字版权价值评估与版权交易。

张树武　男,博士,研究员,博士生导师。主要研究方向:数字版权管理、计算机视觉、跨媒体内容分析、语音和语言信息处理等。

Image texture based adaptive watermarking algorithm

HUANG Ying[1], NIU Baoning[1,*], GUAN Hu[2], ZHANG Shuwu[2]

(1. College of Information and Computer, Taiyuan University of Technology, Taiyuan 030024, China; 2. Digital Content Technology and Media Service Research Center, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: Watermarking is a technique of embedding a mark called a watermark in an image to prove the copyright of this image. This paper proposes an image texture based adaptive watermarking algorithm by taking advantage of the textured regions of an image being easy to hide the watermark. Firstly, a texture measurement method is proposed to reflect the richness of image's texture, and the concepts of global texture value and local texture value are introduced to comprehensively analyze the image texture distribution. The textured regions of an image are located by using the sliding window and judged by the local texture value of the inner window area, and then we only embed the watermark in the textured regions and ensure the visual quality of embedded watermark. The function relationship among the global texture value, the local texture value of the textured regions and the embedding parameter is obtained by multiple regression analysis. It can adaptively adjust the embedding parameters with the texture values of the regions to maximally enhance the imperceptibility and robustness of the watermark. Moreover, embedding the same watermark in multiple non-overlapping textured regions further improves the accuracy of the extracted watermark. The simulation experiments on 100 images show the superiority of the proposed method compared with the state-of-the-art methods in terms of imperceptibility, adaptivity and robustness.

Keywords: image watermarking; image texture; imperceptibility; robustness; adaptivity

Received: 2019-07-09; **Accepted**: 2019-08-20; **Published online**: 2019-08-26 15:08

URL: kns.cnki.net/kcms/detail/11.2625.V.20190826.1052.001.html

Foundation items: National Key R & D Program of China (2017YFB1401000); International Cooperation Project of the Major R & D Program of Shanxi (201603D421015)

* **Corresponding author**. E-mail: niubaoning@tyut.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0371

一种基于曼哈顿距离的帧间加权预测算法

郭红伟[1,2], 朱策[2,*], 李帅[2], 王永华[2]

(1. 红河学院 工学院, 蒙自 661100; 2. 电子科技大学 信息与通信工程学院, 成都 611731)

摘 要: 合并模式通过共享邻域块的运动矢量(MV)来节省编码运动信息比特数,有效提升了编码器率失真性能。然而,当前合并模式中的运动补偿预测(MCP)不够准确。为此,分析了合并模式中的预测残差分布特点,并提出了一种基于曼哈顿距离的加权预测算法作为合并模式的附加候选项。首先,采用邻域合并候选项的运动矢量进行运动补偿预测得到多个预测块;然后,根据候选块位置与像素点的曼哈顿距离对获得的多个预测块进行加权平均得到附加候选项;最后,通过率失真优化(RDO)从附加候选项和原有候选项中选择出最佳的合并模式。实验结果显示:在联合探索测试模型 JEM 7.0 平台上,所提算法在不同的编码器配置下均获得了率失真性能的提升,其中低延迟 P 帧下达到了平均 1.34% 的比特率节省。

关 键 词: 视频编码; 率失真优化(RDO); 合并模式; 运动补偿; 加权预测
中图分类号: TN919.8
文献标识码: A **文章编号**: 1001-5965(2019)12-2415-08

随着电子信息技术的高速发展和各种视频数据采集方式的使用,数字视频成为了多媒体信息的主要载体。然而,未经压缩的数字视频数据量非常巨大,如分辨率为 4 000 × 2 000 的 8 bit RGB 彩色视频,帧率为 30 Hz,则其每小时的数据量高达 2 531 GB。如此之大的数据量给视频的传输和存储带来巨大的挑战,因此自 20 世纪 90 年代以来,视频压缩技术持续成为国内外研究和应用的热点领域[1]。现代视频编码器采用一个包括预测、变换、量化和熵编码等多种压缩工具的混合视频编码框架,能有效去除视频数据中的空域冗余、时域冗余、视觉冗余等,从而实现较高的数据压缩。预测编码包括帧内预测和帧间预测,其中帧间预测利用了视频图像帧间相关性,根据运动矢量(Motion Vector,MV)和参考索引从重建帧中获得当前编码块的预测像素值。进一步地,相邻编码块的运动信息之间存在相关性,为了去除这种相关性从而减少用于编码运动信息的比特数,高效视频编码(High Efficiency Video Coding,HEVC)引入了高级运动矢量预测(Advanced Motion Vector Prediction,AMVP)、合并模式(merge mode)和跳过模式(skip mode)等技术[2]。视频编码器通过率失真优化(Rate Distortion Optimization,RDO)[3-5]为当前编码块选择最佳的预测模式,其数学表述为: $\min\{J\}$, $J = D + \lambda_{mode} R$。其中:J 称为率失真代价;D 和 R 分别为选择某一模式编码时得到的编码失真和编码码率(或比特数);λ_{mode} 为拉格朗日乘子。模式选择过程中,率失真代价最小的预测模式被选择用于编码当前块。

如果当前编码块选择了合并模式,则其采用相邻块的运动矢量从参考帧中获得预测像素,因

收稿日期: 2019-07-09; **录用日期**: 2019-08-13; **网络出版时间**: 2019-08-22 11:09
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190822.1059.003.html
基金项目: 国家自然科学基金(61571102,41761079); 四川省科学技术厅重点项目(2018JY0035); 云南省地方本科高校基础研究联合专项资金(2018FH001-056)
* **通信作者**. E-mail: eczhu@uestc.edu.cn

引用格式: 郭红伟, 朱策, 李帅, 等. 一种基于曼哈顿距离的帧间加权预测算法[J]. 北京航空航天大学学报, 2019, 45(12): 2415-2422. GUO H W, ZHU C, LI S, et al. Manhattan distance based inter-frame weighted prediction algorithm[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2415-2422 (in Chinese).

此不需要重复编码运动信息。合并模式节省了编码运动信息的比特数,在 HEVC 中实现了较高的率失真性能提升[6],该技术已被扩展应用到了下一代视频编码标准的开发中,新标准在联合视频专家组(Joint Video Experts Team,JVET)第 10 次会议后被正式命名为多功能视频编码(Versatile Video Coding,VVC)。近年来,研究人员针对视频编码中的合并模式提出了一些算法改进。HEVC 采用基于四叉树的灵活块划分结构,每个编码树单元(Coding Tree Unit,CTU)被迭代地划分为许多编码单元(Coding Unit,CU),每个 CU 又进一步划分为 1 ~ 2 个预测单元(Prediction Unit,PU)。因此,针对大量预测块的模式选择需要极高的运算量。为了降低 HEVC 编码器的运算复杂度,文献[7-9]提出了不同的早期合并模式决策方法。在文献[7]中,根据不同划分深度子 CU 的模式与最大编码单元(Largest Coding Unit,LCU)最佳模式之间的相关性,设计快速算法。文献[8]提出基于率失真代价的统计模型用于早期合并模式决策,其减少了近一半的编码时间,但率失真性能损失较大。文献[9]分析了不同 CU 深度和量化系数值对应的率失真代价变化特点,采用自适应阈值减少合并模式候选进行 RDO 的次数,以降低编码器运算复杂度。为了便于合并模式在硬件编码器中的实现,文献[10]提出能减少运算量和存储要求的合并模式设计方法。针对三维高效视频编码(3D-HEVC),文献[11-13]提出了相应的早期合并模式判决方法。大部分关于合并模式的改进算法[14-16]都是以损失率失真性能为代价,降低编码器运算复杂度。以提升编码器率失真性能为目的的合并模式改进方法[17]相对较少。

通过分析合并模式的运动补偿预测残差分布特性,本文提出了一个附加的合并模式候选项,其预测像素值由空、时域候选的预测块加权得到。附加的合并模式候选项能获得更精确的预测像素值,减小了预测残差,从而节省编码残差变换系数的比特数。实验结果显示,本文算法能有效地改善编码器压缩性能。

1 视频编码中的合并模式

虽然灵活的四叉树块划分结构更好地适应了视频内容,能有效去除时域冗余,然而细致的划分也导致需要编码的运动信息增加。为此,HEVC 中提出了一种新的模式,即合并模式[6]。合并模式利用了空域和时域相邻编码块的运动相关性,不需要进行运动估计(Motion Estimation,ME),当前待编码块直接使用已编码块的运动信息进行运动补偿预测(Motion Compensation Prediction,MCP)。编码过程中,选择合并模式的 PU 仅需编码邻域 PU 复用索引,有效地减少了比特消耗。图 1 为合并模式的流程。首先,根据空域和时域相邻已编码块的可用运动信息建立合并候选列表;然后,采用每个候选项的运动矢量对当前块进行运动补偿预测并编码计算率失真代价,其率失真代价最小的运动信息对应候选项即为最佳合并模式。

合并模式候选列表来自于空域、时域已编码块的可用运动信息和附加生成的列表项,空域和时域中采用的相邻块位置如图 2 所示。空域运动矢量候选有 5 个,时域运动矢量候选有 2 个,在 HEVC 中最多生成 4 个空域和 1 个时域合并候选列表。

具体地,建立合并候选列表的步骤如图 3 所示。其输出结果是当前块的合并候选列表,每个候选项具有不同的运动信息,被用于当前编码块的运动补偿预测。合并模式中,候选列表的个数由参数 MaxNumMergeCands 控制,其包含在码流中传输到解码端,在 HEVC 中设置合并列表数为 5,在VVC的开发过程中,该候选列表数增加到

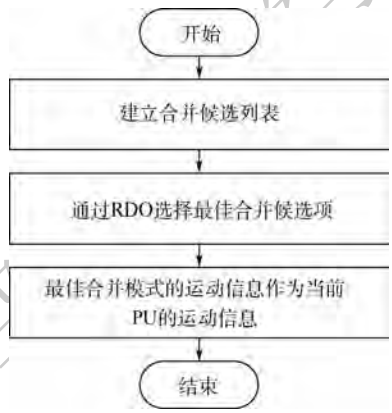


图 1 合并模式流程图

Fig. 1 Flowchart of merge mode

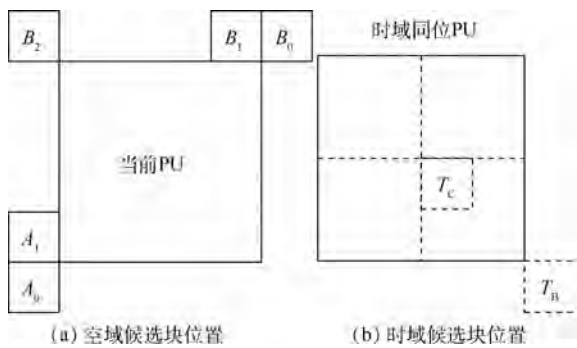


图 2 合并模式候选块位置

Fig. 2 Location of merge mode candidates

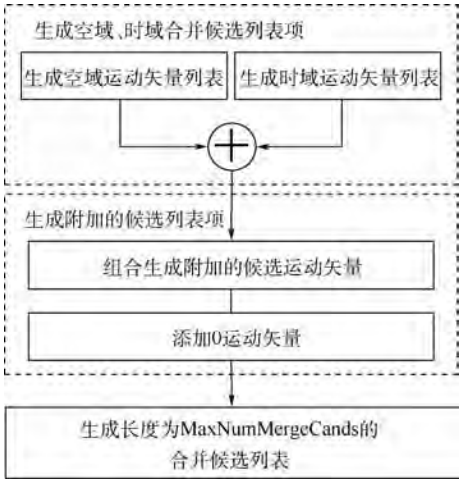


图 3　建立合并候选列表

Fig. 3　Construction of merge candidate list

7 个。参照图 2 和图 3,HEVC 中合并候选列表的生成过程描述如下:

1) 生成空域运动矢量列表

空域运动矢量列表由图 2(a)中相应位置可用的 PU 生成,按照 A_1、B_1、B_0、A_0、B_2的顺序检测每个位置运动信息的可用性,若某个位置的 PU 是帧内预测方式或其位于别的 Slice 中,则其不可用。另外还有 2 种特殊的情况,运动信息不可用。一种情况是检测到当前相邻块的运动信息与之前已检测相邻块的运动信息相同,则标记为不可用,也就是合并列表不添加重复的运动信息。另一种情况是若 CU 划分为左右 2 个 PU 或上下 2 个 PU,则右边 PU 或下方 PU 建立空域运动矢量列表时,对应于图 2(a)中的 A_1或 B_1不可用,这是因为左右2 个PU 运动信息相同或上下 2 个 PU 运动信息相同等价于 CU 采用了 $2N\times 2N$ 的 PU 划分模式。最终按检测顺序至多选择 4 个可用的运动信息生成空域运动矢量列表。

2) 生成时域运动矢量列表

时域运动矢量列表由图 2(b)中邻近已编码帧的同位 PU 生成,首先检测 T_B位置的运动信息是否可用,若 T_B位置 PU 采用帧内预测或与其左上角 PU 不在同一个 CTU 内,则其运动信息不可用,然后检测 T_C位置 PU 的运动信息可用性。合并候选列表中仅生成一个时域运动矢量列表项,该时域运动矢量需要根据当前帧和同位 PU 所属帧各自与其参考帧之间的距离进行缩放。

3) 生成附加的候选列表

如果生成的空域运动矢量列表和时域运动矢量列表总数小于 MaxNumMergeCands,则需要生成附加的候选列表,以使合并候选列表补齐到 MaxNumMergeCands。对于 B Slice,利用之前空域和时域生成的运动信息组合生成附加的候选运动矢量对。若之前步骤生成的合并候选列表长度仍小于 MaxNumMergeCands,则添加 0 运动矢量使列表长度达到 MaxNumMergeCands。

在进行 RDO 选择最佳合并候选项的过程中,每个候选列表项对应的当前块像素值由相应运动矢量通过运动补偿预测得到。式(1)表示单向预测中的运动补偿预测:

$$P(x,y) = F_{\mathrm{ref}}(x+i,y+j) \tag{1}$$

式中:$P(x,y)$和 $F_{\mathrm{ref}}(x,y)$分别为预测图像和参考图像;(i,j)为合并候选项中的运动矢量。双向预测则需要由 2 个运动矢量和对应的参考索引采用加权 2 个方向的预测值获得。

2　加权预测算法

合并模式不需要进行耗时的运动估计,通过复用领域 PU 的运动矢量极大地节省了编码运动信息的比特数,从而为视频编码带来显著的压缩性能提升。然而合并模式仍然有进一步优化以提升编码性能的空间,因为合并模式中的运动补偿预测不够准确。直观地,更准确的运动补偿预测能减小 PU 的预测残差,从而降低编码中的量化失真,减少编码量化系数的比特数。本节首先分析了合并模式中预测残差的分布特征,然后提出了一种基于曼哈顿距离的加权预测作为合并模式的附加候选项,以进一步改善合并模式的性能。

2.1　合并模式中预测残差的分布

合并模式决策过程中,每一个候选列表项的运动信息被用于运动补偿预测获得当前 PU 的残差块,如下:

$$B(x,y) = F(x,y) - P(x,y) \tag{2}$$

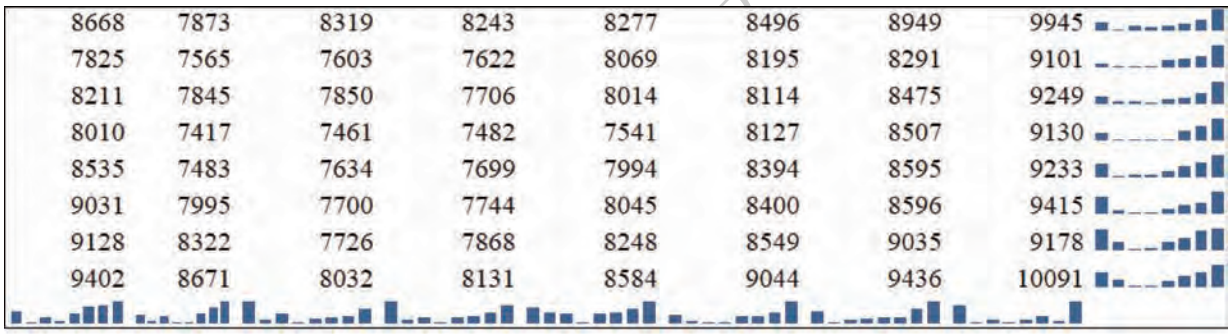
式中:$B(x,y)$、$F(x,y)$和 $P(x,y)$分别为当前 PU 的残差块、原始像素块和预测像素块。

直观地,2 个像素点的运动相关性随着它们之间距离的增大逐渐变小,因此当前 PU 中的像素点运动矢量与合并候选项的运动矢量相关性随着像素点到邻域候选块的距离增大而减小。如式(1)所示的运动补偿预测中,越接近候选块的像素点将能获得越准确的预测像素值,反之,离候选块较远的像素点的预测误差会较大。

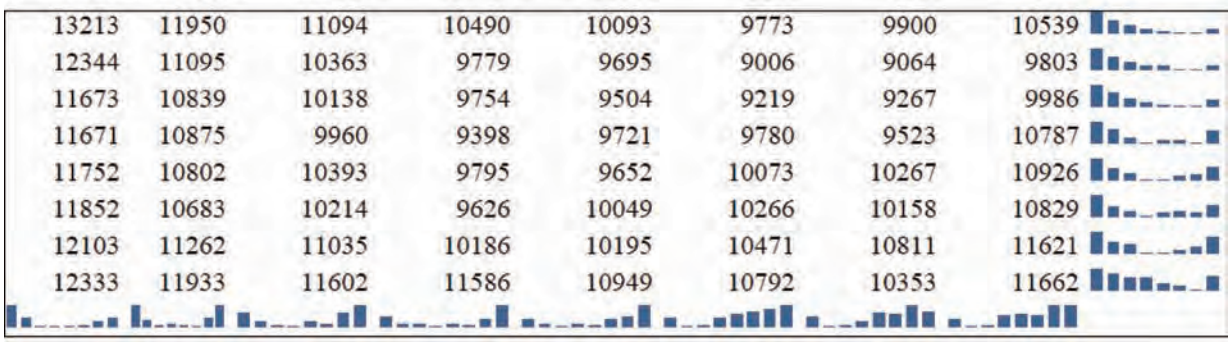
以图 2(a)中空域候选块 B_2为例,由于 B_2位于当前 PU 的左上方,当采用 B_2的运动矢量为当前 PU 进行运动补偿预测,则其残差块的残差幅度将呈现从上到下、从左到右逐渐增大的趋势。

类似的不均匀分布也将出现在其他合并候选项的运动补偿预测残差块中。

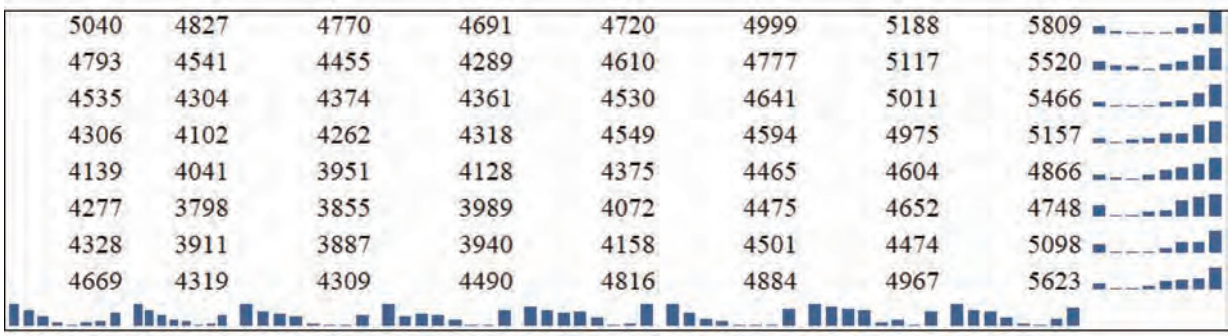
为了验证如上分析,采用 VVC 开发过程中的联合探索测试模型(Joint Exploration Test Model 7.0, JEM 7.0)[18]编码视频序列 BasketballDrill,编码器配置为低延迟 P,量化参数(Quantization Parameter, QP)设置为 22,编码帧数设置为前100 帧。统计编码过程中使用合并模式,运动矢量来自同一空域候选位置的所有 8×8 残差块的残差绝对值之和,该统计数据能反映出残差幅度的分布特点。图 4(a)~(c)分别为运动矢量来自图 2(a)中候选块 B_2、B_0、A_0时的残差幅度统计分布。尽管由于不同位置的候选块被选择作为最佳合并模式的次数不同,导致图 4 中的统计数值差异较大,但它们的残差幅度分布规律与前面的分析基本一致。即随着像素点位置与候选块的距离增加,相应的预测残差幅度变大。



(a) 运动矢量来自当前PU左上方位置候选块B_2



(b) 运动矢量来自当前PU右上方位置候选块B_0



(c) 运动矢量来自当前PU左下方位置候选块A_0

图 4 不同位置合并候选项进行运动补偿预测后的残差幅度统计分布

Fig. 4 Statistical distribution for residual magnitudes after MCP of merge candidates at different locations

2.2 基于曼哈顿距离的加权预测

从 2.1 节的分析可知,由于当前 PU 中不同像素点运动与候选块运动相关程度不同,导致合并模式中的运动补偿预测不够精确。为了充分利用像素间随着距离变化的运动相关性,本节提出一种基于曼哈顿距离的加权预测作为合并模式的

附加候选项,其预测块由可用的不同位置候选项预测块加权得到。所提出附加合并候选项的流程如图 5 所示,具体步骤如下:

1)检测邻域合并候选块

如图 2 中合并模式候选块位置所示,按 A_1、B_1、B_0、A_0、B_2、T_B的顺序检测不同位置候选块是否可用,并生成相应的运动矢量。如果生成的运动矢量数小于 2,则不执行本文算法,否则执行本文算法。需要指出的是,时域候选块 T_C并不包括在所提加权算法中,因为 T_C与当前 PU 中所有像素点的距离相同。

2)运动补偿预测

采用步骤 1)生成的运动矢量根据式(1)进行运动补偿预测获得相应的预测块,记为 $P_i(x,y)$。

3)加权预测

根据 2.1 节的分析可知,由步骤 2)得到的多个预测块具有不同的残差幅度分布特点,随着像素点位置与候选块的距离越远,残差幅度相对变得越大。因此,本文提出使用曼哈顿距离度量像素间随着距离变化的运动相关性。

曼哈顿距离表示 2 个点在标准坐标系上的绝对轴距之和,二维平面上两点 $a(x_1,y_1)$与 $b(x_2,y_2)$之间的曼哈顿距离定义为

$$d = |x_1 - x_2| + |y_1 - y_2| \tag{3}$$

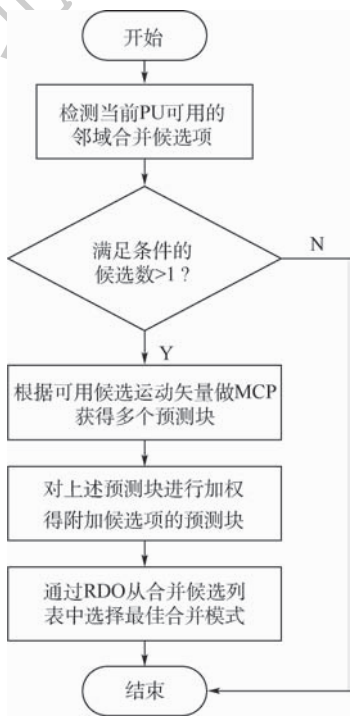


图 5　附加合并候选项的流程图

Fig. 5　Flowchart of additional merge candidate

所提附加合并候选项的预测块由式(4)加权得到:

$$P_w(x,y) = \frac{\sum_{i=1}^{n} W_i(x,y)P_i(x,y)}{\sum_{i=1}^{n} W_i(x,y)}$$
$$n = 2,3,\cdots,6 \tag{4}$$

式中:$W_i(x,y)$为每个预测块 $P_i(x,y)$对应的权重,其设置与候选块的位置有关。每个像素点的权重值随着它与候选块的曼哈顿距离增大而减小,具体设置如表 1 所示,其中 H 和 W 分别为当前预测块的高和宽。

4)最佳合并模式选择

计算所提附加候选项的率失真代价,并与原始合并候选列表中的率失真代价对比选择出最佳合并模式。在联合探索测试模型 JEM 7.0 中,合并候选项的索引号范围为 0 ~ 6,因此所提附加候选项索引号设置为 7。若所提附加候选项最终被选择作为当前 PU 的编码模式,则相应的合并索引号 7 被编码并传至解码端,解码端将采用与编码端一致的加权方法重建编码块。

需要说明是的,若当前 PU 最终选择所提附加合并候选项编码,则当前 PU 没有明确的运动信息。为便于后续编码过程中当前 PU 被用于导出其他编码块的运动矢量,需要对当前 PU 设置一个确定的运动矢量,本文算法采用步骤 1)中检测到的第一个可用候选块运动信息作为当前 PU 的运动信息。

表 1　不同位置候选块对应的权重设置

Table 1　Weight setting for candidates at different locations

候选块的位置	$W_i(x,y)$
A_1	$W-x+y$
B_1	$H-y+x$
B_0	$H-y+x$
A_0	$W-x+y$
B_2	$W-x+H-y+1$
T_B	$x+y+1$

3　实验验证

为验证本文算法的有效性,将提出的基于曼哈顿距离加权预测算法集成到联合探索测试模型 JEM 7.0,该测试软件是由 JVET 在 HEVC 测试模型(HEVC Test Model,HM)基础上开发的,用于

评估新的编码工具和技术改进。测试序列使用 JVET 通用测试条件(Common Test Conditions, CTC)和软件参考配置[19]建议的 Class B、Class C、Class D、Class E 中的 16 个视频和 Class F 的 4 个可选视频。实验测试了低延迟 P(Low Delay P, LDP)、低延迟 B(Low Delay B, LDB)、随机接入(Random Access, RA)3 种编码器配置,依据 CTC 设置 QP 分别为 37、32、27、22 编码每一个测试序列。需要说明的是,Class E 是会议视频序列,根据 CTC 建议,RA 配置下测试序列不包括 Class E 中的 3 个视频。以原始的 JEM 7.0 编码器作为基准,计算本文算法输出 Y、U、V 分量的 BDBR,BDBR 表示在相同的峰值信噪比(Peak Signal-to-Noise Ratio, PSNR)条件下,测试算法相对于基准算法的码率节省百分比,若其值为负则表示性能增益,反之表示性能损失。

表 2 列出了 3 种编码器配置下的全部测试序列 BDBR,数据显示本文算法在 3 种编码器配置下都获得了率失真性能提升。其中,LDP 的性能平均提升达 1.34%,比 LDB 和 RA 配置的性能提升大许多,这是因为 B 帧采用的双向预测比 P 帧的单向预测准确,对于 B 帧可提升的预测准确度空间比较有限。另外,Class F 是屏幕内容视频,其编码块中的像素运动相关性很强,基本不随距离大小变化。因此,对于 Class F 本文基于距离的加权预测获得性能改善不显著。对于实验数据中显示在 LDB 和 RA 配置下,个别测试序列出现率失真性能损失,其原因是由于本文算法为传输附加合并候选索引,增加了合并模式的编码比特,而对于这些特殊的视频,基于距离的加权预测提高的预测精确度比较小。

LDP 编码配置下部分测试序列的率失真性能提升很大,如测试序列 BQTerrace 的码率节省高达 6.98%。图 6 给出了性能改善较为显著的测试序列 BQTerrace、BasketballDrill、BQSquare、Johnny 的率失真曲线对比。可看出,本文算法在低码率和高码率下均优于原始的 JEM 7.0。

表 2 三种编码器配置下本文算法相比于 JEM 7.0 的 BDBR

Table 2 BDBR for proposed algorithm compared with JEM 7.0 under three configurations of encoder %

Class	测试序列	LDP			LDB			RA		
		Y	*U*	*V*	*Y*	*U*	*V*	*Y*	*U*	*V*
B	Kimono	-0.32	-0.29	-0.56	-0.07	-0.10	0.02	-0.07	0.11	0.04
	ParkScene	-0.36	-1.35	-0.83	0.08	0.23	0.30	-0.03	-0.39	-0.42
	Cactus	-1.66	-3.01	-2.57	-0.17	-0.37	-0.90	-0.26	-0.64	-0.46
	BasketballDrive	-1.18	-2.20	-2.27	-0.27	-0.77	-0.52	-0.15	-0.38	-0.02
	BQTerrace	-6.98	-5.93	-5.02	-0.40	-0.76	-0.57	-0.61	-0.71	-0.94
C	BasketballDrill	-1.58	-2.12	-2.69	-0.05	0.02	-0.17	-0.21	-0.02	-0.13
	BQMall	-0.56	-0.86	-0.66	-0.01	-0.79	-0.24	-0.08	-0.40	0.15
	PartyScene	-1.59	-1.66	-1.17	-0.33	-0.43	-0.19	-0.24	0	-0.16
	RaceHorses	-0.89	-1.87	-2.04	-0.08	-0.33	-0.54	-0.13	-0.28	-0.24
D	BasketballPass	-0.29	-0.53	-0.86	-0.01	-1.39	-0.60	-0.08	-0.26	-0.02
	BQSquare	-3.37	-5.00	-3.32	-0.36	-0.20	-1.26	-0.27	-0.03	0
	BlowingBubbles	-0.84	-0.94	-1.24	-0.23	-0.58	-0.67	-0.04	-0.01	-0.22
	RaceHorses	-0.12	-0.12	0.08	-0.01	-0.36	-0.45	0.02	-0.21	-0.23
E	FourPeople	-1.27	-1.17	-0.67	-0.29	-1.60	-2.66	—	—	—
	Johnny	-2.59	-3.99	-3.59	-0.03	-0.59	-0.52	—	—	—
	KristenAndSara	-1.10	-3.75	-3.70	0.09	-0.67	0.09	—	—	—
F	BasketballDrillText	-1.31	-0.98	-1.19	-0.15	-0.40	-0.29	-0.23	-0.04	-0.35
	ChinaSpeed	-0.29	-1.37	-1.02	-0.08	-0.27	-0.24	-0.09	-0.16	-0.25
	SlideEditing	-0.28	-0.15	-0.18	-0.02	-0.31	-0.61	0.04	0.03	0.01
	SlideShow	-0.16	-0.89	0.46	0.49	1.84	0.63	-0.30	-0.41	-0.19
	平均	-1.34	-1.91	-1.65	-0.10	-0.39	-0.47	-0.16	-0.22	-0.20

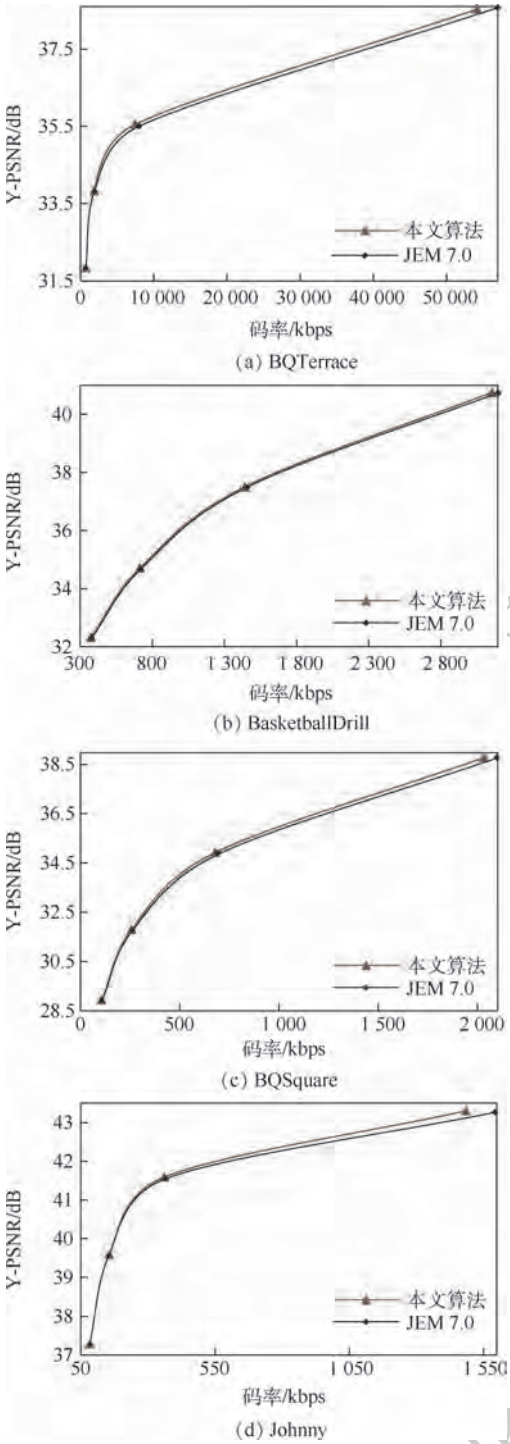


图 6　率失真曲线对比

Fig. 6　Comparison for rate-distortion curves

4　结　论

1）本文分析了合并模式空域、时域不同候选块运动信息与当前待编码 PU 中不同位置像素点运动信息的相关性,并通过实验统计证实随着 PU 中像素点距候选块的距离增加,运动补偿预测后的残差幅度增大。

2）提出了一种基于曼哈顿距离的加权预测算法生成附加合并候选项,其充分利用了所有可用候选块的运动信息,根据每个可用候选块与当前待编码 PU 中像素点的曼哈顿距离设置加权预测的权重,该加权预测获得了更准确的预测块,从而减小量化失真和编码量化系数的比特数。

3）实验结果显示,在 LDP、LDB 和 RA 编码结构下,本文算法均获得了率失真性能提升。特别地,对于 LDP 配置,相比原始 JEM 7.0 编码器,本文算法获得了平均 1.34% 的码率节省,最高达到 6.98%。

参考文献（References）

[1] 马思伟. AVS 视频编码标准技术回顾及最新进展[J]. 计算机研究与发展,2015,52(1):27-37.
MA S W. History and recent developments of AVS video coding standards[J]. Journal of Computer Research and Development, 2015,52(1):27-37(in Chinese).

[2] SULLIVAN G J, OHM J R, HAN W J, et al. Overview of the high efficiency video coding (HEVC) standard[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012,22(12):1649-1668.

[3] GAO Y, ZHU C, LI S, et al. Temporal dependent rate-distortion optimization for low-delay hierarchical video coding[J]. IEEE Transactions on Image Processing, 2017,26(9):4457-4470.

[4] LI S, ZHU C, GAO Y B, et al. Lagrangian multiplier adaptation for rate-distortion optimization with inter-frame dependency[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016,26(1):117-129.

[5] LI S, ZHU C, GAO Y, et al. Inter-frame dependent rate-distortion optimization using Lagrangian multiplier adaption[C]// Proceedings of IEEE International Conference on Multimedia and Expo(ICME). Piscataway, NJ: IEEE Press, 2015:1-6.

[6] HELLE P, OUDIN S, BROSS B, et al. Block merging for quadtree-based partitioning in HEVC[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012,22(12): 1720-1731.

[7] PAN Z Q, KWONG S, SUN M T, et al. Early merge mode decision based on motion estimation and hierarchical depth correlation for HEVC[J]. IEEE Transactions on Broadcasting, 2014, 60(2):405-412.

[8] TARIQ J, KWONG S, YUAN H. Spatial/temporal motion consistency based merge mode early decision for HEVC[J]. Journal of Visual Communication and Image Representation, 2017, 44:198-213.

[9] 蒋洁,叶德周,潘勉. 一种基于自适应阈值的 Merge 模式快速选择方法[J]. 光电子·激光,2016,27(9):980-986.
JIANG J, YE D Z, PAN M. A fast candidate selection method for Merge mode based on adaptive threshold[J]. Journal of Optoelectronics·Laser, 2016,27(9):980-986(in Chinese).

[10] KIM T S, RHEE C E, LEE H J. Merge mode estimation for a hardware-based HEVC encoder[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016,26(1):195-209.

[11] SONG Y X, JIA K B. Early merge mode decision for texture coding in 3D-HEVC[J]. Journal of Visual Communication and Image Representation, 2015, 33: 60-68.

[12] HEO Y S, BANG G, PARK G H. Adaptive merge list construction for 3D-HEVC fast encoder[J]. Electronics Letters, 2016, 52(8): 604-690.

[13] 宋雨新,贾克斌,吴强. 3D-HEVC 合并模式快速判决方法研究[J]. 信号处理, 2016, 32(1): 46-55.
SONG Y X, JIA K B, WU Q. Research on fast merge mode decision method for 3D-HEVC[J]. Journal of Signal Processing, 2016, 32(1): 46-55(in Chinese).

[14] WU J F, GUO B L, HOU J, et al. Fast CU encoding schemes based on merge mode and motion estimation for HEVC inter prediction[J]. KSII Transactions on Internet and Information Systems, 2016, 10(3): 1195-1211.

[15] CHENG Z X, SUN H M, ZHOU D J, et al. Accelerating HEVC inter prediction with improved merge mode handling[J]. IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, 2017, 100(2): 546-554.

[16] VANNE J, VIITANEN M, HAMALAINEN T D. Efficient mode decision schemes for HEVC inter prediction[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(9): 1579-1593.

[17] ZHANG N, FAN X P, ZHAO D B, et al. Merge mode for deformable block motion information derivation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2017, 27(11): 2437-2449.

[18] CHEN J, ALSHINA E, SULLIVAN G J, et al. Algorithm description of joint exploration test model 7(JEM 7)[C]// Joint Video Exploration Team(JVET) of ITU-T SG16WP3 and ISO/IEC JTC 1/SC 29/WG 11 7th Meeting, 2017: 10-15.

[19] SUEHRING K, LI X. JVET common test conditions and software reference configurations[C]// Joint Video Exploration Team (JVET) of ITU-T SG16WP3 and ISO/IEC JTC 1/SC 29/WG11 2nd Meeting, 2016: 1-4.

作者简介:

郭红伟 男,博士研究生,副教授。主要研究方向:视频编码与通信。

朱策 男,博士,教授,博士生导师。主要研究方向:视频编码与通信、视频分析和处理。

李帅 男,博士,副研究员。主要研究方向:视频编码、计算机视觉。

王永华 男,硕士研究生。主要研究方向:视频编码。

Manhattan distance based inter-frame weighted prediction algorithm

GUO Hongwei[1,2], ZHU Ce[2,*], LI Shuai[2], WANG Yonghua[2]

(1. School of Engineering, Honghe University, Mengzi 661100, China;
2. School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China)

Abstract: Merge mode saves the number of bits required to encode motion information by sharing the motion vector (MV) in neighboring blocks, which improves the rate-distortion performance of encoders effectively. However, motion compensation prediction (MCP) is not accurate enough in the merge mode currently. Therefore, this paper analyses the characteristics of residual distribution after MCP in the merge mode, and presents a Manhattan distance based weighted prediction method as an additional candidate for the merge mode. First, several predicted blocks are obtained by MCP with motion vectors in neighboring candidates. Second, the additional candidate is obtained by a weighted average method according to Manhattan distances from the neighboring candidate to the pixel points in the predicted blocks obtained. Finally, the best merge mode is selected by rate distortion optimization (RDO) among the additional candidate and the original candidates. The experimental results show that, on the joint exploration test model 7.0 (JEM 7.0), the proposed method achieves rate distortion performance improvement under the different configurations of encoder, where a bitrate saving of 1.34% on average is obtained under the configuration of low delay P frame.

Keywords: video coding; rate distortion optimization (RDO); merge mode; motion compensation; weighted prediction

Received: 2019-07-09; **Accepted**: 2019-08-13; **Published online**: 2019-08-22 11:09

URL: kns.cnki.net/kcms/detail/11.2625.V.20190822.1059.003.html

Foundation items: National Natural Science Foundation of China (61571102, 41761079); Key Project of Sichuan Provincial Department of Science and Technology (2018JY0035); Yunnan Local Colleges Applied Basic Research Project (2018FH001-056)

* **Corresponding author**. E-mail: eczhu@uestc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0385

基于社团结构节点重要性的网络可视化压缩布局

吴玲达[1], 张喜涛[1,2,*], 孟祥利[1]
(1. 航天工程大学 航天信息学院, 北京 101416; 2. 航天工程大学 航天指挥系, 北京 101416)

摘 要: 为有效展示网络的中观尺度结构,将力导引布局算法与网络社团结构特征相结合,提出了一种基于社团结构节点重要性的网络可视化压缩布局方法。首先,采用 Louvain 算法对网络进行多粒度社团结构划分;然后,通过计算社团结构中节点的拓扑势评估节点的重要性,保留社团结构中的重要节点,合并边缘节点,实现社团结构压缩;最后,采用力导引布局算法布局压缩网络节点,实现网络可视化的压缩布局。实验结果表明:所提方法在压缩节点和连边规模的基础上,能够完整保留原始网络的社团构成,并且通过保留社团结构代表点可以清晰展示社团内部结构,突出社团和重要节点在网络结构中的位置和作用。

关 键 词: 网络可视化; 压缩布局; 力导引布局; 拓扑势; 节点重要性
中图分类号: TP391.9
文献标识码: A **文章编号**: 1001-5965(2019)12-2423-08

网络可视化的目的是辅助用户感知网络结构,理解和探索隐藏在网络数据内部的规律、模式等[1-2]。然而,如果将所有节点和连边的细节信息完全展示在有限尺寸的屏幕中,会导致以下问题:①受屏幕尺寸限制,对于具有大量节点和高密度连边的大规模网络可视化来说,用户会陷入混乱重叠的节点-连接图中,难以识别和感知网络的整体结构,更不能引导用户发现感兴趣的元素;②大规模的数据量必然带来更大的计算压力,对算法和硬件都会有更高的要求。低效、耗时的网络可视化就失去了应有的意义。

因此,为了达到有效展示信息和辅助用户感知网络整体结构的目的,必须对大规模网络进行一定的缩减处理,以降低用户的感知复杂度和布局过程的计算复杂度。根据缩减目的,可将网络缩减处理方法分为 3 类:过滤、抽样和压缩。

过滤就是通过网络中节点或连边的单个属性或多个属性的组合筛选满足条件的节点和连边,从而降低节点数量和连边密度。过滤技术更多的是在对网络结构有了一定了解的基础上,对网络数据进行的进一步细化探索,如根据节点的度、介数中心性、中介中心性及连边的权重等属性,有针对性地将满足条件的元素筛选并显示在屏幕中。

抽样是根据一定的抽样策略筛选有代表性的节点和连边,用尽可能少的节点和连边最大限度地反映原始网络的结构特征。常用的抽样策略包括随机选点、随机选边、随机游走和基于拓扑分层模型的抽样策略等[3-4]。

相对于过滤和抽样,压缩则是通过把具有一定相似性的节点或连边聚合,并采用新的节点来代替聚集,在降低节点规模的同时还能够保证网络的完整性和展示网络的子图构成等特征,更有助于用户从中观尺度感知网络整体结构[5-6]。

本文以有效展示网络的中观尺度结构特征为目标开展网络可视化压缩布局方法研究,将力导引布局算法与网络社团结构特征相结合,提出了

收稿日期: 2019-07-09; **录用日期**: 2019-08-12; **网络出版时间**: 2019-08-20 14:01
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190820.1133.003.html

* 通信作者. E-mail: zxthl0707@163.com

引用格式: 吴玲达, 张喜涛, 孟祥利. 基于社团结构节点重要性的网络可视化压缩布局[J]. 北京航空航天大学学报, 2019, 45(12): 2423-2430. WU L D, ZHANG X T, MENG X L. Compression layout for network visualization based on node importance for community structure[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2423-2430 (in Chinese).

一种基于社团结构节点重要性的网络可视化压缩布局方法。将基于拓扑势的社团结构压缩方法应用于网络拓扑结构可视化,通过原始网络的社团构成和社团内部结构等中观尺度结构的清晰展示,辅助用户分析社团和重要节点在网络结构中的位置和作用。

1 相关工作

在网络可视化压缩布局方法中,由于弹簧模型[7]具有简洁高效和易于理解等优势,国内外学者基于弹簧模型对网络可视化压缩布局方法进行了大量探索。该方法将整个网络模拟为弹簧系统,通过计算节点受到的弹力控制节点间的距离最终达到节点的受力平衡。KK(Kamada-Kawai)算法[8]在弹簧模型的基础上引入胡克定律,根据节点受力状态计算系统能量,将节点最优布局问题转化为系统能量最小化的求解问题,使布局过程的收敛速度有了明显的增加。Davidson 和 Harel[9]在 DH(Davidson-Harel)算法中考虑了节点位置、连边长度和连边交叉等多种美学标准的约束来构建能量函数,通过能量函数模型参数可达到不同的布局效果。FR(Fruchterman-Reingold)算法[10]将节点建模为物理系统中的带电粒子,将粒子间的电荷斥力引入弹簧模型,粒子间距越小斥力越大,在一定程度上降低了密集节点的重叠现象,为使算法快速收敛,通过引入“冷却函数”,采用模拟退火算法,逐步降低节点移动步长,使系统能量快速减小,从而实现快速布局的目标[11-12]。

相比于 KK 算法和 DH 算法,FR 算法在大多数网络数据的可视化应用中具有更好的对称性和聚类布局效果,绘制结果更加符合美学标准,得到广泛的应用。

国内外有关网络压缩的研究主要是基于节点或者连边重要性进行压缩。Gilbert 和 Leychenko[5]基于节点的重要性提出了网络压缩的一般框架,该框架指出,通过删除非重要节点及与之关联的连边可以达到压缩网络拓扑的目的。王晓华等[13]基于几何多尺度分析,提出了网络数据和结构的稀疏表示方式,有效帮助人们凭借尽可能少的信息来分析网络。李甜甜等[14]基于复杂网络中的 k-core 概念划分网络节点,利用力导引布局算法实现压缩后的大规模复杂网络数据的布局显示。上述方法基于网络全局对节点的重要性进行评估,选择重要节点作为网络整体结构的代表节点,同时压缩非重要节点,最终实现网络压缩的目的。这类方法从节点尺度(微观尺度)上保留了网络的核心结构,但是忽略了网络的社团结构,不能从中观尺度展示网络的结构特征。

基于数据场理论,文献[15]通过计算节点的拓扑势刻画了节点在网络中受自身和邻居节点的影响多具有的势值,能够更加真实地描述节点在网络结构构成中的重要性。肖俐平等[15]和王戬[16]将其分别应用于多种真实网络数据,并于基于度、介数和 PageRank 等典型的节点重要性测度指标进行对比,指出拓扑势在考虑节点度重要性的同时,还能突出节点在网络中位置的差异性,在反映节点重要性方面更加精细和真实。

类似地,在同一社团结构中,不同节点对社团结构构成的贡献度不尽相同,度值越大并且离社团中心点越近的节点往往比其他节点更重要。因此,本文考虑根据节点的度和节点之间的最短路径长度确定节点在社团结构中的位置,将基于拓扑势的节点重要性评估应用于社团结构,通过选择社团结构中节点重要性高的节点作为社团结构代表节点,实现网络社团结构压缩。

2 网络可视化压缩布局方法

2.1 多粒度社团结构探测

相比于其他社团结构探测算法,Louvain 算法的运算速度更快,可以快速处理具有数以亿计节点的网络,另外基于层次聚类可以输出不同粒度的社团结构划分结果。因此,本文采用 Louvain 算法实现多粒度社团结构探测。

模块度是刻画网络中社团划分质量的重要指标之一,可通过如下公式计算:

$$Q = \frac{1}{2m}\sum_i \sum_j \left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(c_{v_i}, c_{v_j})$$

式中:m 为网络中的连边总数;A_{ij}为节点对(v_i, v_j)连边的权值;k_i 为节点度;c_{v_i}为节点 v_i 隶属的社团;$\delta(c_{v_i}, c_{v_j})$为狄拉克函数,如果节点隶属于同一社团,则为 1,否则为 0。

基于模块度优化的社团结构探测算法属于凝聚算法的一种,其通过优化模块度增益函数不断地凝聚节点,最终获得社团结构划分结果。文献[17]将模块度增量函数定义为

$$\Delta Q = \left[\frac{C_{\text{in}} + K_{i,\text{in}}}{2M} - \left(\frac{C_{\text{out}} + K_i}{2M}\right)^2\right] - \left[\frac{C_{\text{in}}}{2M} - \left(\frac{C_{\text{out}}}{2M}\right)^2 - \left(\frac{K_i}{2M}\right)^2\right]$$

式中:C_{in}为社团 C 中所有内部边权重的总和;C_{out}

为所有与社团 C 中节点邻接的边权重总和；K_i 为所有邻接节点 i 的边权重的总和；$K_{i,\text{in}}$ 为所有从节点 i 与社团 C 中节点邻接的边权重总和；M 为网络中所有边的权重之和。

如图 1 所示，Louvain 算法主要分为 2 个阶段。首先，初始化每个节点为一个社团，不断遍历网络中的所有节点，计算该点加入到各个社团产生的模块度增量，如果模块度增量大于 0，则从这些大于 0 的模块度增量所对应的社团中选出对应模块增量最大的社团，将该点与之合并。重复上述过程，直到网络中社团两两之间不再合并为止。然后，基于第 1 层社团划分结果，将每个社团抽象为“节点”，并构建新的网络，新节点之间的权重是原始网络中社团之间的权重。重复上一阶段的过程，直到社团之间不能再合并为止。

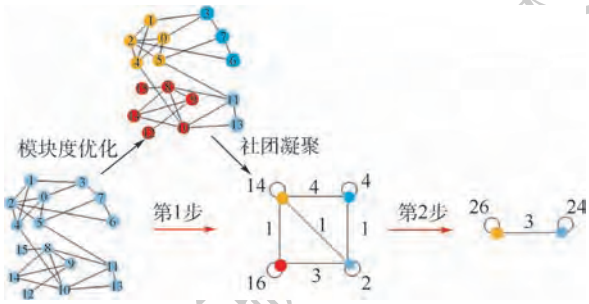


图 1　Louvain 算法示意图

Fig. 1　Schematic diagram of Louvain algorithm

2.2　基于拓扑势的社团结构节点重要性评估

节点拓扑势的大小描述了网络拓扑中的某个节点受自身和近邻节点共同影响所具有的势值。类似地，对于一个给定的社团 $C=(V_C,E_C)$（V_C 为社团 C 中的所有节点集合，E_C 为社团 C 中的所有连边集合），社团中任意一个节点的拓扑势的计算公式为

$$\varphi(v_i)=\sum_{j\neq i}^{|V_C|}k_j\mathrm{e}^{-\left(\frac{d_{ij}}{\delta}\right)^2} \tag{1}$$

式中：$|V_C|$ 为社团 C 中的节点数量；k_j 越大，对节点 v_i 的作用越大；d_{ij} 为节点 v_i 到节点 v_j 的最短路径长度；δ 为影响因子，表示节点之间的相互作用范围。

从式(1)可知，社团中任一节点的势实际上等于社团内其他所有节点对其作用力之和，并且作用力随着到节点距离的增大而逐渐衰减。社团中某点周围的节点越多，距离越短，则该节点的势就越大。因此，可以判断势值最高的节点即为处于社团中心位置的节点，该节点可以看作是社团中心节点；反之，处于社团边缘的节点往往具有的连边较少，势值相对较低。

根据节点的度和节点到社团中心节点的最短路径长度确定的社团结构中节点的拓扑位置反映了节点对社团结构构成的贡献度大小。如图 2 中左图所示，通过计算节点在社团结构中的势来评估节点的重要性，根据重要性排序将社团内部节点分为 3 类：①节点 9 最重要，是拓扑势极值点，也是社团中心点，图 2 中左图用红色标注，用 V_{CA} 表示；②节点 7，19，21，22，25，26，10，16 对拓扑势极值点影响较大，图 2 中左图用黄色标注，用 V_{CB} 表示；③其余节点为边缘节点，图 2 中左图采用灰色标注，用 V_{CC} 表示。选择社团中拓扑势的极值点作为社团代表点，选择对代表点拓扑势贡献大的节点作为相对重要节点。

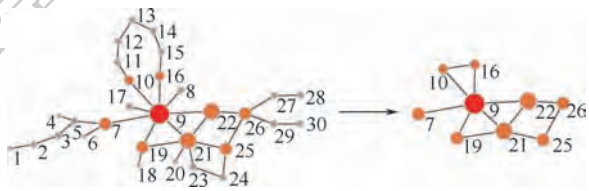


图 2　社团节点分类与社团压缩示意图

Fig. 2　Schematic diagram of community node classification and community compression

2.3　压缩布局算法

在社团结构压缩时，选择尽可能少的节点最大限度地保持社区结构。图 2 中右图为在社团节点分类基础上对社团进行压缩后的社团结构。将社团代表点 V_{CA} 和相对重要节点集合 V_{CB} 作为社团结构代表点集合 V'_{CA}。社团压缩主要是为边缘节点 V_{CC} 设置替代节点，从而将边缘节点与社团结构代表点合并，压缩到网络中只留下社团结构代表点。

对于 $v_i\in V_{CC}$，用 $V_N(v_i)$ 表示节点 v_i 的邻居节点集合，当节点 v_i 满足式(2)时，将节点 v_i 和社区结构代表节点 v_j 压缩为一个节点，并将 v_j 设置为 v_i 的替代节点。

$$v_j=V_N(v_i)\cap V'_{CA} \tag{2}$$

采用广度优先算法为 V_{CC} 集合中的节点查找并设置替代节点，具体方法如算法 1 所示，其中 IfReplaced(v_i) 标记节点 v_i 是否被 V'_{CA} 中的节点代替，用 replace(v_i) 表示节点 v_i 的代替点。

算法 1　广度优先遍历替代节点设置算法。

输入：原始节点 v_i、邻居节点集合 $V_N(v_i)$、社团结构代表点集合 V'_{CA}。

输出：替代节点 replace(v_i)。

1　创建广度优先访问列表 $L(v_i)$，$L(v_i)$.PUSH($V_N(v_i)$)

2　WHILE $L(v_i)$.size > 0 DO

3　　$v_j=L(v_i)$.POP()

```
4     IF (v_j ∈ V'_CA)
5       replace(v_i) = v_j
6       IfReplaced(v_i) = 1
7       BREAK WHILE
8     END IF
9     ELSE
10      L(v_i).PUSH(V_N(v_j))
11    END ELSE
12  END WHILE
```

为 V_{CC} 中的所有节点设置替代节点，生成新的社团结构代表节点集合 V'_C。如算法 1 所示，V_{CC} 中的任一节点 v_i，其替代节点的设置可分为 2 步：

1）如果节点 v_i 的邻居节点 v_j 属于社团结构代表点集合 V'_{CA}，则把节点 v_j 设置为节点 v_i 的替代节点。

2）如果节点 v_j 不是社团结构代表节点，则将节点 v_j 的所有邻居节点添加到访问列表中，遍历查找直到找到节点 v_i 的替代节点。

在对所有社团中的所有边缘节点进行压缩处理后，首先，合并所有新的社团结构代表点集合构建新的网络节点集 V'；然后，遍历原始网络中的所有连边，把连边的端点用替代节点替换，删除重复边，得到压缩网络 $G'=(V',E')$；最后，采用经典的 FR 算法布局压缩网络中的节点，实现网络可视化压缩布局。

3 实验结果与分析

本文选择网络社团分析领域 3 组标准数据集：dolphin、football 及 karat。其中，dolphin 为海豚社会网络，football 为美国大学生足球联赛社会网络，karat 为美国空手道俱乐部社会网络。对应网络的节点数和连边数如表 1 所示。

将本文方法的压缩布局结果分别与基于节点度值（方法 1）、PageRank（方法 2）排序的网络压缩布局方法进行对比。首先，通过节点数、连边数、平均聚类系数和社团数等指标的变化定量评估不同压缩方法的压缩效果；然后，通过对压缩前后拓扑结构的可视化布局效果对比，定性分析本文方法的优势。其中，节点数和连边数的变化量表示网络的压缩规模；平均聚类系数的变化描述了网络中节点间连接关系的紧密程度，平均聚类系数越大，节点连接关系越紧密；社团数的变化反映了压缩网络是否保留了原始网络的社团构成。

实验中选择的节点压缩比为 0.2，即方法 1 和方法 2 按照全网节点总数的 20% 筛选网络代表节点，本文方法按照每个社团中节点总数的 20% 筛选社团结构代表节点。

表 1 为不同方法的压缩结果对比。可知，不同压缩方法通过合并非重要节点，较大程度地降低了节点和连边数量，实现了网络压缩的目的。

另外，压缩网络的平均聚类系数较压缩前有较大程度的升高，也表明了上述 3 种压缩方法都保留联系比较紧密的重要节点，这些节点及其之间的连边反映了网络的骨架结构。其中，基于节点度值（方法 1）、PageRank（方法 2）排序的压缩网络的平均聚类系数相对接近，且都大于本文方法。这是因为前 2 种方法虽然采用的节点重要性评估模型不同，节点重要性排序结果并不一样，但对于整体网络而言，重要节点都分布在排名靠前的区域。因此，基于网络全局节点重要性的压缩获得了相同或基本相同的代表节点集合。而本文方法对不同社团结构中的节点重要性分别排序，保留的节点是各个社团结构中的重要节点，代表了各社团的内部结构。虽然本文方法中压缩网络的平均聚类系数相对较低，但是社团数量没有减少，比较完整地保留了网络的社团构成，从中尺度上保留了网络的整体结构。而前 2 种方法都丢失了部分社团，造成网络整体结构的不完整。

图 3 ~ 图 5 为分别采用上述 3 种方法对不同网络进行可视化压缩布局前后的效果图。将不同方法的布局效果进行对比，可以看出本文方法的优势主要体现在以下 2 个方面：

1）基于社团结构进行压缩，能够保留网络的社团构成，突出中观尺度结构特征。

图 3（b）、（c）、（d）分别为方法 1、方法 2 和本文方法的 dolphin 压缩网络的布局结果，图 4（b）、（c）、（d）分别为方法 1、方法 2 和本文方法的 football 压缩网络的布局结果。与图 3（a）和图 4（a）中

表 1 压缩结果对比

Table 1 Comparison of compression results

数据集	方法	节点数	连边数	平均聚类系数	社团数
dolphin	压缩前	62	159	0.303	5
	方法 1	12	29	0.716	3
	方法 2	12	29	0.716	3
	本文方法	13	33	0.674	5
football	压缩前	115	613	0.403	10
	方法 1	23	100	0.515	9
	方法 2	23	96	0.570	8
	本文方法	26	97	0.534	10
karat	压缩前	34	78	0.588	4
	方法 1	7	17	0.867	3
	方法 2	7	17	0.867	3
	本文方法	8	18	0.733	4

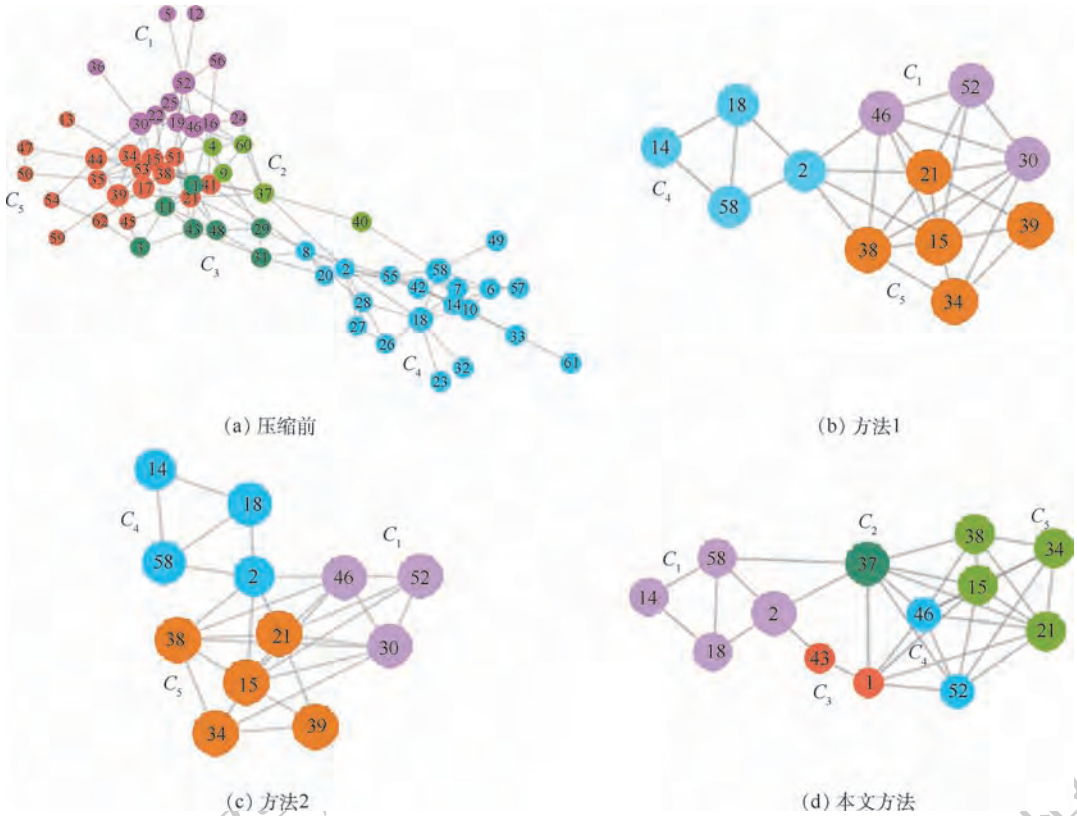


图 3　dolphin 网络压缩前后布局效果

Fig. 3　Layout effect of dolphin network before and after compression

的原始网络布局结果对比,图 3(b)、(c)、(d)和图 4(b)、(c)、(d)中的节点数和连边数都有较大程度的压缩。图 3(a)和图 4(a)中的原始网络受屏幕尺寸限制,节点较小,连边密集,节点和连边重叠交叉现象严重。在混乱重叠的视图中,用户难以感知网络的整体结构。而上述 3 种方法通过缩减节点和连边规模,可以降低节点和连边重叠覆盖造成的视觉杂乱。但是,由于方法 1(图 3(b)和图 4(b))和方法 2(图 3(c)和图 4(c))基于全局结构评估节点的重要性和选择网络代表节点,对节点数量较少或者连边相对稀疏的社团没有保留有效的代表节点。在压缩过程中,这部分社团被压缩合并,造成社团数量减少,无法准确展示原始网络的社团构成。而本文方法(图 3(d)和图 4(d))基于社团结构进行压缩,保证每个社团结构都有代表节点,避免压缩过程中的社团丢失现象,能够反映网络在社团尺度(中观尺度)上的结构特征。

另外,基于社团结构的压缩更有助于感知网络整体结构和社团间的相互作用。以图 3(d)为例,可以发现,原始网络可划分为 5 个社团。其中,社团 C_1、C_5 包含节点数量较多,其他 3 个社团节点数量较少;社团 C_1、C_5 间的交互主要通过 C_2、C_3 这 2 个社团实现。C_2、C_3 社团中的节点虽少,但是承担着全网节点交互的“桥梁”作用,对网络的拓扑构成十分重要。而图 3(b)、(c)中没有保留 C_2、C_3 社团,不能准确反映 C_2、C_3 社团的“桥梁”作用,C_4 直接与 C_1 交互,容易对用户理解网络在中观尺度上的结构特征造成偏差。

2) 采用多粒度社团结构划分算法,展示社团基本结构,并实现网络的多粒度压缩布局。

本文方法在压缩边缘节点的同时保留了必要的社团结构代表节点。由于基于拓扑势对节点的重要性进行分析,保留的节点对社团构成的贡献较大,反映了社团的基本结构。通过保留这部分节点及连接关系,可以清晰地展示社团内部基本结构。同时,通过压缩非重要节点和删除重复连边,降低社团内部连边数量,有助于在视觉上感知社团结构,发现社团或网络结构中的重要节点。

根据社团结构划分的粒度,本文方法通过压缩合并边缘节点,有效降低网络规模、节点数和连边数,实现多粒度的压缩布局。图 5 为对 karat 网络数据进行 3 级压缩布局的效果图。如图 5(d)所示,根据节点压缩比例的不同可以控制网络压缩比例,最多可将一个社团压缩为 1 个节点。如图5(d)中每个节点代表一个社团,“节点”1、32、

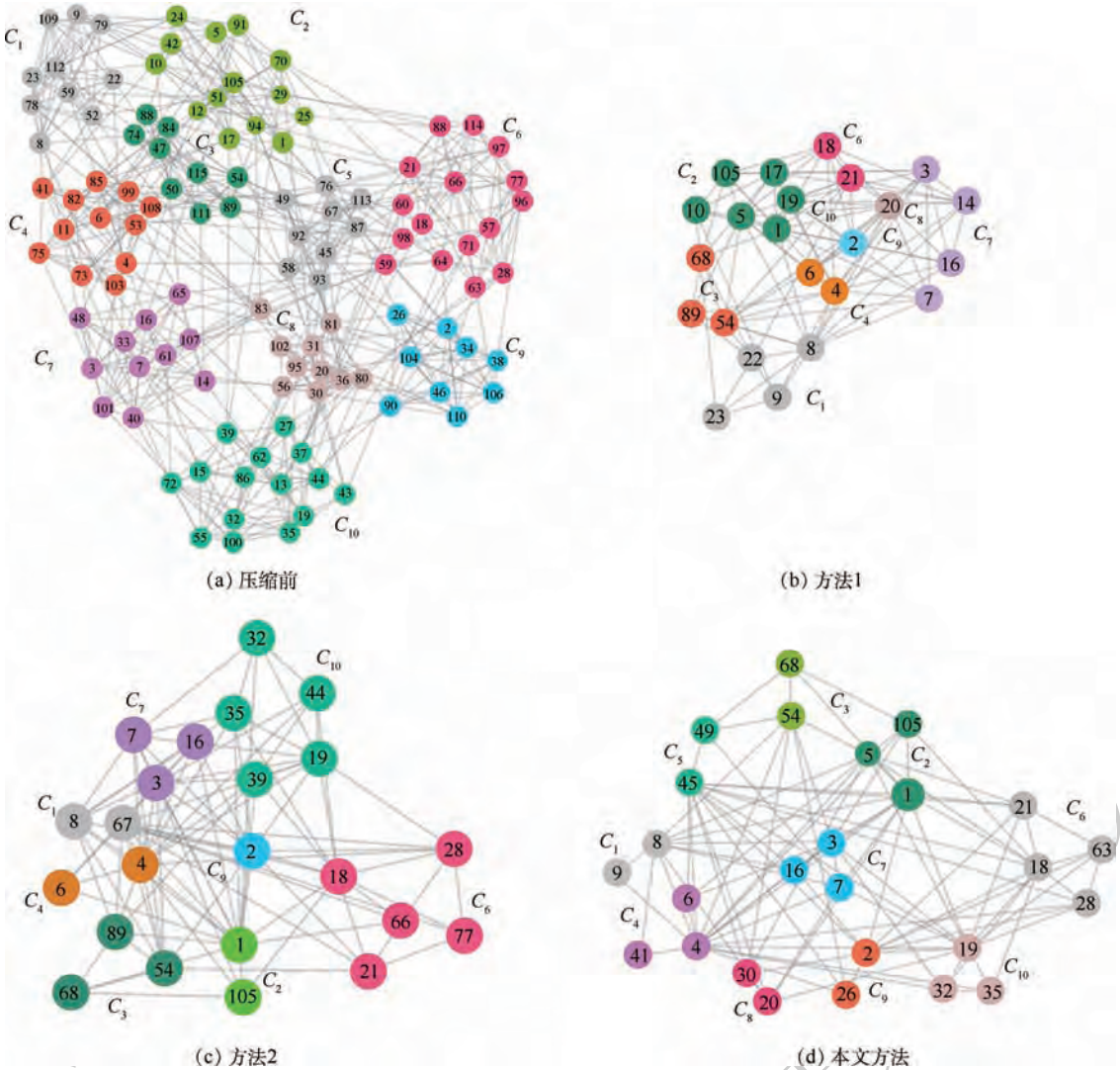


图 4 football 网络压缩前后布局效果

Fig. 4 Layout effect of football network before and after compression

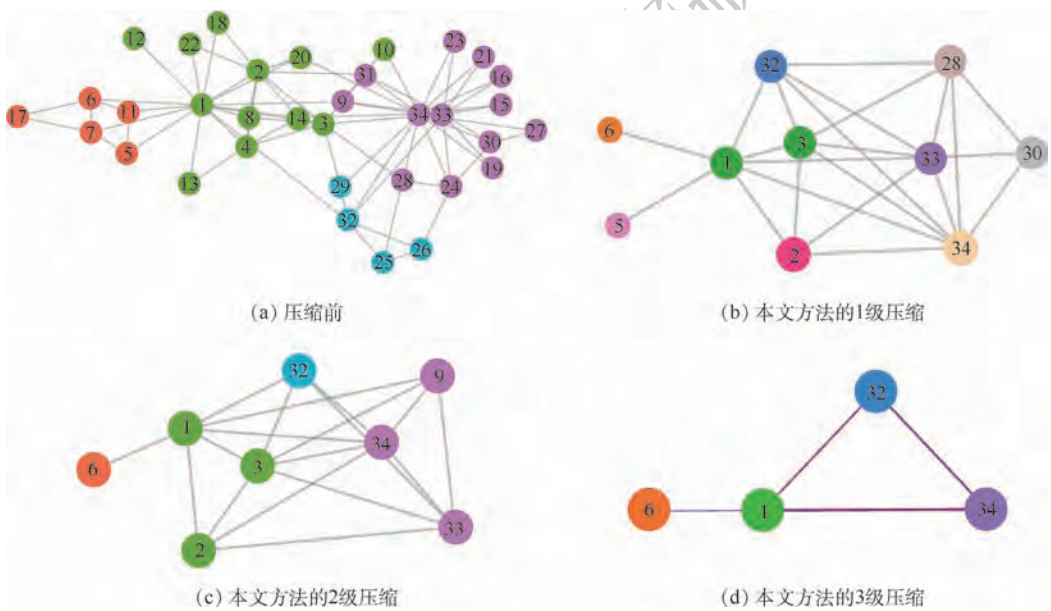


图 5 采用本文方法对 karat 网络进行多粒度压缩前后布局效果

Fig. 5 Layout effect of karat network before and after multi-granularity compression by proposed method

34 之间能够直接交互,而“节点”6 只能通过“节点”1 与其他“节点”交互。

4 结　论

为有效展示网络的中尺度结构特征,将力导引布局算法与网络社团结构特征相结合,提出了一种基于社团结构节点重要性的网络可视化压缩布局方法。实验结果和分析表明,该方法的优势体现在 3 个方面:

1)有效压缩节点和连边数量,降低视觉杂乱现象。

2)比较完整地从中观尺度反映网络的结构构成与交互,便于分析不同社团或节点在网络拓扑中的不同作用。

3)基于多粒度的社团结构探测和不同的节点压缩比可以实现多粒度的压缩展示。本文方法主要通过将网络压缩方法和力导引布局算法结合,从社团构成和社团结构上展示了网络的中观尺度结构,下一步将考虑与其他节点布局算法和交互方法结合,进一步优化布局结果。

参考文献(References)

[1] 张喜涛,吴玲达,于少波. 面向多层网络可视化的多力导引节点自动布局算法[J]. 计算机辅助设计与图形学学报,2019,31(4):639-646.
ZHANG X T,WU L D,YU S B. A multi-force directed layout algorithm for multilayer networks visualization[J]. Journal of Computer-Aided Design & Computer Graphics,2019,31(4):639-646(in Chinese).

[2] 姚中华,吴玲达,宋汉辰. 网络图中边集束优化问题[J]. 北京航空航天大学学报,2015,41(5):871-878.
YAO Z H,WU L D,SONG H C. Problems of network simplification by edge bundling[J]. Journal of Beijing University of Aeronautics and Astronautics,2015,41(5):871-878(in Chinese).

[3] LESKOVEC J, FALOUTSOS C. Sampling from large graphs[C]// Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2006:631-636.

[4] 杜晓林. 大规模社会网络可视化若干问题及算法研究[D]. 哈尔滨:哈尔滨工业大学,2015:18-35.
DU X L. Research on visualization problems and algorithms for large scale social network[D]. Harbin: Harbin Institute of Technology,2015:18-35(in Chinese).

[5] GILBERT A C,LEVCHENKO K. Compressing network graphs[C]// Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM,2004:1-10.

[6] ALMENDRAL J A,CRIADO R,LEYVA I,et al. Introduction to focus issue:Mesoscales in complex networks[J]. Chaos,2011,21(1):016101.

[7] EADES P. A heuristic for graph drawing[J]. Congressus Numerantium,1984,42:149-160.

[8] KAMADA T,KAWAI S. An algorithm for drawing general undirected graphs[J]. Information Processing Letters,1989,31(1):7-15.

[9] DAVIDSON R,HAREL D. Drawing graphs nicely using simulated annealing[J]. ACM Transactions on Graphics,1996,15(4):301-331.

[10] FRUCHTERMAN T M J,REINGOLD E M. Graph drawing by force-directed placement[J]. Software-Practice and Experience,1991,21(11):1129-1164.

[11] 水超,陈洪辉,陈涛,等. 力导向模型的复杂网络社区挖掘算法[J]. 国防科技大学学报,2014,36(4):163-168.
SHUI C,CHEN H H,CHEN T,et al. A community detect algorithm on force-directed model[J]. Journal of National University of Defense Technology,2014,36(4):163-168(in Chinese).

[12] ZHANG X T,WU L D,YU S B. Layer-edge patterns exploration and presentation in multiplex networks:From detail to overview via selections and aggregations[J]. Electronics,2019,8(4):387.

[13] 王晓华,杨新艳,焦李成. 基于多尺度几何分析的复杂网络压缩策略[J]. 电子与信息学报,2009,31(4):968-972.
WANG X H,YANG X Y,JIAO L C. Compression of complex networks based on multiscale geometric analysis[J]. Journal of Electronics & Information Technology,2009,31(4):968-972(in Chinese).

[14] 李甜甜,卢罡,许南山,等. 基于 k-core 的大规模复杂网络压缩布局算法[J]. 计算机工程,2016,42(5):308-312.
LI T T,LU G,XU N S,et al. Compressing layout algorithm for large complex network based on k-core[J]. Computer Engineering,2016,42(5):308-312(in Chinese).

[15] 肖俐平,孟晖,李德毅. 基于拓扑势的网络节点重要性排序及评价方法[J]. 武汉大学学报(信息科学版),2008,33(4):379-383.
XIAO L P,MENG H,LI D Y. Approach to node ranking in a netwrok based on topology potential[J]. Geomatics and Information Science of Wunan University,2008,33(4):379-383(in Chinese).

[16] 王戮. 一种基于拓扑势的社会网络节点重要性评估方法[D]. 哈尔滨:哈尔滨工程大学,2015:18-30.
WANG J. Evaluating the importance of nodes in social networks based on topology potential[D]. Harbin: Harbin Engineering University,2015:18-30(in Chinese).

[17] GACH O,HAO J K. Improving the Louvain algorithm for community detection with modularity maximization[C]// International Conference on Artificial Evolution (Evolution Artificielle). Berlin:Springer,2013:145-156.

作者简介:

吴玲达　女,研究员,博士生导师。主要研究方向:虚拟战场环境构建、信息系统建模与仿真、多媒体与虚拟现实技术。

张喜涛　男,博士,讲师。主要研究方向:网络可视化。

孟祥利　男,博士研究生。主要研究方向:信息系统建模与仿真。

Compression layout for network visualization based on node importance for community structure

WU Lingda[1], ZHANG Xitao[1,2,*], MENG Xiangli[1]

(1. School of Space Information, Space Engineering University, Beijing 101416, China;
2. Department of Space Command, Space Engineering University, Beijing 101416, China)

Abstract: In order to effectively display the mesoscale structure of the network, the force-directed layout algorithm is combined with the network community structure features, and a network visualization compression layout method based on the importance of the node in the community structure is proposed. First, the Louvain algorithm is used to divide the network into multi-granular community structure. Then, the importance of the nodes is evaluated by calculating the topological potential of the nodes in the community structure. Through preserving the important nodes, the community is compressed, while the boundary nodes are merged. Finally, the force-directed layout algorithm is adopted to layout the network to achieve a visual compression layout. The experimental results show that the proposed method can completely preserve the original network community structure on the basis of compression nodes and connected edges, and can clearly display the internal structure of the community by retaining the representative points of the community structure, highlighting the position and role of the community and important nodes in the network structure.

Keywords: network visualization; compression layout; force-directed layout; topological potential; node importance

Received: 2019-07-09; **Accepted**: 2019-08-12; **Published online**: 2019-08-20 14:01
URL: kns.cnki.net/kcms/detail/11.2625.V.20190820.1133.003.html
Foundation item: Fund for Weapons and Equipment Pro-Research (6142010010301)
* **Corresponding author**. E-mail: zxthl0707@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0361

基于蒙特卡罗频率法的葡萄籽总酚含量高光谱测量变量选择

成云玲[1,2], 杨蜀秦[1,2,*]

(1. 西北农林科技大学 机械与电子工程学院, 咸阳 712100; 2. 农业农村部农业物联网重点实验室, 咸阳 712100)

摘 要: 在利用高光谱建立葡萄籽总酚含量的预测模型中,为解决变量过多、模型复杂度高等问题,需依据光谱特点进行有效地数据降维。提出了一种蒙特卡罗频率法(MCF)对高光谱数据进行波长选择,并建立了葡萄籽总酚的支持向量回归(SVR)预测模型。该方法首先采用蒙特卡罗采样(MCS)选择波长子集;然后建立大量 SVR 子模型,并选出均方根误差(RMSE)较小的子模型,统计每个波长出现的频次;最后根据指数递减函数确定波长个数,选取频次最高的波长子集作为特征波长。结果表明,采用 MCF 可以在降维的同时提高模型的预测性能,波长数目由原始的 196 个减少到 9 个,波长范围均在 950 ~ 1 400 nm,RMSE 值从 0.42 减少到 0.37,预测精度优于 SPA 等其他波长选择方法。因此,提出的基于 MCF 在高光谱数据处理中能有效选择特征波长,为准确建立预测模型提供了一种有效的方法。

关 键 词: 变量选择; 蒙特卡罗采样(MCS); 近红外高光谱; 葡萄籽; 总酚含量
中图分类号: S663.1
文献标识码: A **文章编号**: 1001-5965(2019)12-2431-07

近红外高光谱技术是利用物质在近红外光谱区特定的吸收特性,对样品中一种或多种化学成分进行快速检测的方法[1-2]。由于其成本低、速度快、无损检测等优点,已被广泛应用于食品领域[3-5]。近红外高光谱数据具有谱带宽、信号弱和重叠严重的特点,一般由几百到几千个波段组成,相邻波段之间共线性严重,并且包含有大量的冗余信息[6-7]。因此,特征波长选择对于简化模型,提高模型的预测精度和鲁棒性具有重要意义[8-9],这也使得波长选择成为近红外分析领域的一个热点研究课题。

目前常用的变量选择方法有连续投影算法(Successive Projections Algorithm, SPA)[10]、无信息变量消除(Uninformative Variable Elimination, UVE)法[11-12]等。SPA 是一种基于变量投影比较的特征波长选择方法。其通过比较某波长在其他波长上的投影,选择投影向量最大的波长作为待选波长,然后基于校正模型的均方根误差(Root Mean Square Error, RMSE)从待选波长集合中选择最终的特征波长。UVE 在保留特征波长的同时消除无信息变量,是一种基于偏最小二乘(Partial Least Squares, PLS)模型回归系数的波长选择方法。该方法引入稳定性来评价模型中各变量的可靠性,从而确定最终选择的变量,在光谱变量的选择中得到了广泛的应用。这些方法的一个共同特点是,它们试图为给定的数据集选择一个固定的变量子集,而不考虑样本变化对变量选择的影响。

收稿日期: 2019-07-08; **录用日期**: 2019-08-03; **网络出版时间**: 2019-08-12 16:53
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190812.1553.005.html
基金项目: 国家自然科学基金(31501228,61876153); 中央高校基本科研业务费专项资金(2452019180)
* **通信作者**. E-mail: yangshuqin1978@163.com

引用格式: *成云玲, 杨蜀秦. 基于蒙特卡罗频率法的葡萄籽总酚含量高光谱测量变量选择[J]. 北京航空航天大学学报, 2019, 45(12): 2431-2437. CHENG Y L, YANG S Q. Selection of measurement variables for hyperspectra of total phenol content in grape seeds based on Monte Carlo frequency method[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2431-2437 (in Chinese).*

结合蒙特卡罗采样(Monte Carlo Sampling, MCS)技术建立变量选择方法可以有效地解决这一问题。例如,竞争自适应重采样(Competitive Adaptive Reweighted Sampling, CARS)[13]基于 MCS 和 PLS 回归系数选择特征变量。其首先通过 MCS 建立 PLS 模型,然后通过自适应加权采样保留模型中回归系数绝对值权重较大的波长作为新的子集,基于新的波长子集重新建立 PLS 模型,经过多次计算,选择交叉验证均方根误差(Root Mean Square Error of Cross Validation, RMSECV)最小的波长子集作为特征波长,降维性能较好。此外,其他变量选择方法包括蒙特卡罗无信息变量消除(Monte Carlo-Uninformative Variable Elimination, MC-UVE)[14]、模型种群分析(Model Population Analysis, MPA)[15]和变量互补网络(Variable Complementary Network, VCN)[16]等方法表明,结合 MCS 进行变量选择可以得到更好的预测结果。

本文搭建了葡萄籽总酚含量近红外高光谱预测系统,根据模型集群分析的思想,提出了将 MCS 和波长出现频次结合选择特征波长的方法,简称蒙特卡罗频率法(Monte Carlo Frequency, MCF)。该方法能够减少建模过程中的无信息变量及干扰变量,为开发葡萄籽总酚含量检测设备提供理论依据。

1 材料与方法

1.1 样本采集

试验样品来自陕西省杨凌盛唐酒庄,包括霞多丽、贵人香、8802、8803 和雷司令 5 个白葡萄品种。采摘工作从葡萄转色期至成熟期进行,由于不同品种酿酒葡萄的成熟时间存在差异,采摘时间为 2015 年 7 月中旬至 9 月中旬。每个品种从转色期一周后开始,15 天作为一个采摘周期,共采摘 3 次。每次采摘 4 组葡萄,每组 20 个葡萄(籽)作为一个样本。因此,每个品种包括 12 个样本,总计 60 个样本。将采摘的样本快速运送到实验室,手工将葡萄籽分离出来,拍摄其近红外高光谱图像,随后用蛋白质沉淀法[17]测量其总酚含量。

原始图像具有 256 个波长,去除两端包含噪音的波段,最终选择 950 ~ 1 600 nm 之间 196 个波长对应的光谱数据。为了去除高频随机噪声等干扰因素[18],采用 S-G 滤波[19]对数据进行预处理,预处理后的平均光谱如图 1 所示。随机将样本按 4∶1 的比例分为训练集和测试集,葡萄籽总酚含量分布如表 1 所示。

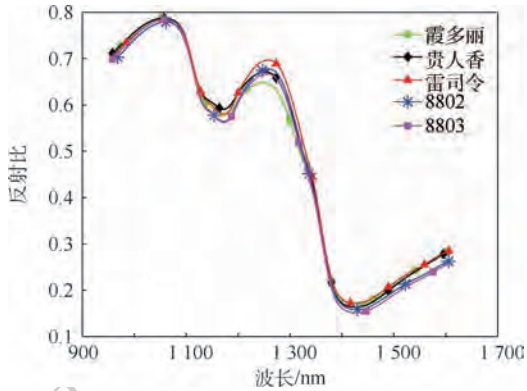


图 1 预处理后的 5 个品种葡萄籽平均光谱

Fig. 1 Average spectra of five types of pretreated grape seeds

表 1 葡萄籽总酚含量分布统计

Table 1 Distribution statistics of total phenol content in grape seeds

参数	训练集	预测集
样本数	48	12
最小值/(g·L^{-1})	1.720 8	2.192 7
最大值/(g·L^{-1})	8.386 9	8.326 3
平均值/(g·L^{-1})	4.531 9	4.972 3
标准偏差/(g·L^{-1})	1.642 9	1.858 6

1.2 仪器设备

本文采用 HyperSIS 高光谱成像系统,包括 IMSpector N17E 型近红外成像光谱仪(Spectral Imaging Ltd., Finland)、320 像素 × 256 像素的 XEVA3616 型面阵 CCD 相机(XenICs Ltd., Belgium)、白光漫反射型卤钨白炽灯(4 个)、高精度电动平移载物台和一台计算机。该系统的光谱分辨率为 2.8 nm,采集的波长范围为 900 ~ 1 700 nm,平台移动速度为 20 mm/s,相机曝光时间为 10 ms。本文方法均在 MATLAB R2017a 中实现。

1.3 预测模型

本文采用支持向量回归(Support Vector Regression, SVR)建立葡萄籽总酚的预测模型,用于比较 MCF 和其他常用变量选择方法的性能。SVR 是一种适用于解决小样本、非线性及高维数据问题的方法[20-21]。SVR 通过核函数,将低维空间向量映射到高维空间,在高维空间中构造线性决策函数来实现原空间中的非线性决策。考虑到高光谱数据和预测变量总酚含量之间映射关系的复杂性和非线性,利用 SVR 建模在一定程度上规避了过拟合风险。本文使用 Libsvm[22]实现 SVR 方法。

1.4 评价指标

本文采用相关系数 R^2 和 RMSE 作为回归模型的评价指标。R^2 越高,RMSE 越低,表明模型的

效果越好。R^2和 RMSE 的计算公式分别为

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y'_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y}_i - y_i)^2} \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(y'_i - y_i)^2}{n-1}} \tag{2}$$

式中:y'_i 和 y_i 分别为第 i 个样本的预测值和真实值;$\bar{y}_i$ 为样本的平均值;n 为样本个数。

2　基于蒙特卡罗频率法选择波长变量

蒙特卡罗方法是一种基于随机数和概率统计来研究问题的技术。本文提出的 MCF 是一种基于 MCS 和波长出现频次的变量选择方法,该方法可以和多种回归方法结合,能够有效选择特征变量。

2.1　模型集群分析基本思想

模型集群分析[15]方法是首先通过 MCS 获取数据子集;然后针对每个子数据集,建立一个子模型;最后从样本空间、变量空间、参数空间或者模型空间中,对所有建立的集群子模型的参数进行统计分析,以获得有用信息。

2.2　基本原理

MCF 选择特征波段主要采用 MCS 选择波长子集,然后利用波长子集建立大量回归子模型;选择 RMSE 较小的子模型,统计每个波长出现的频次;根据指数递减函数选择波长个数,选取频次最高的波长作为特征波长,具体步骤如下:

1) MCS 选择波长子集。设样本的光谱矩阵 $\boldsymbol{X}$ 为 $n \times q$,表示矩阵由 n 个样本和 q 个波长组成;化学值 $\boldsymbol{Y}$ 由向量 $n \times 1$ 表示。根据模型集群分析的建模思想,首先采用 MCS 对所有波长进行采样,每次随机选择 $p(p<q)$ 个波长,可得到 $n \times p$ 的子光谱矩阵。将该过程重复 N 次($N>1\,000$),得到 N 个子数据集$(\boldsymbol{X}_{\mathrm{sub}}, \boldsymbol{Y})_i, i=1,2,\cdots,N$。此过程不仅能得到 N 组不同变量的组合,还能确保每个变量具有相同的采样频率。

2) 将子数据集按 4∶1 随机分为训练集和预测集,建立 N 个回归子模型。SVR 是一种适用于解决小样本及高维数据问题的最常用的建模方法,因此本文采用 SVR 建立预测模型,然后计算 N 个子模型预测集的 RMSE。

3) 计算波长出现的频率。将上述所有子模型按照预测 RMSE 从小到大进行排序,只保留预测结果较好的前 K 个子模型,计算这些模型中各个波段出现的频次 f。一般波段出现频次越高,则认为该波段和化学值相关性越高,根据频次对波长进行重要性排序。f 的计算公式为

$$f(i) = 100 \times \sum_{j=1}^{K} F_i / K \tag{3}$$

式中:i 为波段序号;j 表示保留的子模型;F_i 表示波段 i 是否出现在模型 j 中,若出现则为 1,否则为 0;K 为保留子模型的个数。

4) 根据指数递减函数选择波长个数。建立 m 个 SVR 回归模型,根据模型的预测 RMSE 选择最佳的特征波长个数。指数递减函数[13]定义为

$$r_i = \partial \mathrm{e}^{-ki} \tag{4}$$

式中:r_i 为第 i 次选择的波长个数;∂ 和 k 是由以下 2 个条件决定常数:

1) 在第一次运行中,所有 q 个波长都被用来建模。由于本文共采用了波段裁剪后的 196 个原始波长,因此 $r_1 = 196$。

2) 在第 m(本文取 $m=40$)次运行中,只保留 2 个波长,即 $r_{40} = 2$。

在这 2 个条件下,∂ 和 k 的计算公式分别为

$$\partial = q\left(\frac{q}{2}\right)^{1/(m-1)} = 107.654\,4 \tag{5}$$

$$k = \frac{\ln(q/2)}{m-1} = 0.177\,6 \tag{6}$$

图 2 所示为指数递减函数选择波长个数的过程。可以看出,波长个数呈递减趋势,并分为 2 个阶段,第 1 阶段波长数量下降较快,可快速去除出现频次少的波长;第 2 阶段波长数量下降缓慢,可有效保留出现频次较高的波段。

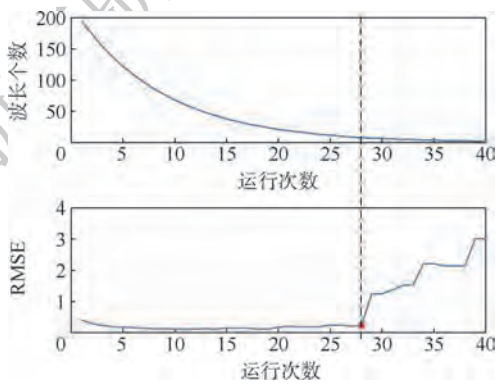


图 2　MCF 特征波长个数选择

Fig. 2　Selection of number of characteristic wavelengths by MCF

3　结果与分析

3.1　不同波长选择方法的预测结果

为了验证提出方法的有效性,将 MCF 与

SPA、CARS 等 2 种方法分别结合 SVR 方法，预测葡萄籽中的总酚含量。

MCF 通过 MCS，每次从训练集中随机选择 100 个波段构建回归子模型。重复 2 000 次，并计算各子模型的 RMSE。然后，对所有子模型进行排序，根据 RMSE 值由小到大，分别选择前 10%、20%、30%、40% 和 50% 的子模型，计算模型中每个波长出现的频率。根据式(4)的指数递减函数选择 RMSE 较小的波长数作为特征波长个数(见图 2)，并找出出现频率最高的波长子集作为特征波长。实验表明，前 30% 的子模型预测性能最佳，因此，本文保留前 30% 子模型用于计算波长频率。各波段频次分布如图 3 所示，最终 MCF 选择 9 个特征波长。

在 SPA 降维算法中，设置波长个数范围为 2～49，根据训练集 RMSE 值确定最佳的光谱变量总数。当波长数量较少时，RMSE 值较大，随着波长个数的增加，RMSE 开始呈下降趋势，当选取 18 个波长时达到最小值。因此，SPA 最终选择的波长个数为 18。

图4为采用CARS进行特征波长选择后，各波长的回归系数路径，设置采样次数为 50 次。每条线反映一个波长系数的变化。星号线处的临界点表示 RMSECV 的最优子集，星号之后由于有效波长的去除，RMSECV 值开始增大。根据 RMSECV 值最小的原则，CARS 共选择了 7 个特征波长。

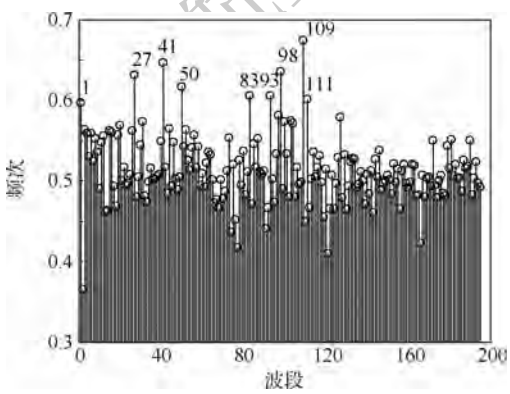


图 3 波段频次分布

Fig. 3 Frequency distribution of spectral bands

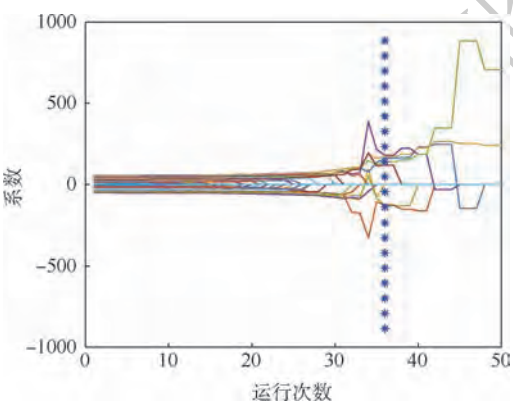


图 4 CARS 方法波长的系数变化

Fig. 4 Coefficient variation of wavelength by CARS

图 5 为 3 种方法选择的变量分布，直观地给出了方法所选变量的波长分布。可以看出，3 种方法波长选择区间大致相同，主要集中在950～1 400 nm。由于高光谱图像光谱分辨率高，光谱曲线几乎连续分布，相邻波长之间数据相关性强，而 MCF 的特征波长分布较为均匀，说明该方法在去除冗余信息方面具有优势。此外，SPA 选取 18 个特征波长，波长个数最多。CARS 选择波长个数最少，其光谱包含的信息量少，因此可能导致模型效果不理想。可进一步根据回归模型的 R^2 和 RMSE 比较 3 个方法的优劣。

以总酚含量为因变量($\boldsymbol{Y}$)，基于波长选择的光谱为自变量(矩阵 $\boldsymbol{X}$)构建 SVR 模型。采用高斯径向基函数(Radial Basis Function，RBF)作为 SVR 的核函数，通过网格寻优算法找到使分类模型最佳的惩罚函数 c 和核函数参数 g，c 和 g 寻优范围取[2^{-8}，2^{16}]。通过训练集和预测集的 R^2 和 RMSE 对模型性能进行评价。

不同降维方法对葡萄籽总酚含量预测结果如表2所示。可以看出，全光谱模型的预测 R^2 和

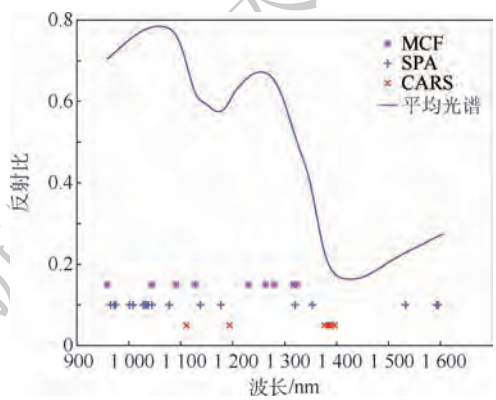


图 5 3 种方法选择的变量分布

Fig. 5 Distribution of variables selected by three methods

表 2 不同降维方法的总酚预测结果比较

Table 2 Comparison of total phenol prediction results with different dimensionality reduction methods

方法	变量数	(c,g)	训练集		预测集	
			R^2	RMSE	R^2	RMSE
SVR	196	(256,84.45)	0.952 4	0.133 8	0.900 4	0.416 3
MCF-SVR	**9**	(**362**,**2 048**)	**0.924 4**	**0.204 9**	**0.905 9**	**0.374 1**
SPA-SVR	18	(2 896,362)	0.920 8	0.216 0	0.886 7	0.476 7
CARS-SVR	7	(256,1 448)	0.812 9	0.504 7	0.791 3	0.829 4

RMSE 分别约为 0.90 和 0.42,表明总酚含量与高光谱数据高度相关。对比 3 种方法,MCF 降维后模型具有最大的预测 R^2(0.91)和最小的 RMSE(0.37),预测结果最好。该方法分别选择了 958、1 044、1 091、1 127、1 230、1 264、1 280、1 317 和 1 323 nm处的特征波长。SPA 选取波段个数最多,其模型结果略低于 MCF,预测 R^2 达到 0.89。CARS 选取波长个数最少,同时预测效果最差,预测相关系数均小于 0.80。此外,MCF 降维后的波长预测结果优于全波段,说明该波长选择方法可以提高模型的预测准确度。

3.2　MCF 性能影响因素

3.2.1　采样次数

为了研究 MCS 次数对 MCF 性能的影响,将波长采样次数分别设置为 1 000、2 000、3 000、4 000和 5 000 次,建立 SVR 子模型并统计各模型的预测 RMSE 值,箱型图如图 6 所示。由图可得,不同采样次数下 N 个模型的 RMSE 最大值、最小值和中值接近,RMSE 分布无明显差别。结果表明,MCS 次数对 MCF 的性能没有显著影响。因此,本文采用 2 000 次作为默认波长采样次数。

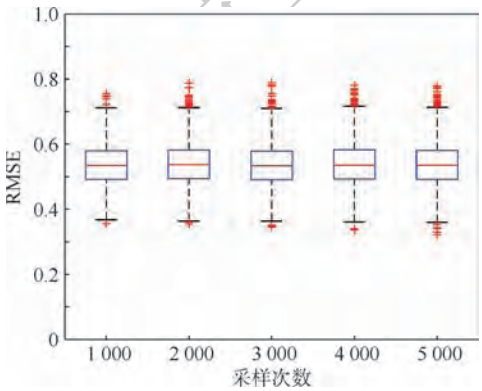


图 6　MCF 不同采样次数的箱型图

Fig. 6　Box graph of MCF with different sampling times

3.2.2　MCF 结合不同回归方法

对 MCF 结合不同回归方法的性能进行比较。除 SVR 之外,还采用最小二乘回归(Partial Least Squares Regression,PLSR)法、RBF 神经网络建立子模型选择特征波段。为了比较不同回归方法的波段选择效果,采用蒙特卡罗对波长采样 2 000 次,分别用 SVR、PLSR 和 RBF 这 3 种方法建立回归子模型,选择出现频次最高的前 9 个波长作为特征波长。用特征波长建立葡萄籽总酚的 SVR 预测模型,表 3 为不同模型的预测结果。MCF 结合 3 种回归方法进行波段选择,预测 R^2 达到 0.85 ~ 0.91,RMSE 约为 0.37 ~ 0.55。而其中采用 SVR 建立子模型进行波段选择时,预测效果最佳。

表 3　MCF 结合不同回归方法的总酚预测结果比较

Table 3　Comparison of total phenol prediction results of MCF combined with different regression methods

回归方法	(c,g)	训练集		预测集	
		R^2	RMSE	R^2	RMSE
SVR	**(362,2 048)**	**0.924 4**	**0.204 9**	**0.905 9**	**0.374 1**
PLSR	(20.55,5 793)	0.884 7	0.366 2	0.853 9	0.546 6
RBF	(862,724)	0.872 2	0.340 6	0.875 6	0.428 9

4　结　论

本文提出了一种基于 MCS 和波长出现频次结合的变量选择方法,简称蒙特卡罗频率法(MCF)。

1)针对葡萄籽总酚近红外高光谱,利用 MCF 进行特征波长选择,波长数目由 196 个减少到 9 个。

2)采用 SVR 建立总酚的回归模型,预测 R^2 和 RMSE 分别约为 0.91 和 0.37。与其他变量选择方法相比,MCF 在减少无信息变量和干扰变量的同时提高了模型的预测结果。

3)讨论了 MCS 次数和不同回归方法对 MCF 性能的影响。结果表明,采样次数对 MCF 波长选择无显著影响,采用 SVR 建立子模型进行波段选择时,模型效果最佳。

因此,MCF 可以作为一种有效的波长选择工具应用于高光谱数据分析,具有良好的预测性能。

致谢　感谢西北农林科技大学葡萄酒学院刘旭副教授及其团队在样本采集和葡萄籽总酚含量测量中的贡献。

参考文献(References)

[1] 张保华,李江波,樊书祥,等. 高光谱成像技术在果蔬品质与安全无损检测中的原理及应用[J]. 光谱学与光谱分析,2014,34(10):2743-2751.
ZHANG B H,LI J B,FAN S X,et al. Principle and application of high spectral imaging technology in nondestructive testing of fruit and vegetable quality and safety[J]. Spectroscopy and Spectral Analysis,2014,34(10):2743-2751(in Chinese).

[2] CHEN S,ZHANG F,NING J,et al. Predicting the anthocyanin content of wine grapes by NIR hyperspectral imaging[J]. Food Chemistry,2015,172:788-793.

[3] EIMASRY G,SU D W,ALLEN P,et al. Near-infrared hyperspectral imaging for predicting colour,pH and tenderness of fresh beef[J]. Journal of Food Engineering,2012,110(1):127-140.

[4] LEO L,ROGER J M,HERRERO-LANGREO A,et al. Comparison of multispectral indexes extracted from hyperspectral images

for the assessment of fruit ripening[J]. Journal of Food Engineering,2011,104(4):612-620.

[5] GMES V M,FERNANDES A M,FAIA A,et al. Comparison of different approaches for the prediction of sugar content in new vintages of whole port wine grape berries using hyperspectral imaging[J]. Computers and Electronics in Agriculture,2017,140:244-254.

[6] LI W,PRASAD S,FOWLER J E,et al. Locality-preserving dimensionality reduction and classification for hyperspectral image analysis[J]. IEEE Transactions on Geoscience & Remote Sensing,2012,50(4):1185-1198.

[7] 宦克为,刘小溪,郑峰,等. 基于蒙特卡罗特征投影法的小麦蛋白质近红外光谱测量变量选择[J]. 农业工程学报,2013,29(4):266-271.
HUAN K W,LIU X X,ZHENG F,et al. Selection of variables for wheat protein near infrared spectroscopy based on monte carlo characteristic projection[J]. Journal of Agricultural Engineering,2013,29(4):266-271(in Chinese).

[8] 郝勇,孙旭东,潘圆媛,等. 蒙特卡罗无信息变量消除方法用于近红外光谱预测果品硬度和表面色泽的研究[J]. 光谱学与光谱分析,2011,31(5):1225-1229.
HAO Y,SUN X D,PAN Y Y,et al. Monte-carlo method of elimination of uninformed variables was used to predict fruit hardness and surface color by near infrared spectroscopy[J]. Spectroscopy and Spectral Analysis,2011,31(5):1225-1229(in Chinese).

[9] 顾章源,刘翔,苏枫,等. 基于流形学习的多光谱优化波段选择算法研究[J]. 上海航天,2017,34(3):40-46.
GU Z Y,LIU X,SU F,et al. Research on multi-spectral optimal band selection algorithm based on manifold learning[J]. Aerospace Shanghai,2017,34(3):40-46(in Chinese).

[10] ARAUJO M C U,SALDANHA T C B,GALVAO R K H,et al. The successive projections algorithm for variable selection in spectroscopic multicomponent analysis[J]. Chemometrics & Intelligent Laboratory Systems,2001,57(2):65-73.

[11] CENTNER V,MASSART D L,NOORD O E D,et al. Elimination of uninformative variables for multivariate calibration[J]. Analytical Chemistry,1996,68(21):3851-3858.

[12] MOROS J,KULIGOWSKI J,QINTAS G,et al. New cut-off criterion for uninformative variable elimination in multivariate calibration of near-infrared spectra for the determination of heroin in illicit street drugs[J]. Analytica Chimica Acta,2008,630(2):150-160.

[13] LI H,LIANG Y,XU Q,et al. Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration[J]. Analytica Chimica Acta,2009,648(1):77-84.

[14] CAI W,LI Y,SHAO X. A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra[J]. Chemometrics & Intelligent Laboratory Systems,2008,90(2):188-194.

[15] LI H,LIANG Y,XU Q,et al. Model population analysis for variable selection[J]. Journal of Chemometrics,2010,24(7-8):418-423.

[16] LI H,XU Q,ZHANG W,et al. Variable complementary network:A novel approach for identifying biomarkers and their mutual associations[J]. Metabolomics,2012,8(6):1218-1226.

[17] HARBERTSON J F,PICCIOTTO E A,ACKERMANN K. Phenolic and anthocyanin assay for use with spectrophotometer[D]. Davis,CA:University of California,2005.

[18] 褚小立,袁洪福,陆婉珍. 近红外分析中光谱预处理及波长选择方法进展与应用[J]. 化学进展,2004,16(4):528-542.
CHU X L,YUAN H F,LU W Z. Progress and application of spectral pretreatment and wavelength selection methods in nir analysis[J]. Progress in Chemistry,2004,16(4):528-542(in Chinese).

[19] GORRY P A. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method[J]. Analytical Chemistry,1990,62(6):570-573.

[20] ZHANG C,GUO C,LIU F,et al. Hyperspectral imaging analysis for ripeness evaluation of strawberry with support vector machine[J]. Journal of Food Engineering,2016,179:11-18.

[21] NI Z,XU L,XIAODUO J,et al. Determination of total iron-reactive phenolics,anthocyanins and tannins in wine grapes of skins and seeds based on near-infrared hyperspectral imaging[J]. Food Chemistry,2017,237:811-817.

[22] CHANG C C,LIN C J. LIBSVM: A library for support vector machines[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(3):27.

作者简介:

成云玲 女,硕士研究生。主要研究方向:高光谱技术在农业信息领域的应用。

杨蜀秦 女,博士,副教授,硕士生导师。主要研究方向:计算机视觉和模式识别。

Selection of measurement variables for hyperspectra of total phenol content in grape seeds based on Monte Carlo frequency method

CHENG Yunling[1,2], YANG Shuqin[1,2,*]

(1. College of Mechanical and Electronic Engineering, Northwest A&F University, Xianyang 712100, China;
2. Key Laboratory of Agricultural Internet of Things, Ministry of Agriculture and Rural Affairs, Xianyang 712100, China)

Abstract: In order to solve the problems of too many variables and high model complexity, it is necessary to effectively reduce the dimension of the data according to the characteristics in establishing the prediction model of total phenol content in grape seeds by using hyperspectral data. In this paper, a Monte Carlo frequency (MCF) method was proposed to select the wavelength of hyperspectral data, and the support vector regression (SVR) prediction model of grape seed total phenols was established. The method uses Monte Carlo sampling to select wavelength subset, then establishes a large number of SVR sub-models, and selects sub-models with smaller root mean square error (RMSE) to count the frequency of each wavelength. Finally, the number of wavelengths is determined by exponential decline function, and the wavelength subset with the highest frequency is selected as the characteristic wavelength. The results show that the prediction performance of the model can be improved by using MCF method at the same time of dimensionality reduction. The number of wavelengths can be reduced from 196 to 9, the range of wavelengths is between 950 and 1400 nm, and the RMSE value can be reduced from 0.42 to 0.37. The prediction accuracy is better than other wavelength selection methods such as SPA. The results show that the proposed MCF method can effectively select characteristic wavelengths in hyperspectral data processing, which provides an effective method for the accurate establishment of prediction model.

Keywords: variable selection; Monte Carlo sampling (MCS); near infrared hyperspectra; grape seeds; total phenol content

Received: 2019-07-08; **Accepted**: 2019-08-03; **Published online**: 2019-08-12 16:53

URL: kns.cnki.net/kcms/detail/11.2625.V.20190812.1553.005.html

Foundation items: National Natural Science Foundation of China (31501228,61876153); the Fundamental Research Funds for the Central Universities (2452019180)

* **Corresponding author**. E-mail: yangshuqin1978@163.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0366

基于美学评判的文本生成图像优化

徐天宇，王智*
（清华大学 计算机科学与技术系，深圳 518055）

摘 要：在对抗生成网络(GAN)这一概念的诞生及发展推动下，文本生成图像的研究取得进展和突破，但大部分的研究内容集中于提高生成图片稳定性和解析度的问题，提高生成结果美观度的研究则很少。而计算机视觉中另一项经典的课题——图像美观度评判的研究也在深度神经网络的推动下提出了一些成果可信度较高的美观度评判模型。本文借助美观度评判模型，对实现文本生成图像目标的GAN模型进行了改造，以期提高其生成图片的美观度指标。首先针对StackGAN++模型，通过选定的美观度评判模型从美学角度评估其生成结果；然后通过借助评判模型构造美学损失的方式对其进行优化。结果使得其生成图像的总体美学分数比原模型提高了3.17%，同时Inception Score提高了2.68%，证明所提方法具有一定效果，但仍存在一定缺陷和提升空间。

关 键 词：文本生成图像；对抗生成网络（GAN）；美观度评判；StackGAN++；美学损失

中图分类号：TP391.41；TP183

文献标识码：A **文章编号**：1001-5965(2019)12-2438-11

基于给定文本生成对应图像是计算机视觉领域一项经典且富有挑战性的任务，顾名思义，即给出一句描述确定内容的文本（可描述某件物体或某个场景环境），通过一定架构的模型生成与文本内容相对应的图像，使其尽可能做到逼近现实，能够迷惑人眼的识别或一些模型的判断。该任务需要在理解文本含义的基础上，根据文本的内容构建出合理的像素分布，形成一幅完整的、真实的图片。因为给出的文本所包含的信息量在通常情况下都远少于其所对应生成的图像（文本通常只对图像中主体部分大致进行了描述，图像则还包含主体所处背景、图像全局特征等额外信息），所以一句给定的文本可能会对应许多符合其描述的图像结果，这是文本生成图像任务的难点所在[1]。在如今生活、制造等多方面迈向智能化发展的时期，文本生成图像这一任务在实际生产生活当中具有非常广的应用价值和潜力，比如应用于平面广告设计领域，可以为广告制作团队生成广告用的配图，从而不必再专门雇用插画制作人员；家具、日用品生产领域，可以通过给出一段产品描述，利用模型批量生成大量的产品概念图，从而给设计者提供了可供选择的样例空间，降低设计环节的工作量。

如今，基于给定文本生成图像任务的实现都是基于从大量的图像数据中学习并模拟其数据的分布形式来生成尽可能接近真实的图像，尤其在对抗生成网络（Generative Adversarial Networks，GAN）[2]的火热发展下，借助其来实现文本生成图像的任务已经成为了主流选择，目前也有许多生成效果优秀的模型被提出。在这一研究方面，

收稿日期：2019-07-09；**录用日期**：2019-08-14；**网络出版时间**：2019-08-27 11:09
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190827.1036.003.html
基金项目：国家自然科学基金（61872215，61531006）
* **通信作者**. E-mail：wangzhi@sz.tsinghua.edu.cn

引用格式：徐天宇，王智. 基于美学评判的文本生成图像优化[J]. 北京航空航天大学学报，2019，45(12)：2438-2448.
XU T Y，WANG Z. Text-to-image synthesis optimization based on aesthetic assessment[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2438-2448 (in Chinese).

研究者所关注的重点是如何能够提高生成模型生成图片的真实性、清晰度、多样性、解析度等问题,这些将直接影响生成模型的质量和性能,并关系到生成模型能否有效投入到实际应用当中。

然而如果考虑到实际应用,图像好看,或者有足够的美观度也是一项重要的需求。比如为平面广告设计配图,对图像的要求不仅是清晰、真实,还应该拥有较高的美观度,从而能够吸引人的眼球,提高广告的关注度。可以说,如果能够实现提高此类模型生成图片的美观度,则在实际应用场景中将会给用户带来更加良好的使用体验,从而提高此类应用的质量。遗憾的是,现在对文本生成图像 GAN 的研究很少关注生成图像的美观质量,现有文献中也并未发现有将美学评判与图像生成相结合的研究,这成为了本文研究的动机。

由此引出另一个问题:如何评判一幅图像的美观度。图像的美观度评判实际上是一项带有主观性质的任务,每个人因不同的阅历、审美观甚至所处环境、情感状态等多方面因素的影响,对同一幅图像有可能会给出完全不同的评价。然而,面对互联网空间与日俱增的图片数量,借助人力对其进行美观度的评价是不切实际的。因此,研究借助计算机进行自动化图像美观度评判成为了计算机视觉领域另一项研究课题,至今也有许多研究者提出了实现原理各异且效果优良的美观度评判模型。借助这些模型,可以对目标图像进行分类或评分,给出尽可能接近符合多数人评价标准的评判结果。

借此,本文致力于研究从美观度的角度对文本生成图像 GAN 的生成结果进行优化的方法。本文的贡献和创新点如下:

1)从实际应用的角度出发,将生成结果美观度加入评价文本生成图像 GAN 模型生成结果的评价指标,以目前受到较高认可度的文本生成图像 GAN 模型——StackGAN++[3] 为基础,从美观度的角度对其生成结果进行评估,以观察其生成结果的美观度质量。

2)将美观度评判模型融入该 GAN 的生成模型当中,通过增添美学损失的方式改造生成模型,从而在模型训练过程中加入美学控制因素,引导模型生成美观度更高的结果。本文提出的改进方法使得模型生成图像的总体美学质量(以 IS(Inception Score)为评价指标[4])提高了 2.68%,其生成图像结果整体的美观度指标提高了 3.17%。

1　相关工作

1.1　美观度评判模型

随着网络空间中图片数量的急速增长,在图片检索领域为了能够更好地为用户甄选返回图像的质量、给用户返回更高质量的搜索结果,对图片按美学质量进行分类的需求逐渐增加。图片所附带的数据标签(如喜欢该图的人的数量、图片内容等)可以作为美观度评价的一类较为有效的标准,但大部分的图片并不存在类似这样的标签,虽然如今有许多研究已能够做到给图片准确高效地进行标签标注[5],然而即使每幅图片均被标注了足够用以进行评判的标签,图片庞大的数量又使得人工评判工作量巨大,因此需要能够对图片进行美观度评判的模型,由计算机来完成这一任务。

受到心理学、神经科学等领域中对人类美学感知的研究成果启发,计算机视觉领域的研究者们通过模拟、复现人类处理接收到的图像视觉信息的过程,设计实现了一系列自动评判图片美学质量的模型[6]。图像美观度评判模型一般遵循一个固定的流程:首先对输入图像进行特征提取,然后借助提取的特征,利用训练好的分类或回归算法获得相应的结果。

特征提取则是其中非常重要的一环,因为特征信息是对图像美学质量的概括,其决定了美观度评判模型的精确度。选取得当的特征既能提高模型评判的精确度,又能减少不必要的计算量,因为不同特征对于图像美学质量的贡献度是不同的[7]。早期的研究中,研究者们通常选择以绘画、摄影所用的美学规则理论和人的直观感受为依据,自主设计所要提取的特征,比如清晰度、色调、三分规则等。这类方法的好处是直观、易于理解,但缺点在于所设计的特征通常不能很全面地描述图像美学信息,而且设计特征对于研究者的工程能力和相关领域知识了解程度都有较高的要求。而随着深度学习领域的不断发展,将卷积神经网络(Convolitional Neural Networks,CNN)应用于图像处理这一方式展现出了卓越的效果。借助 CNN 能够从大量的图像数据中学习到有力的图像特征表示,其所包含的信息量远超人工特征设计所设定的特征[8],从而使得 CNN 处理图像的方式在图像处理领域得到广泛应用,并逐渐成为主流选择的方法。深度学习方法应用于图像美观度评判的特征提取环节,主要有 2 种方式:第 1 种是借助已有的深度学习图像处理模型,利用其中间层特征作为评判依据,采用传统的分类或回归方

法进行美观度评判;第 2 种是对已有的模型进行改造,使得其能够从图像数据中学习到新的隐藏的美学特征,并借此对图像的美观度作出评判。

本文采用的是 Kong 等[9]设计的美观度评判模型。该模型随 AADB(Aesthetics and Attributes Database)数据集一同提出,其基于 AlexNet[8]改造得来,通过提取图片的内容特征以及自定义的属性标签特征来帮助判断图像的美观度。此外,该模型吸收了 Siamese 网络[10]的结构,实现了接收两幅一组的图像作为输入并给出它们之间相对评分的功能,同时提出了 2 种对图像进行成对采样的训练方式来辅助增加结果的精确度。实验结果表明该模型在 AVA(Aesthetic Visual Analysis)数据集上的判别准确率达到77.33%,超过了当时已有的许多模型的表现。作者并未对该模型进行命名,为方便说明,下文中统一用“AADB 模型”对其进行代指。

1.2 文本生成图像 GAN

GAN 的提出是机器学习领域一项重大的突破,其为生成模型的训练提供了一种对抗训练的思路。相比于传统的生成模型如变分自编码器、玻尔兹曼机,GAN 优势有:其训练只需借助反向传播而不需要马尔可夫链、能够产生全新的样本以及更加真实清晰的结果、简化任务设计思路等,因此,其成为了现今机器学习领域十分火热的研究课题。

GAN 的结构一般可分为两部分:生成器部分,负责接收一段随机噪声作为输入来生成一定的结果;判别器部分,负责接收训练数据或生成器生成的数据作为输入,判断输入是来自哪一方。生成器的最终目标是生成能够彻底欺骗判别器的数据,即判别器无法区分输入数据来自真实数据分布还是生成器拟合的数据分布;而判别器的最终目标是有效区分其输入来源,识别出来自生成器的输入。GAN 的训练正是基于这种博弈的过程,令生成器和判别器二者之间进行对抗,交替更新参数,当模型最终达到纳什均衡时,生成器即学习到了训练数据的数据分布,产生相应的结果。

虽然 GAN 拥有良好的表现力和极大的发展潜力,但其本身还存在一些缺点,比如训练困难、无监督使得生成结果缺少限制、模式崩溃、梯度消失等问题。后续许多研究者对 GAN 从结构[11]、训练方法[12]或实现方法[13]上进行了改进,逐渐提高了 GAN 训练的稳定性和生成效果。此外,CGAN(Conditional GAN)[14]将条件信息与生成器和判别器的原始输入拼接形成新的输入,用以限制 GAN 生成和判别的表现,使得 GAN 生成结果的稳定性得到提高。

利用 GAN 来实现文本生成图像任务也是基于 CGAN 的思想,以文本-图像组合为训练数据,文本作为输入数据的一部分,在生成器中与随机噪声拼接作为生成器的整体输入,在判别器中则用于形成不同的判断组合——真实图片与对应文本、真实图片与不匹配文本、生成器生成图片与任意文本并进行鉴别。文本数据通常会借助其他编码模型将纯文字信息转化为一定维数的文本嵌入向量,用以投入模型的训练计算当中。最先利用 GAN 实现文本生成图像任务的是 Reed 等[15]提出的 GAN-INT-CLS 模型,其吸收了 CGAN 和 DCGAN(Deep Convolutional GAN)[11]的思想,同时提出改进判别器接收的文本-图像组合输入(新增真实图像与不匹配文本的组合)以及通过插值的方式创造新的文本编码向量两种方法来提高生成结果的质量和丰富度,生成了 64×64 大小的图像。随后该领域的一项重要突破是 Zhang 等[16]提出的 StackGAN 模型,该模型通过使用 2 个生成器的方式生成图像,首次实现了只借助给定文本的条件下生成 256×256 大小的图像。该模型中,第 1 个生成器接收随机噪声与文本向量的拼接来生成 64×64 大小的中间结果,第 2 个生成器则使用该中间结果与文本向量作为输入,这种方式可以实现利用文本信息对中间结果进行修正和细节补充,来获得质量更高的 256×256 大小图像的结果。

在 StackGAN 的理论基础上,Zhang 等[3]提出了 StackGAN++模型。该模型使用 3 个生成器-判别器组以类似树状的方式连接,其中 3 个生成器分别对应生成 64×64、128×128、256×256 大小的图像,第 1 个生成器以文本向量和随机噪声的拼接为输出,之后每一个生成器接收前一个生成器生成的图像结果与文本向量作为输入,生成下一阶段的图像结果;每一个判别器接收对应阶段的生成器的输出与文本向量进行判别,计算条件生成损失。此外,Zhang 等[3]引入了无条件生成损失,即计算在不使用文本信息的情况下生成图片的损失,与条件生成损失相结合,引导模型的训练,最终进一步提高了生成图片的质量。本文即选用了该模型进行基于美学评判的优化改进研究。

此后文本生成图像 GAN 的研究多在类似 StackGAN++的多阶段生成模式基础上,通过加入各种辅助信息来帮助生成器生成更好的结果,

如 AttnGAN(Attentional GAN)[17]引入了注意力机制,分析对比生成图像与对应文本之间的特征相似度,并利用对比结果辅助生成器的训练;Cha 等[18]则通过引入感知损失的方式,从图像特征层面进行对比来辅助生成器更好地学习到训练数据的分布。

2　StackGAN++的美学质量分析

在提出基于美学评估的对 StackGAN++模型的优化方法之前,需要了解该模型目前生成结果的美学质量如何。本节将利用 AADB 模型对其进行初步测量。

本节实验使用的 StackGAN++模型是基于 Caltech-UCSD Birds 200 鸟类图像数据库 2011 版训练的鸟类图像生成模型,其测试数据集中包含 2933 张图像,每张图像对应 10 条文本说明,其中文本数据需经过 char-CNN-RNN 模型编码。Zhang 等[3]给出了其模型源码的 github 地址(https://github.com/hanzhanggit/StackGAN-v2)。

本文实验运行于 Ubuntu 16.04 操作系统,使用 GeForce GTX 1080 Ti 显卡进行训练。软件环境方面,本实验利用 Adaconda2 搭建 python2.7 虚拟环境,并需要安装 Pytorch1.0 以及 caffe1.0(分别对应 StackGAN++以及 AADB 模型运行所需)。

2.1　测试数据集生成结果的美观度分布

首先针对测试数据集所产生的样本进行美观度评判,观察其分布状况。理论情况下,训练数据集中包含了 29 330 条语句对应的嵌入向量,经由生成模型后获得 29 330 张图像结果,实际运行中由于 StackGAN++模型所采用的批处理训练策略,最终生成图像数量为 29 280 张,但从整体数量的规模来看并不影响对于其整体美观度评价的判断。利用 AADB 模型获得生成图像的美学分数,其分布如图 1 所示。

由 AADB 模型计算得出的美学分数集中于[0,1]区间,在特殊情况下会超过 1。为了便于标注美学分数的分布区间,在绘制区间分布柱状图时,将由 AADB 模型获取的美学分数(超过 1 的截断至 0.999 9)乘以 10,这种表示方法也符合实际生活中人工评判时的常用取值范围选择;在展示降序分布时则直接采用模型输出的结果范围来标注分数坐标轴。图 1(a)表明,原始 StackGAN++在测试数据集上生成图像的美学分数集中在 5~8 的区间段内,占总体的 78.6%,其中 6~7 区间段内的图像数量最多,占整体结果数量的33.9%。而图 1(b)表明,在 5~8 区间段内,图像的美学分数变化呈现出均匀平缓的变化趋势,并没有出现在某一节点大幅变动的情况。

29 280 张生成结果的平均美学分数为 0.628 28。根据 AADB 模型作者给出的评判标准,一张图片的分数超过 0.6 则可以认为是一张好图片,低于 0.4 则认为是一张差图片,在两者之间认为是一张一般性质图片,而本文出于后续实验样本划分的考虑,将好图片的下限标准提高至 0.65,差图片的上限标准提高至 0.5。由此来看,模型的平均结果处于一般质量的区间,说明原模型的整体生成结果从美观度的角度来讲仍然存在可以提升的空间。本文从全部生成结果中选择美学分数最高以及最低的图片各 10 张的结果,交由真人进行主观评判,其结果均与美学分数表现出对应关系,即认为最高分数的 10 张图片拥有较高的美观度,而最低分数的 10 张图片则评价一般或交叉表明 AADB 模型给出的美学分数对图像美观度的评价能较好地符合人的直观感受。

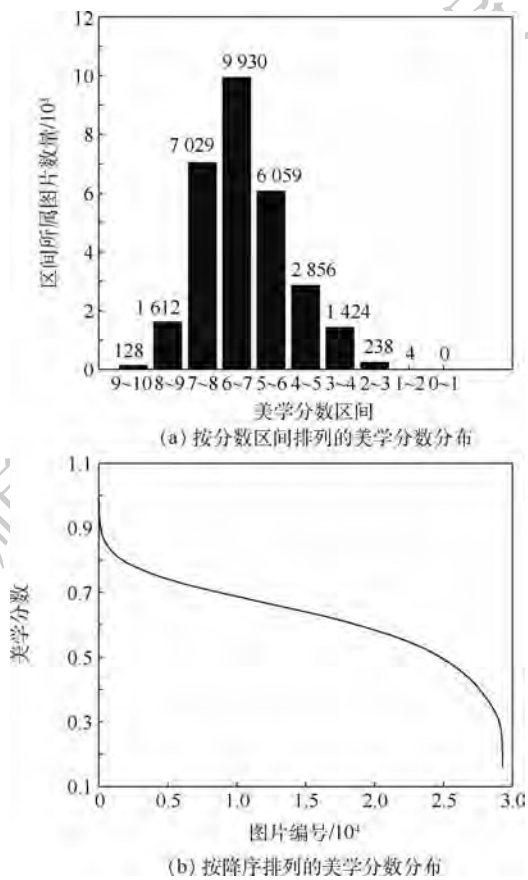


图 1　StackGAN++在测试数据集上的美学分数分布

Fig. 1　Distributions of aesthetic score of images generated by StackGAN++ using test dataset

2.2　固定文本批量生成图像结果的美观度分布

如果想要达成提高生成模型美观度质量的目

标,一个简单的想法是,可以对同一条语句,一次性批量生成大量的图片,按美观度模型给出的分数进行降序排序,从中选出分数最高图像作为输出结果,或以分数排序最靠前(分数最高)的一定数量的图像作为输出结果,再交由模型使用者自行判断选择最终的结果。这种方法虽然存在严重的效率问题,但易于实现且非常直观。其中的问题在于确定生成图像的数量,因为随着生成图像数量的增加,其多样性也会随之增加。也更容易出现更多美观度高的图像结果,但进行美观度评判以及排序选择的时间消耗也会随之增加,因此若选择此种做法作为优化方法,需要在生成结果质量以及模型运行效率之间寻求一个平衡点。

基于以上考虑,除对测试集整体进行美观度评判以外,还从中选择一批(实验设定为 24)数量的文本输入数据,针对每一条文本数据生成不同数量的图片来观察其美学分数的分布。选择 100、200、350、500、750、1 000 共 6 种生成数量,针对选定的文本数据生成对应数量的图像,利用 AADB 模型计算生成结果的美学分数。图 2 展示了其中一条文本的结果。结果表明,美学分数在

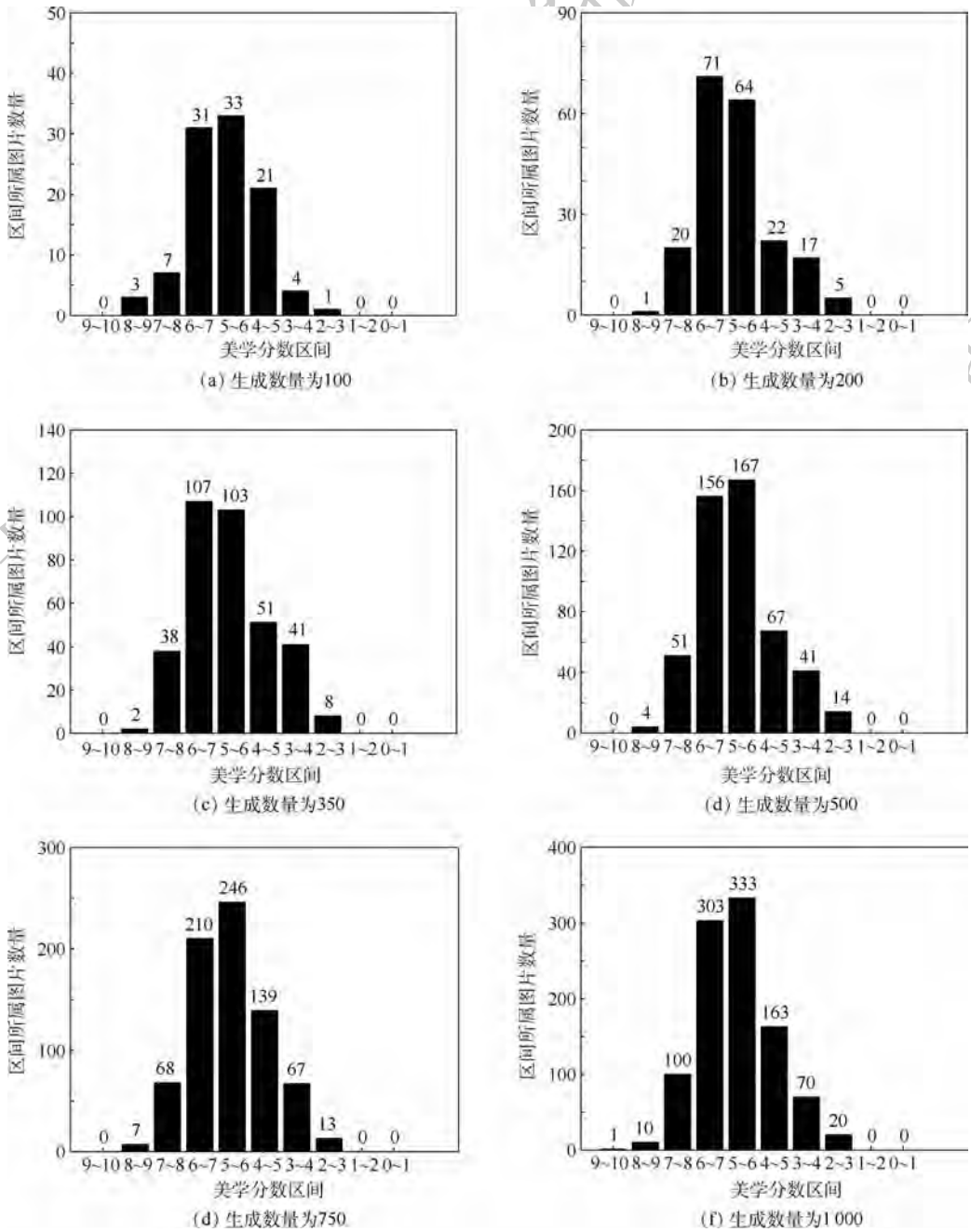


图 2 原模型输入同一文本在不同生成数量情况下的美学分数分布

Fig. 2 Disrtributions of aesthetic score of images generated by the same text in original model with different generating quantities

各个区间的分布状况是相近的,基本不受一次性生成数量的影响。6 组结果都表现出生成图像的美学分数集中于 5 ~ 7 的区间内的分布状况,且随着生成数量的增加,高分图像的出现频率也越来越高。表 1 展示了 6 组分布结果中最高分数图像的分数与分数前 10 高图像的平均分数,表明了一次性生成数量越多,即使是处于高分分段的图像其整体的质量也会得到提高,也验证了本节第一段所述的情况。但面对最高分数的情况,因生成模型会以随机噪声作为输入来生成图像,这导致了其对生成结果的不可控性,所以生成结果会出现一定的扰动,使得最高分图像的分数与生成数量之间并不存在确定的正相关关联性。

表 1　不同生成数量情况下最高分数与前 10 平均分数

Table 1　Highest aesthetic score and average of top 10 aesthetic score when generating different quantities of images

图像生成数量	最高分数	前 10 平均分数
100	0.892 8	0.769 8
200	0.804 2	0.775 3
350	0.845 1	0.788 9
500	0.843 9	0.797 8
750	0.858 4	0.815 5
1 000	0.954 8	0.859 2

3　基于美学评判的图像生成优化

原始的 StackGAN ++ 模型采用了 3 组生成器-判别器组合,以类似树状的方式进行连接,每一个生成器生成不同尺寸的图像,并作为下一个生成器输入数据的一部分。其中每一个生成器的损失 $L_{G_i}(i=1,2,3)$ 计算式为

$$L_{G_i} = -E_{s_i\sim p_{G_i}}[\ln D_i(\boldsymbol{s}_i)] - E_{s_i\sim p_{G_i}}[\ln D_i(\boldsymbol{s}_i,\boldsymbol{c})] \tag{1}$$

式中:p_{G_i} 为生成器 G_i 学习到的数据分布;$\boldsymbol{s}_i$ 为生成器 G_i 生成的结果;$\boldsymbol{c}$ 为文本向量;D_i 为与生成器 G_i 对应的判别器,其接收单个输入 $\boldsymbol{s}_i$ 或双输入 $\boldsymbol{s}_i$ 和 $\boldsymbol{c}$,输出相应的判别结果;$E[\cdot]$ 表示期望函数。

该损失计算方式由两部分组成,前一部分计算生成器不利用文本向量生成图像的损失,即无条件损失,该部分用以监督生成器生成更加真实的、使判别器认为来自于真实数据分布的数据;后一部分计算生成器利用了文本向量生成图像的损失,即条件损失,该部分用来监督生成器生成符合输入文本描述的图像,即保证文本与图像之间的一致性。在 StackGAN ++ 的理论描述中,Zhang 等[3]认为每一个生成器生成的图像虽然大小不同,但都是基于同一条文本生成的,所以它们彼此之间应该保持相似的色彩和基本结构,并提出了色彩一致性损失用来保证 3 个生成器生成图像之间拥有较高的色彩一致性。但经过实验作者发现在基于文本生成的模式下,色彩一致性所起到的作用十分微弱,因为其对生成结果的约束力要远远小于文本-图像一致性的约束,即式(1)中的 $E_{s_i\sim p_{G_i}}[\ln D_i(\boldsymbol{s}_i,\boldsymbol{c})]$。式(2)为生成器的总体损失(下文称为对抗损失)计算公式,用于训练过程中的梯度计算。

$$L_G = \sum_{i=1}^{3} L_{G_i} \tag{2}$$

受到 Johnson 等[19]提出的感知损失的启发,本文将 AADB 模型与 StackGAN ++ 的生成器结合,用于在生成模型训练过程中提供辅助训练信息,达成从美学角度来优化生成模型的目的。具体地,在式(2)的基础上,加入一项新定义的损失——美学损失 L_{aes},其计算式为

$$L_{\text{aes}} = E_{s_3\sim p_{G_3}}[1 - \text{Aes}(\boldsymbol{s}_3)]^2 \tag{3}$$

式中:Aes 函数表示使用 AADB 模型计算生成结果 $\boldsymbol{s}_3$ 的美学分数。2.1 节中,AADB 模型计算得出的美学分数存在超过 1 的情况,因此在计算美学损失时,会对模型返回的美学分数进行判断,如果其超过了 1,则将其截断至 0.999 9。该损失实际计算了最后一阶段生成器 G_3 生成结果的美学分数与 1 之间的欧几里得距离,最小化该损失即最小化生成结果美学分数与 1 之间的差距,代表了生成结果美学质量的提升。最后,加入了美学损失后新的生成器损失计算公式为

$$L'_G = L_G + \beta L_{\text{aes}} \tag{4}$$

式中:β 为美学损失的权重,用来控制其在总体损失中所占的比例,β 越大则美学损失所占的比例越大。β 为 0 时,模型即还原为 StackGAN ++。

由于美学损失的作用是引导生成器生成美观度更高的图像,而对抗损失则是控制整个训练过程以及生成结果的关键,保证了生成器能够生成符合文本描述的真实图像,这是文本生成图像模型最基本的目的,因此 β 值的选择应当在保证在训练过程中美学损失起到的调控作用不会压过对抗损失的前提下对生成结果的美观度产生影响。

4　实验与性能评估

选取不同的美学损失权重 β 进行训练,以 IS 作为训练获得模型的质量的衡量指标,在保证 IS

与原模型相比不降低的前提下,观察其生成结果的美观度分布情况。IS 是借助 Inception Model[20] 计算得出的用来衡量 GAN 图像生成效果的最常用指标之一,通常情况下其数值越大代表 GAN 生成的图像具有更高的多样性和真实性,进而代表生成图像的总体质量更好。在文本生成图像 GAN 领域,IS 被广泛用来进行不同 GAN 之间的效果对比。

本节所使用的环境与第 2 节对 StackGAN++ 本身进行美学质量分析的实验环境相同,故此处不再赘述。模型训练过程采用批训练策略,每个批包含 24 条文本嵌入向量,每一个时期(epoch)中包含 368 个批的训练过程,下文将一个批完成一次训练的过程称为一步(step)。训练过程包含 600 个时期,并于每 2 000 步的时间节点保存一次模型参数,以便于训练完成后根据保存时模型的表现选取效果最好的模型。本文提出的优化方法的实现流程如图 3 所示。

本文选取 $\beta=45,0,0.000\,1$,分别进行了训练。选择 45 是因为,观察 StackGAN++ 训练时生成器的对抗损失发现对抗损失与美学损失的比值在 50 左右。因此,当 $\beta=45$ 时,对抗损失与经权重放大的美学损失在数值上比较接近;1 与 0.000 1 是基于经验的选择,取 $\beta=1$ 时美学损失与对抗损失平权,而 $\beta=0.000\,1$ 则是参考了 Cha 等[18] 提出的感知损失的权重选取。训练完成后,对应每个 β 取值各形成了一组于不同时间节点保存的模型,分别从中选取 IS 分数最高的模型作为对应取值下的结果模型。在对选定的模型进行美学质量评判之前,需要先考察它们所生成的图片的总体质量,以确保在引入了美学损失后没有出现模型生成图像质量下降的情况。表 2 展示了 3 种取值对应模型与原模型的 IS 数据,其中 β 为 0 即代表未引入美学损失的原始 StackGAN++ 模型。

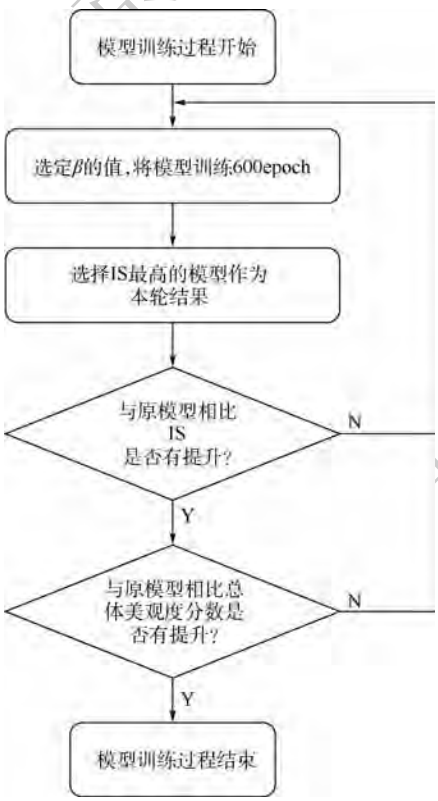


图 3 基于美学评判的文本生成图像优化方法实现流程

Fig. 3 Procedure of text-to-image synthesis optimization based on aesthetic assessment

通过对比,当 $\beta=0.000\,1$ 时,模型在 IS 上取得最高的数值,并且超过了原始模型的 IS,表明美学损失的引入还起到了提高模型生成效果的正面效应。这是可以理解的,因为当生成器生成了一幅效果很差的图像,例如模糊不清或主体扭曲变形,此时美观度评判模型将会给出较低的分数,使得美学损失增大并导致生成器总损失增大。此外,当 $\beta=45$ 时,模型的 IS 分数降低,表现为生成器生成图像的质量有所下降。对 $\beta=45$ 时获得的模型所生成的图像进行人工评判的结果也反映出这时生成图像出现了更多的模糊、失真等不良结果。因此,$\beta=45$ 的情况已无继续讨论的价值,此后美学层面的实验和数据统计也不再考虑此种情况。当 $\beta=1$ 时,模型的 IS 与原模型相比十分接近,还需通过美学分数的分布对比来确定在此情况下美学损失是否起到了优化的作用。

为了验证美学损失是否对生成模型结果的美学质量起到了优化作用,接下来计算了使用 $\beta=1,0.000\,1$ 这 2 种情况的模型在测试数据集上生成的 29 280 张图像的美学分数分布情况;同时针对一个批的 24 条文本嵌入向量,每条文本生成 1 000 张图像,计算其美学分数的分布,数据结果如图 4 所示(这里选出一条文本生成的 1 000 张图像的美学分数分布进行展示)。表 3 展示了 2 种 β 取值下模型在测试数据集上的生成结果的美学分数,同时一并列出了原模型在测试数据集上生成结果的美学分数作为对比。从表中可知,当 $\beta=0.0001$ 时,由测试数据集生成的图像其平

表 2 不同 β 取值对应模型的 IS

Table 2 IS of models using different β

β	IS
0	4.10±0.06
45	4.05±0.03
1	4.13±0.05
0.000 1	4.21±0.03

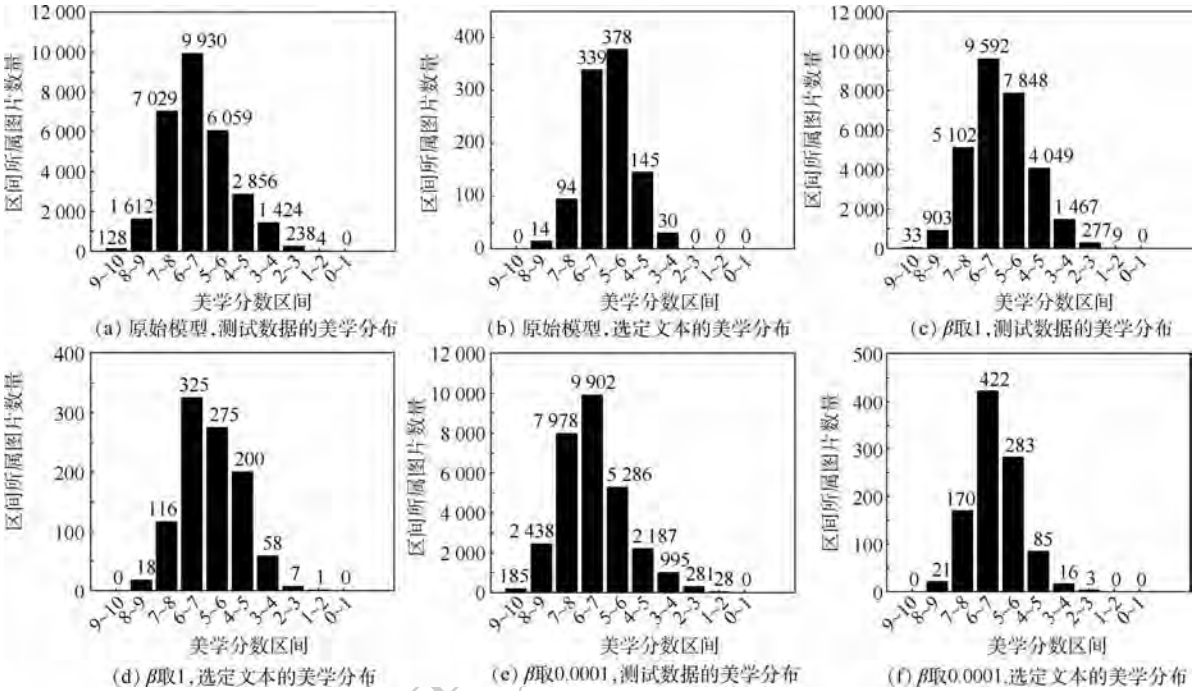


图 4　不同 β 取值情况下测试数据生成结果与选定一条文本的生成结果美学分数分布

Fig. 4　Distributions of aesthetic scores of images generated by models with different β using texts from test dataset and chosen text

表 3　不同 β 取值对应模型的美学分数对比

Table 3　Comparison of aesthetic scores of models using different β

β	测试集平均分数
0	0.628 3
1	0.600 6
0.000 1	0.648 2

均美学分数与原模型相比提高了 3.17%；表 4 给出了原模型与 β = 0.000 1 优化模型分别生成的 24 组针对同一条文本的 1 000 幅图像平均美学分数对比情况，也可以发现大部分文本生成结果的美学分数与原模型的生成结果相比有所提高。同时由图 4 所示的美学分数分布情况也能看出，此时高分段图像的数量增加，较低分段图像的数量减少，表明美学损失起到了调控生成结果美观度的作用。图 5 展示了原模型与 β = 0.000 1 的优化模型使用 4 条文本对应生成的1 000 张图像中等距抽取 10 张图像的结果（每个分图第 1 行为原模型，第 2 行为优化模型，每个模型对每条文本均生成1 000 张图像），每行图像从左到右按美学分数从高到低的顺序排列，从中可以直观感受到，经过美学优化的生成模型所生成的图像结果在色彩对比度、整体色调、背景虚化简单化等方面均有一定优势，反映了其美观评价相比原模型有所提升。

表 4　原模型与优化模型使用 24 条文本生成 1 000 张图像的平均美学分数对比

Table 4　Comparison of average aesthetic score of 1 000 images generated by original models and optimized models using 24 different texts

编号	平均美学分数		对比提升量
	原模型	β = 0.000 1	
1	0.592 9	0.619 2	+0.026 3
2	0.586 6	0.618 0	+0.031 4
3	0.567 1	0.617 0	+0.049 9
4	0.574 1	0.543 3	−0.030 8
5	0.581 9	0.551 5	−0.030 4
6	0.551 0	0.518 8	−0.032 2
7	0.602 4	0.622 1	+0.019 7
8	0.568 8	0.546 7	−0.022 1
9	0.598 7	0.627 9	+0.029 2
10	0.616 6	0.589 0	−0.027 6
11	0.600 8	0.613 8	+0.013 0
12	0.551 1	0.622 7	+0.071 6
13	0.700 7	0.691 3	−0.009 4
14	0.586 2	0.659 6	+0.073 4
15	0.586 8	0.623 2	+0.036 4
16	0.595 1	0.622 7	+0.027 6
17	0.621 7	0.634 5	+0.012 8
18	0.555 0	0.579 2	+0.024 2
19	0.585 9	0.597 4	+0.011 5
20	0.568 9	0.567 3	−0.001 6
21	0.598 8	0.640 5	+0.041 7
22	0.579 4	0.547 7	−0.031 7
23	0.558 7	0.593 0	+0.034 3
24	0.589 0	0.541 9	−0.047 1

图 5 从原模型与优化模型对 4 条文本各生成的 1 000 幅图像中等距抽取图像对比

Fig. 5 Comparison of systematic sampling results of images generated by original model and optimized model using 4 chosen texts (1 000 images for each text and each model)

5 结 论

本文提出了一种基于美学评判的文本生成图像 GAN 的优化方法,利用美观度评判模型获得生成器生成图像的美学分数,计算该生成图像的美学损失,与模型本身的对抗损失以适当的权重关系相结合,作为该生成器新的损失并重新训练模型,最后对获得的新模型生成的图像进行了美学质量的统计与和原模型的对比。实验所得结论如下:

1) 经过本文方法获得的生成模型,其生成结果的美观度与原模型相比得到了提升,同时 IS 分数也有所提高,表明美学损失能够起到提高生成模型质量的作用。

2) 该方法同时也存在一定的局限性:首先,利用本文方法选取的美学损失权重只适用于本文实验所用的鸟类数据集,而如果需要在其他数据集上应用本文方法,除了原始模型需要更换以外,权重的选取也需要全部从头开始,这也与文本生成图像类 GAN 和美学模型本身的性质有关;其次,权重选取尚未找到一个公式化的方法,都是借助经验来试探性的选择。

未来的工作将分为 2 个部分,一是基于文中的实验继续调整美学损失的权重,寻找最佳的参数以使得生成结果能进一步提高;二是寻找其他美学损失的计算方式,以使得美学因素能够取得更好的训练调控效果。

致谢　感谢耀萃基金数据智能创新联合实验室(InfleXion Lab)提供的帮助!

参考文献(References)

[1] BODNAR C. Text to image synthesis using generative adversarial networks[EB/OL]. (2018-05-02)[2019-07-08]. https://arxiv.org/abs/1805.00676.

[2] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2014: 2672-2680.

[3] ZHANG H, XU T, LI H, et al. Stackgan++: Realistic image synthesis with stacked generative adversarial networks[EB/OL]. (2018-06-28)[2019-07-08]. https://arxiv.org/abs/1710.10916.

[4] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2016: 2234-2242.

[5] LI Z, TANG J, MEI T. Deep collaborative embedding for social image understanding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 41(9): 2070-2083.

[6] DENG Y, LOY C C, TANG X. Image aesthetic assessment: An experimental survey[J]. IEEE Signal Processing Magazine, 2017, 34(4): 80-106.

[7] DATTA R, JOSHI D, LI J, et al. Studying aesthetics in photographic images using a computational approach[C]// European Conference on Computer Vision. Berlin: Springer, 2006: 288-301.

[8] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2012: 1097-1105.

[9] KONG S, SHEN X, LIN Z, et al. Photo aesthetics ranking network with attributes and content adaptation[C]// European Conference on Computer Vision. Berlin: Springer, 2016: 662-679.

[10] CHOPRA S, HADSELL R, LECUN Y. Learning a similarity metric discriminatively, with application to face verification[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2005: 539-546.

[11] RADFORD A, METZ L, CHINTALA S. Unsupervised representation learning with deep convolutional generative adversarial networks[EB/OL]. (2016-01-07)[2019-07-08]. https://arxiv.org/abs/1511.06434.

[12] SALIMANS T, GOODFELLOW I, ZAREMBA W, et al. Improved techniques for training gans[C]// Advances in Neural Information Processing Systems. Cambridge: MIT Press, 2016: 2234-2242.

[13] ARJOVSKY M, CHINTALA S, BOTTOU L. Wasserstein gan[EB/OL]. (2017-12-06)[2019-07-08]. https://arxiv.org/abs/1701.07875.

[14] MIRZA M, OSINDERO S. Conditional generative adversarial nets[EB/OL]. (2014-11-06)[2019-07-08]. https://arxiv.org/abs/1411.1784.

[15] REED S, AKATA Z, YAN X, et al. Generative adversarial text to image synthesis[EB/OL]. (2016-06-05)[2019-07-08]. https://arxiv.org/abs/1605.05396.

[16] ZHANG H, XU T, LI H, et al. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 5907-5915.

[17] XU T, ZHANG P, HUANG Q, et al. Attngan: Fine-grained text to image generation with attentional generative adversarial networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 1316-1324.

[18] CHA M, GWON Y, KUNG H T. Adversarial nets with perceptual losses for text-to-image synthesis[C]// 2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP). Piscataway, NJ: IEEE Press, 2017: 1-6.

[19] JOHNSON J, ALAHI A, LI F. Perceptual losses for real-time style transfer and super-resolution[C]// European Conference on Computer Vision. Berlin: Springer, 2016: 694-711.

[20] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 2818-2826.

作者简介:

徐天宇　男,硕士研究生。主要研究方向:数据挖掘与深度学习。

王智　男,博士,副教授。主要研究方向:多媒体网络。

Text-to-image synthesis optimization based on aesthetic assessment

XU Tianyu, WANG Zhi*

(Department of Computer Science and Technology, Tsinghua University, Shenzhen 518055, China)

Abstract: Due to the development of generative adversarial network (GAN), much progress has been achieved in the research of text-to-image synthesis. However, most of the researches focus on improving the stability and resolution of generated images rather than aesthetic quality. On the other hand, image aesthetic assessment research is also a classic task in computer vision field, and currently there exists several state-of-the-art models on image aesthetic assessment. In this work, we propose to improve the aesthetic quality of images generated by text-to-image GAN by incorporating an image aesthetic assessment model into a conditional GAN. We choose StackGAN++, a state-of-the-art text-to-image synthesis model, assess the aesthetic quality of images generated by it with a chosen aesthetic assessment model, then define a new loss function: aesthetic loss, and use it to improve StackGAN++. Compared with the original model, the total aesthetic score of generated images is improved by 3.17% and the inception score is improved by 2.68%, indicating that the proposed optimization is effective but still has several weaknesses that can be improved in future work.

Keywords: text-to-image synthesis; generative adversarial networks (GAN); aesthetic assessment; StackGAN++; aesthetic loss

Received: 2019-07-09; **Accepted**: 2019-08-14; **Published online**: 2019-08-27 11:09

URL: kns.cnki.net/kcms/detail/11.2625.V.20190827.1036.003.html

Foundation items: National Natural Science Foundation of China (61872215, 61531006)

* **Corresponding author**. E-mail: wangzhi@sz.tsinghua.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0367

基于边缘和结构的无参考屏幕内容图像质量评估

魏乐松，陈俊豪，牛玉贞*
（福州大学 数学与计算机科学学院，福州 350100）

摘 要：屏幕内容图像(SCI)是一种与传统自然图像不同的图像，具有更多的文本、图形以及特殊的布局。考虑文本、图形、图像和布局对屏幕内容图像质量的影响，提出了针对屏幕内容图像的基于边缘和结构的无参考质量评估(BES)算法。文本、图形和图像具有大量边缘，并且人类视觉系统对边缘高度敏感，因此BES算法首先使用Gabor滤波器的虚部提取边缘并计算每张屏幕内容图像的边缘特征。其次，提取一个结构特征来表示屏幕内容图像的布局。具体而言，利用Scharr滤波器计算得到一个局部二值模式(LBP)图，接着利用LBP图计算得到结构特征。最后，应用随机森林回归算法将边缘和结构特征映射为主观分数。实验结果表明，在数据库SIQAD和SCID上，所提出BES算法性能的皮尔森线性相关系数(PLCC)相对于对比算法中最先进的无参考算法，分别提高了2.63%和11.22%，甚至高于一些全参考算法。

关 键 词：屏幕内容图像；无参考质量评估；随机森林回归；边缘特征；结构特征
中图分类号：TP399
文献标识码：A **文章编号**：1001-5965(2019)12-2449-07

随着互联网技术的快速发展，屏幕内容图像(Screen Content Image，SCI)被广泛应用于现代多媒体应用，如无线显示、远程教育、屏幕共享、实时通信等。由于技术和设备的缺陷，在压缩、传输、获取等过程中不可避免地会引入各种失真，影响用户的体验。例如，当利用智能手机中的相机创建屏幕内容图像时，由于相机运动和不同的环境，可能会产生噪声和模糊失真。对于通过因特网传输屏幕内容图像，为了进行有效传输，可能由于图像编码而产生压缩失真。因此，迫切需要屏幕内容图像的视觉质量评估算法，用来优化多媒体应用系统。

客观质量评估算法根据参考原始图像的程度，评估算法可以分为3类：全参考(Full Reference，FR)、半参考(Reduce Reference，RR)和无参考(No Reference，NR)。其中，全参考图像质量评估算法使用了原始图像作为失真图像的参照图像；半参考图像质量评估算法中仅使用了部分参照图像的信息；无参考图像质量评估算法没有使用任何的参照图像中的信息作为先验数据。然而，在许多实际应用中，特别是在大数据应用领域，获取全部或部分的参考图像信息是非常昂贵的，甚至是不可能的。而无参考图像质量评估不依靠任何参考图像的信息，比全参考图像质量评估和半参考图像质量评估具有更大的实际应用前景。

在过去的几十年中，图像质量评估领域得到很大的发展，已经有针对视觉内容设计的各种图像质量评估算法。大多数传统的评估指标都是全参考算法，诸如峰值信噪比算法(Peak Signal-to-Noise Ratio，PSNR)和均方误差(Mean Square Error，MSE)，它们通过简单地计算参考和失真图像

收稿日期：2019-07-09；**录用日期**：2019-08-03；**网络出版时间**：2019-08-12 16:20
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190812.1333.001.html
基金项目：国家自然科学基金（61672158）；福建省自然科学基金（2019J2006）
*通信作者. E-mail：yuzhenniu@gmail.com

引用格式：魏乐松，陈俊豪，牛玉贞. 基于边缘和结构的无参考屏幕内容图像质量评估[J]. 北京航空航天大学学报，2019，45(12)：2449-2455. WEI L S，CHEN J H，NIU Y Z. Blind quality assessment for screen content images based on edge and structure[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2449-2455 (in Chinese).

之间的像素差异来预测图像的视觉质量。由于其简单有效,这些方法在工业界和学术界得到广泛的应用。其没有考虑人类视觉系统的属性,因此,无法获得与人类感知较一致的质量预测结果。为了解决这个问题,许多研究人员根据人类视觉系统的特点,提出了各种不同的评估算法,如结构相似性(Structural Similarity,SSIM)算法[1]、梯度幅度相似性偏差(Gradient Magnitude Similarity Deviation,GMSD)算法[2]、自然图像质量评估(Natural Image Quality Evaluator,NIQE)算法[3]、以及盲/无参考图像空间质量评估(Blind/Referenceless Image Spatial Quality Evaluator,BRISQUE)算法[4]。

但是,上述的评估算法是专为自然图像设计的,对屏幕内容图像效果不好。通常,屏幕内容图像由计算机生成,由文本、图形和图像组成,具有特殊的布局,导致这 2 种图像在统计特征上存在明显差异。对于屏幕内容图像的研究,Yang 等[5]进行了主观实验并构建了一个 SIQAD 数据库,基于该数据库,提出了各种针对屏幕内容图像的评估算法,且提出了 SPQA(SCI Perceptual Quality Assessment)算法,该算法通过分析图像区域和文本区域的质量感知特征来考虑图像整体质量。Wang 等通过考虑视野自适应和局部信息内容加权,提出了 SQI(SCI Quality Index)算法[6],且基于主要视觉信息和不可预测的不确定性提取图像的统计特征,然后提出了半参考模型[7]。Shao 等[8]通过利用稀疏表示框架提出了 BLIQUP-SCI(Blind Quality Predictor for SCI)算法。Fang 等[9]通过对图像亮度和纹理特征的局部和全局表示,提出了 NRLT(No Reference quality assessment method by incorporating statistical Luminance and Texture features)算法。

现有的针对屏幕内容图像质量评估的无参考算法[8-9]不能与主观感知产生较高的一致性,因此,针对屏幕内容图像设计有效的无参考质量评估算法仍然存在挑战。本文结合文本、图形、图像和布局对屏幕内容图像质量的影响,提出了针对屏幕内容图像的基于边缘和结构的无参考质量评估(Blind quality assessment for screen content images based on Edge and Structure,BES)算法。

与自然图像不同,屏幕内容图像由具有大量边缘的文本、图形和图像组成,并且人类视觉系统对边缘高度敏感。因此,BES 算法首先对失真图像的亮度分量进行双三次插值处理,然后使用 Gabor 滤波器的虚部对插值后的亮度分量提取边缘,并计算每个失真图像的边缘特征。

屏幕内容图像以独特的布局显示文本、图形和图像,因此,BES 算法提取结构特征来表示屏幕内容图像的布局,首先使用双三次插值对失真图像进行插值,然后使用 Scharr 滤波器在插值后的失真图上计算得到局部二值模式(Local Binary Pattern,LBP)图,接着通过 LBP 图计算得到结构特征。与其他算法中直接使用频率直方图描述全局信息不同[9],BES 算法通过累加图像中具有相同 LBP 模式的像素的梯度值来表示失真图像的结构特征。

利用相应的方法提取边缘特征和结构特征,将随机森林回归(Random Forest Regression,RFR)算法作为映射函数,将边缘和结构特征映射为主观质量分数。

1 BES 算法

Guo 等[10]研究表明,通过双三次插值处理,可以减少屏幕内容图像和自然图像之间的统计特征差异,使得输入的屏幕内容图像在统计特征上更加类似于自然图像,以便更好地表示图像。Ni 等[11]通过实验证明了 Gabor 滤波器的虚部可以有效地提取边缘信息。文献[1]中表明,图像结构携带重要的视觉信息,人类视觉系统可以通过获取图像结构信息来感知和理解图像。因此,通过双三次插值处理之后,结合屏幕内容图像的边缘和结构特征来表示图像。最后,利用随机森林回归算法将从多尺度中提取的边缘和结构特征映射为主观质量分数。

1.1 提取边缘特征

根据生理学实验发现,二维 Gabor 滤波器可以有效地模拟哺乳动物视觉皮层中的简单细胞感受野剖面[12-13],说明了 Gabor 滤波器可以有效地表征人类视觉系统感知。因此,利用 Gabor 滤波器提取屏幕内容图像的边缘特征。

BES 算法首先将失真的屏幕内容图像从 RGB 颜色空间转换为 LMN 颜色空间[14],以便提取亮度分量。这里选择 LMN 颜色空间的原因是颜色空间转换过程中的权重针对人类视觉系统进行了优化[15]。接着使用双三次插值来对每个输入的失真屏幕内容图像的亮度分量进行插值,将其进行放大,平滑图像中的边缘。双三次插值的表达式为

$$p(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} a_{ij}x^i y^j \tag{1}$$

式中:a_{ij}为一个邻近像素的权重系数,这个权重系数是根据像素分布导数计算得来的。

在空间域中,二维 Gabor 滤波器可以描述为由正弦平面波调制的高斯核函数,其虚部是奇对称的并且是用于检测边缘的有效工具。文献[18-19]表明,水平或垂直方向的视觉灵敏度高于其他方向。因此,选择水平和垂直方向,即 $\theta=0$ 和 $\theta=\pi/2$(θ 为 Gabor 滤波器中的方向参数),以获得水平和垂直方向上的 Gabor 滤波器,分别表示为 $\boldsymbol{g}_{\mathrm{h}}$ 和 $\boldsymbol{g}_{\mathrm{v}}$。将通过双三次插值处理过的输入图像的亮度分量与每个 Gabor 滤波器进行卷积,以获得水平和垂直方向的边缘图,卷积计算公式为

$$\boldsymbol{e}_{\mathrm{h}}=\boldsymbol{g}_{\mathrm{h}}\otimes\boldsymbol{l} \tag{2}$$

$$\boldsymbol{e}_{\mathrm{v}}=\boldsymbol{g}_{\mathrm{v}}\otimes\boldsymbol{l} \tag{3}$$

式中:"⊗"表示卷积计算;$\boldsymbol{l}$ 为通过双三次插值处理过的输入图像的亮度分量;$\boldsymbol{e}_{\mathrm{h}}$ 和 $\boldsymbol{e}_{\mathrm{v}}$ 分别为水平和垂直方向的梯度图,接着将这 2 个结果相加得到最终的梯度图,表示为 $\boldsymbol{e}$,即

$$\boldsymbol{e}=\boldsymbol{e}_{\mathrm{h}}+\boldsymbol{e}_{\mathrm{v}} \tag{4}$$

这里将 $\boldsymbol{e}$ 的绝对值直方图作为输入图像的边缘特征,直方图分组设置为 10,因此用一个 10 维向量$\{f_1,f_2,\cdots,f_k\}$来表示边缘特征,向量中第 $n(1\leqslant n\leqslant k=10)$ 个元素的计算式为

$$f_n=\frac{1}{M}\sum_{i=1}^{M}\beta(|e_i|,Q(n)) \tag{5}$$

$$\beta(a,b)=\begin{cases}1 & a\in b\\ 0 & \text{其他}\end{cases} \tag{6}$$

式中:$Q(n)$ 为第 n 个分组的取值范围;M 为图像中的像素个数;e_i 为图像中第 i 个像素。

1.2　提取结构特征

文献[1]中表明,图像结构信息的有效提取和描述对图像感知质量评估有很大帮助。本文提出的算法中,先利用双三次插值对失真图像进行插值,接着分别使用水平和垂直方向的 Scharr 滤波器与插值后的图像进行卷积计算,得到梯度图,计算公式为

$$\boldsymbol{t}_{\mathrm{h}}=\boldsymbol{s}_{\mathrm{h}}\otimes\boldsymbol{p} \tag{7}$$

$$\boldsymbol{t}_{\mathrm{v}}=\boldsymbol{s}_{\mathrm{v}}\otimes\boldsymbol{p} \tag{8}$$

$$\boldsymbol{t}=\sqrt{(\boldsymbol{t}_{\mathrm{h}})^2+(\boldsymbol{t}_{\mathrm{v}})^2} \tag{9}$$

式中:$\boldsymbol{p}$ 为插值后的失真屏幕内容图像;$\boldsymbol{s}_{\mathrm{h}}$ 和 $\boldsymbol{s}_{\mathrm{v}}$ 分别为水平和垂直方向的 Scharr 滤波器;$\boldsymbol{t}_{\mathrm{h}}$ 和 $\boldsymbol{t}_{\mathrm{v}}$ 分别为插值后的图像和对应方向的 Scharr 滤波器经过卷积计算后得到的水平和垂直方向的插值后图像的梯度图;$\boldsymbol{t}$ 为最终的插值后图像的梯度图。

接着,基于上述得到的梯度图利用 LBP 算子来提取插值后图像的结构信息。LBP 算子是用来描述中心像素和其周围像素的关系[16],常规的 LBP 算子表达式为

$$\mathrm{LBP}_{I,R}=\sum_{i=0}^{I-1}u(t_i-t_c)2^i \tag{10}$$

$$u(t_i-t_c)=\begin{cases}1 & t_i-t_c\geqslant 0\\ 0 & t_i-t_c<0\end{cases} \tag{11}$$

式中:I 和 R 分别为周围像素数量和邻域的半径;t_c 为局部区域中心位置像素的梯度值;t_i 为邻接位置的梯度值。为了获得旋转不变性,局部旋转不变均匀 LBP 算子[16]定义为

$$\mathrm{LBP}'_{I,R}=\begin{cases}\sum_{i=0}^{I-1}u(t_i-t_c) & \rho(\mathrm{LBP}_{I,R})\leqslant 2\\ I+1 & \text{其他}\end{cases} \tag{12}$$

$$\rho(\mathrm{LBP}_{I,R})=\|u(t_{I-1}-t_c)-u(t_0-t_c)\|+\sum_{i=0}^{I-1}\|u(t_i-t_c)-u(t_{i-1}-t_c)\| \tag{13}$$

式中:ρ 为按位旋转的像素数量。局部旋转不变均匀 LBP 具有 $I+2$ 种模式,可以表示局部显著的梯度结构。受文献[17]的启发,该算法在梯度图中累加具有相同 LBP 模式的像素的梯度值,并将其作为结构特征,计算公式为

$$w(v)=\sum_{j=1}^{N}t_j f(\mathrm{LBP}_{I,R}(j),v) \tag{14}$$

$$f(d,r)=\begin{cases}1 & d=r\\ 0 & \text{其他}\end{cases} \tag{15}$$

式中:N 为图像的像素个数;$v\in[0,V]$ 为可能的 LBP 模式;t_j 为 LBP 模式的权重。在本文中,设置 $I=8$,邻域半径设置为 $R=1$,这样一张失真图的结构特征将由一个 10 维向量来表示。在图 1 中给出了 SIQAD 数据库中典型的失真图,及其对应的边缘图、LBP 图和表示特征的直方图。

1.3　回归模型

给定一张失真屏幕内容图像,根据上述算法,在每个尺度上可以获得 20 维特征,包括 10 维边缘特征和 10 维结构特征。文献[18]表明,人类视觉系统可以从粗略到细致地获得图像信息。因此,为了使提取的图像特征更符合人类视觉系统的特点,本文对图像进行 4 次下采样,并在这 4 个尺度上以及原尺度上提取特征。因此,对于每张失真图像,总共提取 100 维特征。在实验中,将双三次插值因子设置为 2,并将随机森林回归算法作为映射函数,把质量感知特征映射为主观质量分数。将随机从数据库中选择的 80% 图像用来训练模型,然后将剩余图像作为测试集并计算出其视觉质量分数。该实验进行 1 000 次,然后将其中位数作为最终的结果。

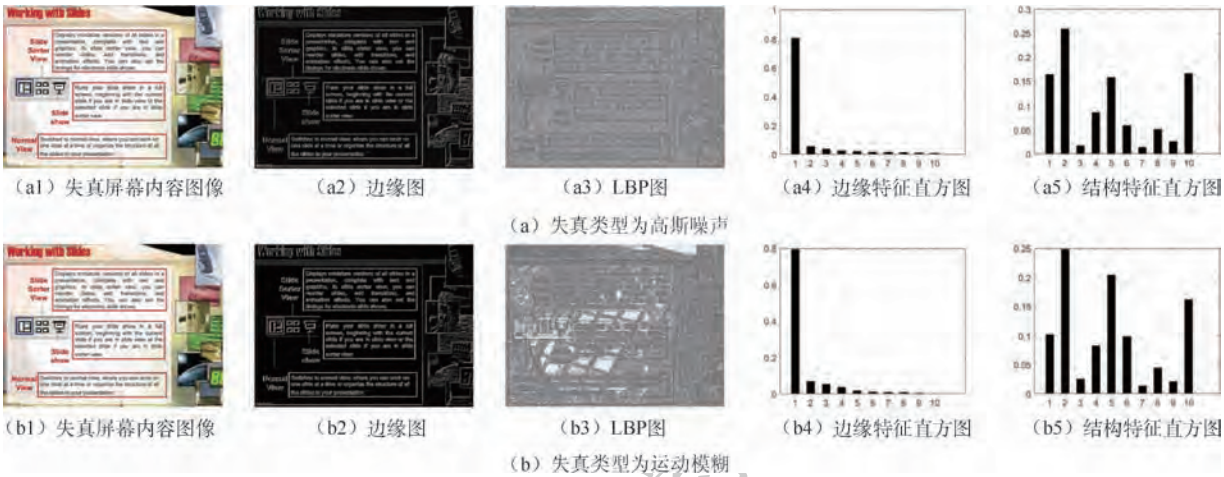


图 1 失真屏幕内容图像以及对应的特征图和特征直方图示例

Fig. 1 Examples of distorted screen content images, their feature maps, and feature histograms

2 实验结果

为了测试所提算法的性能,本文在 2 个数据库 SIQAD[5] 和 SCID[19] 上,将所提算法与其他评估算法进行对比实验,实验结果如表 1 和表 2 所示。SIQAD 数据库包含 20 张原始图像和 980 张失真图像,980 张失真图像有 7 种失真类型,每种失真类型有 7 种失真等级。这 7 种失真类型包括高斯噪声(GN)、高斯模糊(GB)、运动模糊(MB)、对比度变化(CC)、JPEG 压缩(JPEG)、JPEG2000 压缩(J2K),基于层划分的压缩(LSC)。SCID 数据库具有 40 张原始图像和1 800 张失真图像,失真图像中有 9 种失真类型,每种失真类型有 5 种失真等级。这 9 种失真类型包括 GN、GB、MB、CC、JPEG、J2K,颜色饱和度变化(CSC)、具有抖动的颜色量化(CQD),高效视频编码标准(HEVC-SCC)。这 2 个数据库都提供了平均主观得分差(DMOS)作为主观评分。

本文使用 3 个常用指标来评估主观和客观评分之间的一致性:皮尔森线性相关系数(Pearson Linear Correlation Coefficient, PLCC),斯皮尔曼等级相关系数(Spearman Rank-order Correlation Coefficient, SRCC)和根均方根误差(Root Mean Squared Error, RMSE)。PLCC 可用于评估预测的准确性,SRCC 可用于评估预测的单调性,RMSE 是评估预测一致性的一种方法。PLCC 和 SRCC 值越高,算法的性能越好。相反,较小的 RMSE 表示更好的精度。

不同的图像质量评估算法可能会产生不同范围的分数,为了将各种评估算法进行比较,就需要把评估分数映射到共同的分数空间,这里使用逻辑斯蒂函数对评估分数进行映射:

$$z(q) = p_1\left(\frac{1}{2} - \frac{1}{1+\mathrm{e}^{[p_2(q-p_3)]}}\right) + p_4 q + p_5 \quad (16)$$

式中:p_1、p_2、p_3、p_4 和 p_5 是 5 个拟合参数;q 为拟合前数据集;$z(q)$为拟合后数据集。

为了证明本文提出的算法的优越性,本文将其与以下经典的质量评估算法进行比较:PSNR、SSIM[1]、GMSD[2]、MAD(Most Apparent Distortion)[20]、SPQA[5]、SQI[6]、ESIM(Edge Similarity)[19]、GFM(Gabor Feature Model)[11]、NIQE[3]、IFC(Information Fidelity Criterion)[21]、BRISQUE[4]、GWH-GLBP(Gradient-Weighted Histogram of Local Binary Pattern calculated on the Gradient map)[17]、GSIM(Gradient Similarity)[22] 以及 NRLT[11]。在这些评估算法中,SPQA、SQI、ESIM、GFM 和 NRLT 是针对屏幕内容图像设计的算法,其余的算法是为自然图像设计的。

表 1 和表 2 分别列出了上述图像质量评估算法在 SIQAD 数据库和 SCID 数据库上每种失真类型以及整体性能的测试结果。在 2 个表列出来的全参考算法中,每个测量指标(即 PLCC、SRCC 和 RMSE)的最佳值用黑色粗体显示,而无参考算法中的每个测量指标的最佳值用黑色粗体加下划线显示。这些用来比较的算法的程序源代码都是从原始地址下载的。在表 1 中,对于 SQI 算法,失真类型上的 RMSE 值没有提供。

从表 1 中可以看出,针对屏幕内容图像设计的全参考算法,即 SPQA、SQI、ESIM 和 GFM,实现了比针对自然图像设计的全参考算法更好的性能,即 PSNR、SSIM、MAD 和 GMSD。在数据库 SIQAD 中,本文 BES 算法在所有无参考算法中取得最佳的性能,其中除了 NRLT 算法,其余算法都

表 1　SIQAD 数据库上的实验结果

Table 1　Experimental results on SIQAD database

指标	失真类型	全参考								无参考				
		PSNR	SSIM	MAD	GMSD	SPQA	SQI	ESIM	GFM	NIQE	BRISQUE	GWH-GLBP	NRLT	BES
PLCC	GN	**0.9053**	0.8806	0.8852	0.8956	0.8921	0.8829	0.8891	0.8990	0.8634	0.9045	0.8493	0.8983	**0.9090**
	GB	0.8503	0.9104	0.9120	0.9094	0.9058	0.9202	**0.9234**	0.9143	0.7560	0.8909	0.9098	0.8826	**0.9224**
	MB	0.7044	0.8060	0.8361	0.8436	0.8315	0.8789	**0.8886**	0.8662	0.5487	0.8571	0.8320	0.8611	**0.8694**
	CC	0.7401	0.7435	0.3933	0.7827	0.7992	0.7724	0.7641	**0.8107**	0.3555	0.5417	0.4108	0.7878	**0.8224**
	JPEG	0.7545	0.7487	0.7662	0.7746	0.7696	0.8218	0.7999	**0.8398**	0.5980	0.2887	0.5729	**0.7812**	0.7576
	J2K	0.7893	0.7749	0.8344	**0.8509**	0.8252	0.8271	0.7888	0.8486	0.5165	0.3509	0.7163	0.7489	**0.8182**
	LSC	0.7805	0.7307	0.8184	**0.8559**	0.7958	0.8310	0.7915	0.8288	0.5869	0.2865	0.5512	0.7309	**0.7575**
	ALL	0.5858	0.7561	0.5467	0.7291	0.8584	0.8644	0.8788	**0.8828**	0.4261	0.3591	0.7874	0.8482	**0.8705**
SRCC	GN	0.8790	0.8694	0.8721	**0.8856**	0.8823	0.8602	0.8757	0.8795	0.8429	0.8875	0.8287	0.8747	**0.8851**
	GB	0.8573	0.8921	0.9087	0.9119	0.9017	**0.9244**	0.9239	0.9132	0.6494	0.8715	0.8941	0.8656	**0.9045**
	MB	0.713	0.8041	0.8357	0.8441	0.8255	0.881	**0.8938**	0.8699	0.4272	0.8653	0.8229	0.8432	**0.8571**
	CC	0.6828	0.6405	0.3907	0.6378	0.6154	0.6677	0.6108	**0.7038**	0.1324	0.3135	0.2460	0.5996	**0.6582**
	JPEG	0.7568	0.7576	0.7674	0.7712	0.7673	0.8189	0.7989	**0.8434**	0.5109	0.2208	0.4803	**0.7386**	0.7072
	J2K	0.7746	0.7603	0.8382	0.8436	0.8152	0.8169	0.7827	**0.8444**	0.3238	0.3018	0.7017	0.7315	**0.7939**
	LSC	0.793	0.7371	0.8154	**0.8592**	0.8003	0.8432	0.7958	0.8445	0.3944	0.1844	0.5049	0.6675	**0.7291**
	ALL	0.5570	0.7566	0.5831	0.7305	0.8416	0.8548	0.8632	**0.8735**	0.3827	0.662	0.7194	0.8245	**0.8488**
RMSE	GN	**6.3372**	7.7044	6.9391	6.6354	6.7394		6.8272	6.6835	7.4083	6.0137	7.6829	6.3100	**6.1062**
	GB	7.7376	6.3619	6.2269	6.9816	6.4301		**5.8270**	6.1459	9.8026	6.4792	6.1466	6.6892	**5.5974**
	MB	9.2287	7.0600	7.13229	6.9816	7.2223		**5.9639**	6.5184	10.6136	6.1982	6.9347	6.4284	**6.4323**
	CC	8.4591	**6.8184**	11.565	7.8297	7.6184		8.1141	7.3638	11.4678	9.966	11.198	7.5601	**6.7135**
	JPEG	6.1165	5.6406	6.038	5.9427	6.0000		5.6401	**5.1009**	7.2869	8.6037	7.3238	**5.7121**	5.9880
	J2K	6.3819	6.1804	5.7276	**5.4591**	5.8706		6.3877	5.4985	8.3318	9.094	6.7750	6.5058	**5.7048**
	LSC	5.3336	4.9379	4.9025	**4.4121**	5.1664		5.2150	4.7736	6.8159	7.8712	7.0289	5.7836	**5.3908**
	ALL	11.601	10.855	11.986	9.7972	7.3421	7.1982	6.8310	**6.7234**	12.7941	11.717	8.6726	7.4447	**6.9147**

表 2　SCID 数据库上的实验结果

Table 2　Experimental results on SCID database

指标	失真类型	全参考					无参考				
		PSNR	SSIM	MAD	IFC	GSIM	NIQE	BRISQUE	GWH-GLBP	NRLT	BES
PLCC	GN	**0.9530**	0.9354	0.9315	0.8897	0.9170	0.8324	0.9645	0.8619	**0.9681**	0.9398
	GB	0.7772	**0.8711**	0.8559	0.8406	0.8449	0.5912	0.5847	0.7144	0.6672	**0.8553**
	MB	0.7615	**0.8794**	0.8362	0.3372	0.8383	0.5390	0.6434	0.6736	0.5906	**0.9314**
	CC	0.7435	0.6903	0.4987	0.1198	**0.8675**	0.2840	0.4734	0.2520	0.4993	**0.8122**
	JPEG	0.8393	0.8581	0.9251	0.8762	**0.9373**	0.6824	0.5879	0.7694	**0.8512**	0.8205
	J2K	0.9176	0.8586	0.9381	0.8570	**0.9441**	0.7099	0.5515	0.6645	**0.8326**	0.7954
	CSC	0.0622	0.0890	**0.1296**	0.0764	0.0560	0.2196	0.1879	0.2134	0.1963	**0.3110**
	HEVC-SCC	0.7991	0.8635	**0.8953**	0.7918	0.8835	0.4289	0.4418	**0.6142**	0.5334	0.5875
	CQD	**0.9210**	0.8668	0.9014	0.7655	0.8974	0.5787	0.7146	0.6034	0.7250	**0.8303**
	ALL	0.7622	0.7579	**0.7719**	0.6285	0.7042	0.3392	0.6303	0.6647	0.7060	**0.7852**
SRCC	GN	**0.9424**	0.9171	0.9262	0.8877	0.9112	0.8505	0.9871	0.8669	**0.9612**	0.9333
	GB	0.702	**0.8698**	0.8603	0.8351	0.8420	0.3462	0.5162	0.7019	0.6250	**0.8479**
	MB	0.7375	**0.8588**	0.8296	0.4477	0.8194	0.3691	0.6265	0.6419	0.5123	**0.7883**
	CC	0.7265	0.6564	0.4784	0.1198	**0.8304**	0.1004	0.3096	0.2314	0.3031	**0.4773**
	JPEG	0.8231	0.8490	0.9242	0.8770	**0.9366**	0.6291	0.5469	0.7430	**0.8382**	0.7957
	J2K	0.9074	0.8439	0.9330	0.8457	**0.9349**	0.6540	0.4997	0.6398	**0.7916**	0.7634
	CSC	0.0908	0.0963	**0.1440**	0.0521	0.1214	0.0401	0.0143	0.0986	0.1009	**0.1098**
	HEVC-SCC	0.8074	0.8263	**0.8771**	0.7869	0.8730	0.3921	0.2444	**0.4851**	0.4524	0.4654
	CQD	**0.9080**	0.7766	0.9024	0.7368	0.8634	0.4397	0.5580	0.5046	0.5923	**0.7548**
	ALL	0.7512	0.7146	**0.7576**	0.5799	0.6993	0.2727	0.6066	0.6211	0.6866	**0.7613**
RMSE	GN	**3.8093**	4.4458	4.5714	5.7380	5.0127	6.9602	3.2369	6.3623	**3.1679**	4.5162
	GB	6.6633	**5.1998**	5.4475	5.7354	5.6648	8.4422	8.4767	7.3110	7.7567	**5.4365**
	MB	7.0843	**5.2044**	5.9947	10.2905	5.9607	8.9849	8.2437	7.9311	8.6979	**6.4253**
	CC	5.9867	6.4767	7.7590	8.8876	**4.4524**	8.4821	7.7178	8.5698	7.6891	**5.1836**
	JPEG	8.1718	7.7179	5.7076	7.2431	**5.2369**	10.9001	12.1386	9.4948	**7.8897**	8.6342
	J2K	6.3222	8.1562	5.5103	8.1986	**5.2462**	11.1673	13.1257	11.5863	**8.7148**	9.2540
	CSC	9.8203	9.8003	**9.7564**	9.8106	9.8239	9.3489	9.4966	9.5234	9.5356	**9.2654**
	HEVC-SCC	8.4009	8.5037	**6.1988**	8.4969	6.5176	12.3695	12.3545	**10.8270**	11.5562	11.1382
	CQD	**4.9814**	7.9855	5.5354	8.2269	5.6406	10.3763	8.7083	10.0025	8.8508	**7.1015**
	ALL	9.1682	9.6133	**8.9739**	11.0157	10.0552	13.3271	10.9967	10.5205	9.9811	**8.8319**

是针对自然图像提出的,只考虑了图像部分的特征,忽略了文字部分的特点,使得性能不佳;NRLT 算法结合亮度和结构特征来表示图像,但是图像中存在大量文字,文字对亮度的敏感度低于图像部分,而本文 BES 算法利用图像中存在大量边缘这一特性,结合图像的特殊布局来表示图像,PLCC 指标相比 NRLT 算法提高了 2.63%。

本文 BES 算法的整体性能甚至高于专门为屏幕内容图像设计的全参考算法 SPQA 和 SQI。其中,SPQA 算法主要是考虑图像的亮度以及锐度特征,SQI 算法结合图像局部结构相似性以及局部信息内容加权来计算图像分数,但是这 2 个算法都没有考虑图像中文字存在大量边缘这一特点,导致性能较低。在单个失真类型上,BES 算法的性能除了在 JPEG 失真类型上略低于 NRLT 算法外,在其他失真类型上的性能都是无参考算法中最好的。

从表 2 中可以看出,在 SCID 数据库中,BES 算法的整体性能在 3 个指标上不仅高于对比的无参考算法,而且高于经典的全参考算法,即 PSNR、SSIM、MAD、IFC 和 GSIM。其中,NRLT 是对比方法中是先进的无参考算法,所提算法 PLCC 指标相比其提高了 11.22%。在单个失真类型上,除了在 GN、JPEG、J2K 以及 HEVC-SCC 这4 种失真类型上性能比其他无参考算法低以外,对于其余 5 种失真类型,BES 算法的性能在无参考算法中都是最高的。

3 结 论

本文根据人类视觉系统特点提出了一种新的针对屏幕内容图像的基于边缘和结构的无参考质量评估(EBS)算法,此算法是基于边缘信息和结构信息。提出的 BES 算法经实验验证,得到:

1) 本文算法考虑到屏幕内容图像具有特殊布局以及丰富边缘这 2 个特征,利用 Gabor 滤波器和 LBP 算子分别提取边缘和结构特征,并从 5 个尺度上提取特征,实现了与主观感知较高的一致性。

2) 本文算法可以实现对多种失真类型的屏幕内容图像较好的质量评估效果,在数据库 SIQAD 和 SCID 上的性能都优于经典的评估算法,甚至优于一些全参考算法。

通过实验验证,本文算法的屏幕内容图像质量评估效果取得一定的提升,提高了与主观感知的一致性。在未来,可以分别从图像区域和文字区域考虑,进一步提高该图像的质量评估效果。

参考文献(References)

[1] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[2] XUE W, ZHANG L, MOU X, et al. Gradient magnitude similarity deviation: A highly efficient perceptual image quality index [J]. IEEE Transactions on Image Processing, 2014, 23(2): 684-695.

[3] MITTAL A, SOUNDARARAJAN R, BOVIK A C. Making a 'completely blind' image quality analyzer[J]. IEEE Signal Processing Letters, 2013, 20(3): 209-212.

[4] MITTAL A, MOORTHY A K, BOVIK A C. No-reference image quality assessment in the spatial domain[J]. IEEE Transactions on Image Processing, 2012, 21(12): 4695-4708.

[5] YANG H, FANG Y, LIN W. Perceptual quality assessment of screen content images[J]. IEEE Transactions on Image Processing, 2015, 24(11): 4408-4421.

[6] WANG S, GU K, ZENG K, et al. Objective quality assessment and perceptual compression of screen content images[J]. IEEE Computer Graphics and Applications, 2016, 38(1): 47-58.

[7] WANG S, GU K, ZHANG X, et al. Reduced-reference quality assessment of screen content images[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2016, 28(1): 1-14.

[8] SHAO F, GAO Y, LI F, et al. Toward a blind quality predictor for screen content images[J]. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2017, 48(9): 1-10.

[9] FANG Y, YAN J, LI L, et al. No reference quality assessment for screen content images with both local and global feature representation[J]. IEEE Transactions on Image Processing, 2018, 27(4): 1600-1610.

[10] GUO X, HUANG L, GU K, et al. Naturalization of screen content images for enhanced quality evaluation[J]. IEICE Transactions on Information and Systems, 2017, 100(3): 574-577.

[11] NI Z, ZENG H, MA L, et al. A Gabor feature-based quality assessment model for the screen content Images[J]. IEEE Transactions on Image Processing, 2018, 27(9): 4516-4528.

[12] DAUGMAN J G. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters[J]. Journal of the Optical Society of America A, 1985, 2(7): 1160-1169.

[13] JONES J P, PALMER L A. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex[J]. Journal of Neurophysiology, 1987, 58(6): 1233-1258.

[14] GEUSEBROEK J M, VAN DEBOOMGAARD R, SMEULDERS A W M, et al. Color invariance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2001, 23(12): 1338-1350.

[15] GEUSEBROEK J M, VAN DEBOOMGAARD R, SMEULDERS A W M, et al. Color and scale: The spatial structure of color images[C]// Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2000: 331-341.

[16] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local

binary patterns[J]. Classification with Local Binary Patterns, 2002,24(7):971-987.

[17] LI Q, LIN W, FANG Y. No-reference quality assessment for multiply-distorted images in gradient domain[J]. IEEE Signal Processing Letters,2016,23(4):541-545.

[18] HUGHES H C, NOZAWA G, KITTERLE F. Global precedence, spatial frequency channels, and the statistics of natural images[J]. Journal of Cognitive Neuroscience, 1996, 8(3): 197-230.

[19] NI Z, MA L, ZENG H, et al. ESIM: Edge similarity for screen content image quality assessment[J]. IEEE Transactions on Image Processing,2017,26(10):4818-4831.

[20] ERIC C L, DAMON M C. Most apparent distortion: Full-reference image quality assessment and the role of strategy[J]. Journal of Electronic Imaging,2010,19(1):011006.

[21] SHEIKH H R, BOVIK A C, DE VECIANA G, et al. An information fidelity criterion for image quality assessment using natural scene statistics[J]. IEEE Transactions on Image Processing,2005,14(12):2117-2128.

[22] LIU A, LIN W, NARWARIA M. Image quality assessment based on gradient similarity[J]. IEEE Transactions on Image Processing,2012,21(4):1500-1512.

作者简介:

魏乐松　男,硕士研究生。主要研究方向:多媒体。

陈俊豪　男,博士研究生。主要研究方向:多媒体、计算机图像学和人工智能。

牛玉贞　女,博士,教授,博士生导师。主要研究方向:计算机视觉、多媒体和计算机图像学。

Blind quality assessment for screen content images based on edge and structure

WEI Lesong, CHEN Junhao, NIU Yuzhen*

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350100, China)

Abstract: A screen content image (SCI) has great difference compared with a natural image, and an SCI contains more text, graphic, and special layout. Considering the influences of texts, graphics, pictures, and layout on quality of an SCI, a blind quality assessment metric for SCIs based on edge and structure (BES) has been proposed. Since texts, graphics, and pictures have a large number of edges and the human visual system is highly sensitive to edges, the BES metric first extracts edges using the imaginary part of the Gabor filter and computes an edge feature for each SCI. Second, a structure feature is extracted to represent the layout of an SCI. Specifically, the Scharr filter is exploited to calculate a local binary pattern (LBP) map which is used to compute a structure feature. Finally, a random forest regression algorithm is applied to map the edge and structure features to subjective scores. The experimental results show that in the database SIQAD and SCID, the Pearson linear correlation coefficient (PLCC) of the performance of the proposed BES index is 2.63% and 11.22% higher than the latest no reference index in the comparison respectively, and even higher than some full reference indexes.

Keywords: screen content image; blind quality assessment; random forest regression; edge feature; structure feature

Received: 2019-07-09; **Accepted**: 2019-08-03; **Published online**: 2019-08-12 16:20

URL: kns.cnki.net/kcms/detail/11.2625.V.20190812.1333.001.html

Foundation items: National Natural Science Foundation of China (61672158); Natural Science Foundation of Fujian Province (2019J2006)

* **Corresponding author**. E-mail: yuzhenniu@gmail.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0383

基于 HTTP 自适应流媒体传输的 3D 视频质量评价

翟宇轩，刘怡桑，徐艺文，陈忠辉，房颖*，赵铁松
（福州大学 物理与信息工程学院，福州 350108）

摘 要：3D 视频网络服务的关键在于提高用户的体验质量(QoE)，而体验质量往往会由于网络环境的变化及视频内容的不同而受到影响。传统的 2D 视频传输可以采用基于 HTTP 的自适应流媒体(HAS)速率自适应机制有效地利用网络带宽，提高用户体验质量。因此对于如何利用动态自适应流媒体技术实现至少需要传输两路视频流的 3D 网络视频服务已经越来越被关注。HAS 技术的关键在于媒体质量级别的动态转换策略，主要研究了 3D 视频中不同视点比特率的变化对用户观看体验质量的影响。首先，建立一个主观数据库探讨块级客观质量与 3D 视频的视觉体验质量之间的关系，块级客观质量将随着比特率的变化而变化。其次，提出了一种基于卷积神经网络(CNN)的 QoE 模型，该模型可以通过块级客观质量有效地评估 QoE，模型预测值和平均意见分(MOS)的皮尔森线性相关系数(PLCC)为 0.906，可在自适应流媒体应用中为 3D 视频传输中不同视点的码率调整提供指导。

关 键 词：3D 视频；体验质量；视频质量评价；卷积神经网络(CNN)；HTTP 自适应流媒体
中图分类号：TN919.8
文献标识码：A **文章编号**：1001-5965(2019)12-2456-07

近年来，网络技术的发展推动了在线 3D 视频服务的兴起。相比于传统的 2D 视频，3D 视频即使进行了有效的压缩其数据量仍然较大，而有限的网络带宽资源导致了 3D 视频传输中的质量波动，从而使得 3D 视频网络服务的用户体验质量(Quality of Experience，QoE)[1]不高。因此，3D 视频的传输引入了基于 HTTP 的自适应流媒体(HTTP Adaptive Streaming，HAS)技术。HAS 可以根据不同的网络带宽情况提供不同码率的视频，在避免卡顿的同时尽可能利用有限的带宽提升视频质量，改善用户体验质量。

由于 3D 视频需要额外考虑视点间的码率分配，3D 视频的 HAS 技术比传统 2D 视频的更加复杂。目前，针对 3D 视频 HAS 的相关工作还比较少，文献[2]提出了交互式多视点 HAS 的最优传输策略来平衡编码的失真和渲染合成失真，该算法同时考虑了视频内容特征和用户交互度。文献[3]利用软件定义网络(Software Defined Network，SDN)提出了基于视频块的流媒体传输框架，这个方案可以根据感兴趣区域改善用户体验质量。文献[4]提出了一种基于 HTTP 动态自适应流媒体(Dynamic Adaptive Streaming over HTTP，DASH)的高效 3D 自适应流媒体服务方法，其能为用户提供流畅的立体视频。尽管上述工作已经为 3D 流媒体应用提出了有效的方案，但是为了给用户提供最佳的观看体验，评估不同流媒体自适应方案的用户体验质量至关重要[5]。目前，对于基于 HAS 技术的 3D 视频传输中的用户体验

收稿日期：2019-07-09；**录用日期**：2019-08-14；**网络出版时间**：2019-08-22 11:24
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190822.1122.004.html
基金项目：国家自然科学基金（61671152）
* **通信作者**. E-mail：fangying@fzu.edu.cn

引用格式：翟宇轩，刘怡桑，徐艺文，等. 基于HTTP自适应流媒体传输的3D视频质量评价[J]. 北京航空航天大学学报，2019，45(12)：2456-2462. ZHAI Y X，LIU Y S，XU Y W，et al. 3D video quality evaluation based on adaptive streaming transmission over HTTP[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2456-2462 (in Chinese).

的研究越来越受到关注。QoE 模型反映了客观质量与用户体验质量之间的关系,可以极大地帮助 3D HAS 系统的设计和优化。

文献[6]发现相对于视频质量的瞬间急剧变化,视频质量由低到高缓慢变化时的用户观看体验质量更高。基于这一特性,提出了一种适用于 DASH 的 QoE 自适应算法。文献[7]通过分析 2D 视频和 3D 视频的自适应流媒体传输策略,发现由 3D 到 2D 的转换可能是降低比特率的最佳选择,而相反的由 2D 到 3D 的转换并没有明显改善用户的体验质量。虽然文献[7]定性分析了 3D 视频质量切换对感知质量的影响,但是仍然缺乏可以用于指导 3D 视频传输时质量切换的 QoE 的量化模型。另外,已有大量工作致力于 3D 图像的客观质量评价[8-10],这些评价算法可以准确反映 3D 内容的感知质量,但无法用于表达视频质量切换导致的 QoE 变化。

为了研究 3D 视频传输过程中网络质量波动(Network Quality Fluctuation,NQF)对用户体验质量的影响,本文设计了主观实验用于获取用户在 NQF 情况下观看视频的体验质量。主观实验特别考虑了单视点和双视点的视频质量改变分别对 3D 视觉感知质量的影响。最后,提出了一个基于卷积神经网络(Convolutional Neural Networks,CNN)的 QoE 模型,该模型体现了块级客观质量与用户对 3D 视频的观看体验质量的映射关系,可用于指导 3D 视频自适应传输中的视点间码率分配。

1　网络质量波动的影响

双目立体 3D 视频(Stereoscopic 3D video)通常包含左右 2 个视点,其所需的带宽远大于传统的 2D 视频。3D 视频的自适应传输在带宽不足情况下会面临视频质量的突降以及左右视点的比特率平衡等问题。本节通过主观实验来分析单视点和双视点视觉质量的改变对 3D 感知质量的影响。文献[11]表明,主视眼的不同对总体感知质量的影响可忽略不计。因此,实验中单视点的图像质量变化均基于左视点的图像。

实验总共使用了 13 个 3D 视频序列,其中包括 3MV-HEVC 数据库[12] 中的 CP (Carpark)、SK (Shark)、ST (Street)、GF (Gtfly)、KD (Kendo)、LB (Lovebird)、BN (Balloons)、BA (Bookarrival),以及数字音频编解码技术标准工作组(AVS)数据库[13] 中的 BM (Badminton)、JL (Jinli)、DB (Dubai)、AG (Asiangame)、WS (Wushu)。图 1 为各个序列的截图,表 1 为相应的空间信息(Spatial Information,SI)和时间信息(Temporal Information,TI)[14]。所有视频序列时长为 10 s,视频质量的变化发生在视频序列第 5 s 末,即每个序列的前5 s和后 5 s 拥有不同的视频质量。为了避免引入由分辨率不同造成的体验质量差别,高分辨率视频均采用了下采样处理,所有序列的分辨率为1 024 × 768,帧率为 25 帧/s。

NQF 实验按照 4 种比特率编码视频:

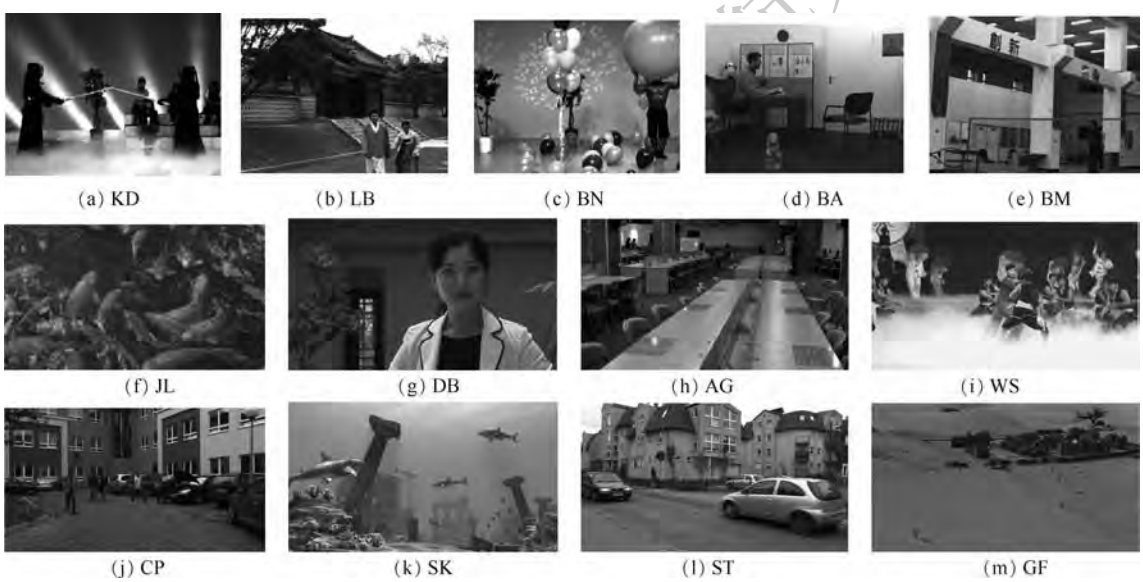


图 1　3D 序列截图[12-13]

Fig. 1　Snapshots of 3D sequences[12-13]

1)“全比特率”表示足够的带宽使得视点质量近似原画质,视点比特率设置为 1 000 kbit/s。

2)“高比特率”模拟网络轻微阻塞时的视点比特率,设置为 200 kbit/s。

3)“中比特率”模拟网络遭受中等阻塞时的视点比特率,设置为 100 kbit/s。

4)“低比特率”模拟网络遭受严重阻塞时的视点比特率,设置为 50 kbit/s。

由表 2 可知,模拟的 NQF 类型包含 4 种质量的切换,分别为单视点质量上升、双视点质量上升、单视点质量下降和双视点质量下降。其中质量上升和下降过程被细分为 6 种比特率变化:低到全、中到全、高到全;全到低、全到中、全到高。使用 3D-HEVC 标准参考软件 HTM 16.0 作为编码器,13 个原始序列依据 12 种视频质量切换类型共生成 156 组 3D 测试序列。实验中采用的观看设备为华硕 PG278 3D 屏幕和 NVIDIA 3D 眼镜。本次实验共有 34 名受试者,包括 23 名男性和11名女性,年龄介于21 ~25岁。所有受试者都通过了视力测试并且在观看 3D 视频中没有产生不适感。在主观测试之前,受试者都已熟悉 3D 视频显示方式和实验流程,并将显示屏和眼镜调整到舒适的位置。主观测试遵循 ITU-R BT.500[15] 建议书推荐的单激励(Single Stimulus,SS)方法和五级损伤量表[15]。在测试期间,所有 NQF 测试序列在随机打乱顺序后连续显示,每个序列结束后都有 5 s 的间隔用于评分。

表 1 测试数据集[12-13]

Table 1 Test dataset[12-13]

分辨率	序列名称	空间信息	时间信息
1 024 × 768	KD	40.07	13.96
	LB	57.83	4.05
	BN	37.48	8.39
	BM	56.15	9.07
1 920 × 1 080	JL	42.56	24.83
	DB	35.72	2.56
	AG	78.73	17.88
	WS	54.57	17.27
1 920 × 1 088	CP	65.11	5.98
	SK	23.65	15.05
	ST	57.58	8.47
	GF	51.96	16.11

表 2 网络质量波动类型

Table 2 Network quality fluctuation types

质量变化	视点	编号	前 5 s 视频质量(左/右)/(kbit·s^{-1})	后 5 s 视频质量(左/右)/(kbit·s^{-1})
上升	单视点	1	50/1 000	1 000/1 000
		2	100/1 000	1 000/1 000
		3	200/1 000	1 000/1 000
	双视点	4	50/50	1 000/1 000
		5	100/100	1 000/1 000
		6	200/200	1 000/1 000
下降	单视点	7	1 000/1 000	50/1 000
		8	1 000/1 000	100/1 000
		9	1 000/1 000	200/1 000
	双视点	10	1 000/1 000	50/50
		11	1 000/1 000	100/100
		12	1 000/1 000	200/200

为了提高数据的可靠性,每个受试者的测试都引入重复序列。数据结果采用 ITU-R BT.500 建议书中的可靠性原则[15]来排除不可靠的分数。24 名受试者对视频质量的评分值被保留至后续的数据分析。为了检查选择的样本量是否足以产生稳定的结果,以“数据饱和度”作为指导原则[16]。受试者人数上升导致的平均意见得分(Mean Opinion Score,MOS)数据饱和曲线如图 2 所示,每个受试者对 13 个序列的主观评分为 $\boldsymbol{a}_n$,选取 m 个受试者的主观评分均值为

$$\boldsymbol{x}_m = \frac{1}{m}\sum_{n=1}^{m}\boldsymbol{a}_n \tag{1}$$

所有 24 个主观评分均值为 $\boldsymbol{s}$,$\boldsymbol{x}_m$ 和 $\boldsymbol{s}$ 之间的皮尔森线性相关系数(Pearson Linear Correlation Coefficient,PLCC)随着选取人数 m 增加而增大,“饱和值”出现在受试者人数达到 20 时,这表明本次实验采用 24 个样本值已足够。

根据 ITU-R BT.500 建议书[15],所有受试者的 MOS 表现了主观评分等级。图 3 给出了不同比特率切换时单目和双目质量波动的主观评分。图 3(a)和(b)分别表示单视点和双视点质量上升的 MOS 值,可以看出,单视点质量切换比双视点质量切换引起了更小的用户体验质量下降;图 3(c)和(d)的比较同样可以发现单视点质量切换对体验质量的影响更小。该结果符合双目视觉的掩蔽特性,当一个视点质量不变,另一个视点质量下降至一定范围内,人眼无法察觉到失真[17]。

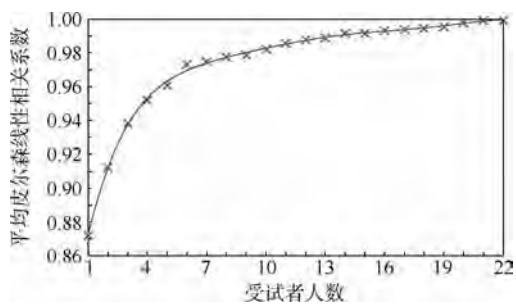


图 2 受试者人数上升导致的 MOS 数据饱和

Fig. 2 MOS data saturation caused by increased number of subjects

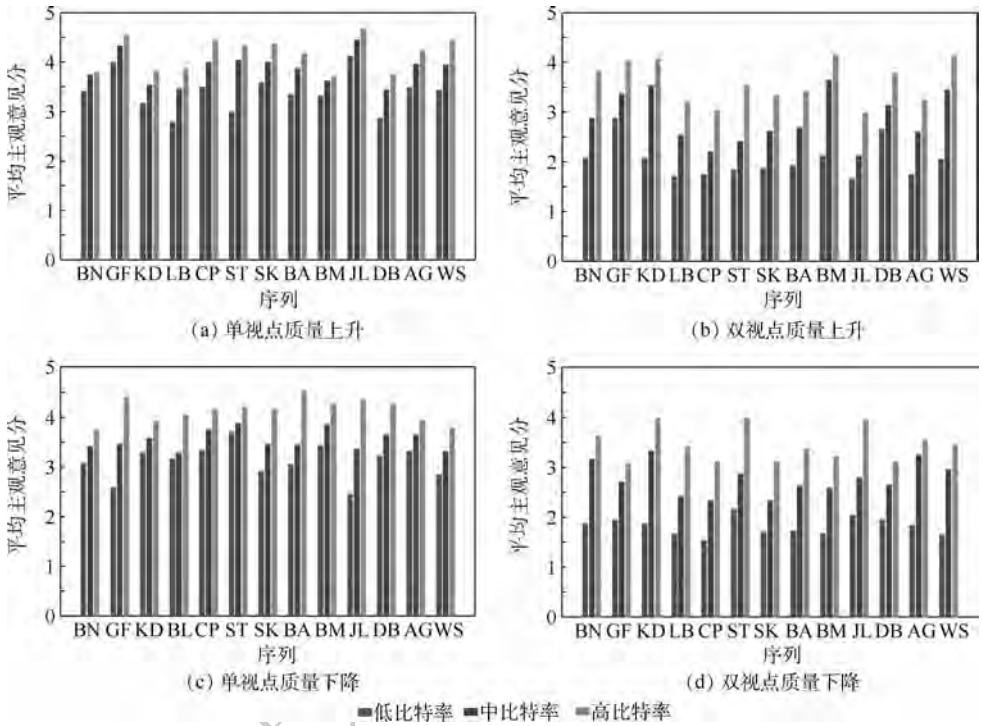


图 3　NQF 主观实验结果

Fig. 3　Subjective experimental results of NQF

该结论已应用至 3D 视频的非对称编码[18-19]来减少视觉冗余,也成为 3D 视频非对称码率传输的基础[20]。

为了研究上升和下降 2 种视频质量切换的区别,表 3 显示了同种类型质量变化下所有 13 个视频序列 MOS 的平均值。由表 3 可知,无论何种视频质量(低、中、高)和视点(单视点、双视点),视频质量上升的评分总高于质量下降的评分。原因是,在视频质量上升类型中,10 s 序列的后 5 s 为高画质,短时记忆的影响使得受试者的最终评价分数更接近后出现的高画质的分数;同理,质量下降类型中后出现的低画质降低了整体评分。这个现象是由于用户体验质量会受到心理学中的近因效应影响,即前一时刻的体验质量会对之后一段时间内的观看体验造成影响[21-22]。当用户接受高质量视频时会产生更高视频质量的期望,因此视频质量的下降会使得用户更加沮丧,加速用户体验质量的下降;相反,观看低质量视频的用户则对视频质量的上升更加敏感,提高了体验质量的评分。本文主观实验中,视频质量的上升和下降 2 种切换表现出了不同体验质量评分,这一结果与现有时变视频质量研究[21-22]一致。这说明在 3D 视频自适应传输策略方面:带宽受限时,视频质量可以避免出现突然剧烈的下降;带宽充足时,可以快速提升视频的质量。

表 3　视频质量波动类型的 MOS 均值

Table 3　Average MOS of video quality fluctuation types

视频质量类型	单视点升	单视点降	双视点升	双视点降
低质量	3.39	3.11	2.00	1.83
中质量	3.88	3.55	2.87	2.78
高质量	4.17	4.11	3.61	3.47

2　3D QoE 模型

目前,主观测试虽然是衡量用户体验质量的最可靠方法,但存在高成本,低速度和无法进行实时评估等缺点。客观 QoE 模型是一种获得近似主观评分的低成本方法。利用第 1 节方法所获得的视频主观质量数据库及 CNN 构建了 3D 视频的客观 QoE 模型。将 3D 测试序列和原始序列的每一帧划分为 64×64 子块,并计算左右视图中测试序列的子块与相对应原始序列的子块的结构相似性(Structural Similarity, SSIM)[23-24],左右视点所有帧的块级 SSIM 分别表示为 $\boldsymbol{q}_{\mathrm{l}}$ 和 $\boldsymbol{q}_{\mathrm{r}}$。所得到的 $\boldsymbol{q}_{\mathrm{l}}$ 和 $\boldsymbol{q}_{\mathrm{r}}$ 输入到 CNN 网络中预测用户的体验质量值。因此,本文提出的 QoE 模型为

$$Q = f(\boldsymbol{q}_{\mathrm{l}}, \boldsymbol{q}_{\mathrm{r}}) \tag{2}$$

式中:Q 表示用户体验质量的预测值,预测函数基于 CNN 完成。

如图 4 所示,本文设计的 CNN 模型由 2 层卷积层和 2 层全连接层组成。卷积核的大小分别设置为 5×5、3×3。图中:M 和 N 分别为视频的宽

和高，K 为视频帧数。输入为测试序列的块级 SSIM，输出为用户体验质量的预测值 Q。在网络训练过程中，从 156 组 3D 测试序列中随机选取 141 个样本作为训练集，用于训练并验证 MOS 和 3D 视频内容的块级 SSIM 之间的关系，剩余 15 个样本作为测试集用作最后的测试。

在图像视频质量评价中，通常通过斯皮尔曼秩相关系数（Spearman Rank Order Correlation Coefficient，SROCC）、肯德尔秩相关系数（Kendall Rank Order Correlation Coefficient，KROCC）、PLCC 来评价所提模型的性能。其相关系数能够用于反映客观质量评价与主观 MOS 值的相关程度，值越接近于 1，则说明模型的性能越好。由表 4 可知，测试集中所有序列的主观评价 MOS 值与模型预测值之间的 SROCC、KROCC、PLCC 分别为 0.927、0.775、0.906，评估结果说明基于 CNN 的 QoE 模型能够较好地预测用户体验质量。

由于缺少同类型数据库和模型进行比较，本文设置了 3 组对比实验：①将每个视频序列的 2 个视点 PSNR 的平均值作为预测 QoE；②将每个视频序列的 2 个视点 SSIM 的平均值作为预测 QoE；③采用文献[10]的 3D 质量评价算法。实验测试使用了同样配置的电脑（Intel Core i5-7500，8GB RAM，Windows 10 64-bit）。实验结果如表 4 所示，可以看出，SSIM、PSNR 无法很好地反映用户体验质量，文献[10]的 3D 质量评价算法在本文的数据库上表现同样不佳，但优于 PSNR、SSIM，这是由于 3D 质量评价方法针对于评价 3D 立体图像和视频的压缩失真，考虑了图像失真类型和人眼双目视觉特性而没有考虑真实传输过程中的视频质量的变化。本文还记录了不同算法测试一个 10 s 序列的运行时间，本文模型需要计算块级 SSIM，导致运行时间大于直接计算 3D 视频的 PSNR、SSIM，但是性能远好于其他 3 种方法。

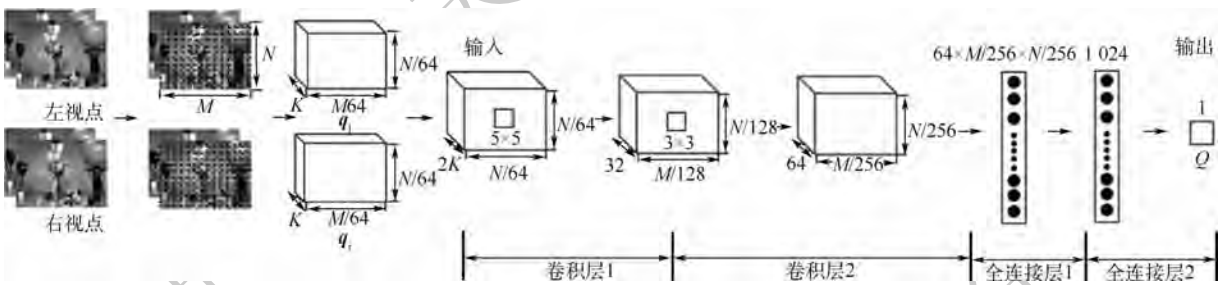


图 4 3D QoE 模型框架

Fig. 4 Framework of 3D QoE model

表 4 QoE 模型和其他方法的性能比较

Table 4 Comparison of performance between QoE model and other methods

指标	QoE 模型	PSNR	SSIM	文献[10]
SROCC	0.927	0.357	0.464	0.550
KROCC	0.775	0.257	0.371	0.390
PLCC	0.906	0.411	0.460	0.441
时间/s	842	6.8	29.0	4 770

3 结 论

本文研究了 3D 自适应流媒体应用中的用户体验质量，并设计了网络带宽不稳定情况下的 3D 视频质量变化的主观实验。

实验结果表明受试者对单视点质量变化不敏感，并且视频质量上升状态的用户体验质量更高。通过训练主观评分数据，本文建立了基于 CNN 的 QoE 模型用于评估 3D 视频在自适应传输中的用户体验质量。该模型提供了高精度的 QoE 预测值，可用于改进 3D 自适应传输和平衡 3D 视频传输中两视点间比特率。在流媒体视频服务中，代理服务器可以根据该模型预测得到的 QoE 为用户提供不同码率的 3D 视频，从而有效分配网络带宽资源。

参考文献（References）

[1] MILLER K, QUACCHIO E, GENNARI G, et al. Adaptation algorithm for adaptive streaming over HTTP[C]// Proceedings of the 19th International Packet Video Workshop. Piscataway, NJ: IEEE Press, 2012: 173-178.

[2] ZHANG X, TONI L, FROSSARD P, et al. Adaptive streaming in interactive multiview video systems[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 29(4): 1130-1144.

[3] ZHAO S, MEDHI D. SDN-Assisted adaptive streaming framework for tile-based immersive content using MPEG-DASH [C]// Proceedings of the IEEE Conference on Network Function Virtualization and Software Defined Networks. Piscataway, NJ: IEEE Press, 2017: 1-6.

[4] PARK G, LEE J, LEE G, et al. Efficient 3D adaptive HTTP streaming scheme over internet TV[C]// Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting. Piscataway, NJ: IEEE Press, 2012: 1-6.

[5] GARCIA M N, SIMONE F D, TAVAKOLI S, et al. Quality of Experience and HTTP adaptive streaming: A review of subjective studies[C]// Proceedings of the 6th International Workshop on Quality of Multimedia Experience, QoMEX. Piscataway, NJ: IEEE Press, 2014: 141-146

[6] MOK R, LUO X, CHAN E, et al. QDASH: A QoE-aware DASH system[C]// Proceedings of the 3rd Multimedia Systems Conference. New York: ACM Press, 2012: 11-22.

[7] TAVAKOLI S, GUTIERREZ J, GARCIA N. Subjective quality study of adaptive streaming of monoscopic and stereoscopic video[J]. IEEE Journal on Selected Areas in Communications, 2014, 32(4): 684-692.

[8] SHAO F, LIN W, GU S, et al. Perceptual full-reference quality assessment of stereoscopic images by considering binocular visual characteristics[J]. IEEE Transactions on Image Processing, 2013, 22(5): 1940-1953.

[9] LIN Y H, WU J L. Quality assessment of stereoscopic 3D image compression by binocular integration behaviors[J]. IEEE Transactions on Image Processing, 2014, 23(4): 1527-1542.

[10] ZHANG Y, CHANDLER D M. 3D-MAD: A full reference stereoscopic image quality estimator based on binocular lightness and contrast perception[J]. IEEE Transactions on Image Processing, 2015, 24(11): 3810-3825.

[11] BATTISTI F, CARLI M, L E CALLET P, et al. Toward the assessment of quality of experience for asymmetric encoding in immersive media[J]. IEEE Transactions on Broadcasting, 2018, 64(2): 392-406.

[12] RUSANOVSKYY D, MÜLLER K, VETRO A. Common test conditions of 3DV core experiments[EB/OL]. (2013-08-02)[2019-07-06]. https://phenix.int-evry.fr/jct2/doc_end_user/current_document.php?id=1344.

[13] AVS Video Group. General test conditions of stereoscopic video coding[EB/OL]. (2010-12-16)[2019-07-06]. http://www.avs.org.cn/FileList.asp?meetingid=53&filetype=output.

[14] ITU-Recommendation. BT. 1788. Methodology for the subjective assessment of video quality in multimedia applications[S]. Geneva: ITU, 2007.

[15] ITU-Recommendation. BT. 500-11. Methodology for the subjective assessment of the quality of television pictures[S]. Geneva: ITU, 2012.

[16] ZHANG W, LIU H. Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications[J]. IEEE Transactions on Image Processing, 2017, 26(5): 2424-2437.

[17] ZHAO Y, CHEN Z, ZHU C, et al. Binocular just-noticeable-difference model for stereoscopic images[J]. IEEE Signal Processing Letters, 2010, 18(1): 19-22.

[18] SHAO F, JIANG G, WANG X, et al. Stereoscopic video coding with asymmetric luminance and chrominance qualities[J]. IEEE Transactions on Consumer Electronics, 2011, 56(4): 2460-2468.

[19] HUANG Y H, HUANG Y H, LIU W C. Quality-efficient upsampling method for asymmetric resolution stereoscopic video coding with interview motion compensation and error compensation[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2014, 24(3): 430-442.

[20] SAYGILI G, GURLER C G, TEKALP A M. Evaluation of asymmetric stereo video coding and rate scaling for adaptive 3D video streaming[J]. IEEE Transactions on Broadcasting, 2011, 57(2): 593-601.

[21] CHEN C, CHOI L K, DE VECIANA G, et al. Modeling the time-varying subjective quality of HTTP video streams with rate adaptations[J]. IEEE Transactions on Image Processing, 2014, 23(5): 2206-2221.

[22] ZHENGFANG D M, MA K D, WANG Z. Quality-of-experience for adaptive streaming videos: An expectation confirmation theory motivated approach[J]. IEEE Transactions on Image Processing, 2018, 27(12): 6135-6146.

[23] WANG Z, BOVIK A C, SHEIKH H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4): 600-612.

[24] ZHANG W, QU C, MA L, et al. Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network[J]. Pattern Recognition, 2016, 59: 176-187.

作者简介:

翟宇轩　男,硕士研究生。主要研究方向:视频传输、质量评价。

房颖　女,博士,讲师。主要研究方向:视频质量评价。

3D video quality evaluation based on adaptive streaming over HTTP

ZHAI Yuxuan, LIU Yisang, XU Yiwen, CHEN Zhonghui, FANG Ying*, ZHAO Tiesong

(College of Physics and Information Engineering, Fuzhou University, Fuzhou 350108, China)

Abstract: The key for 3D video network service is to improve the quality of experience (QoE) of users, which can be, however, affected by mutable network conditions and video contents. For conventional 2D videos, the HTTP adaptive streaming (HAS) technique has demonstrated its significance in improving user QoE by utilizing dynamically switched bitrates, while for 3D video transmission with at least two video streams, this technique has not yet been extensively explored. Dynamic conversion policy of the video quality level is the core of HAS technique. In this work, we investigate the impact on user QoE when introducing dynamic bitrates to different 3D videos. A subjective database is built to illustrate the connection between block-level objective quality, which changes with bitrates, and the QoE of 3D vision. Through this, we propose a convolutional neural network (CNN) based QoE model that effectively assesses the QoE by block-level objective quality. The Pearson linear correlation coefficient (PLCC) of the model predictive value and the mean opinion score (MOS) is 0.906. The proposed framework can provide guidance to inter-view bitrate balancing of HAS for 3D video transmission.

Keywords: 3D video; quality of experience; video quality assessment; convolutional neural network (CNN); HTTP adaptive streaming

Received: 2019-07-09; **Accepted**: 2019-08-14; **Published online**: 2019-08-22 11:24

URL: kns.cnki.net/kcms/detail/11.2625.V.20190822.1122.004.html

Foundation item: National Natural Science Foundation of China (61671152)

* **Corresponding author**. E-mail: fangying@fzu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0384

基于视频的三维人体姿态估计

杨彬[1,2], 李和平[3], 曾慧[1,2,*]

(1. 北京科技大学 自动化学院, 北京 100083; 2. 北京市工业波谱成像工程技术研究中心, 北京 100083;
3. 中国科学院自动化研究所, 北京 100190)

摘 要: 已有的三维人体姿态估计方法侧重于通过单帧图像来估计人体的三维姿态,忽略了视频中前后帧之间的相关性,因此,通过挖掘视频在时间维度上的信息可以进一步提高三维人体姿态估计的准确率。基于此,设计了一种可以充分提取视频时序信息的卷积神经网络结构,在获得高精度的同时也具有消耗计算资源小的优点,仅仅使用二维关节点的坐标为输入即可恢复完整的三维人体姿态。然后提出了一种新的损失函数利用相邻帧间人体姿态的连续性,来改进视频序列中三维姿态估计的平滑性,同时也解决了因缺少帧间信息而导致准确率下降的问题。通过在公开数据集 Human3.6M 上进行测试,实验结果表明本文方法相比目前的基准三维姿态估计算法的平均测试误差降低了 1.2 mm,对于视频序列的三维人体姿态估计有着较高的准确率。

关 键 词: 三维人体姿态; 卷积神经网络; 视频序列; 损失函数; 平滑
中图分类号: V221+.3; TB553
文献标识码: A **文章编号**: 1001-5965(2019)12-2463-07

人体姿态估计是指还原给定图片或者视频中人体关节点位置的过程,其对于描述人体姿态,预测人体行为起到至关重要的作用。近年来,随着深度学习技术的发展,人体姿态估计越来越广泛地运用到计算机视觉的各个领域之中,例如人机交互、行为识别以及智能监控等等。现如今,二维人体姿态估计算法的日渐成熟,三维的人体姿态估计开始受到更多研究者的关注,其在二维姿态估计的基础上加入了深度信息,这也进一步扩大了姿态估计的应用场景。早期的研究过多关注于利用人体的几何约束为主要特征来估计三维人体姿态[1-3],例如使用梯度方向直方图以及层次物体识别模型提取特征来对三维姿态进行预测,这种方法保证了输出结果的合理性,不过由于不同个体之间存在差异,往往难以获得精确的结果。当前的研究算法大多通过单幅 RGB 图像[4-8]以及利用已知二维姿态方法[9-15]来恢复人体的三维姿态,前者将姿态估计由回归问题转化为在离散空间中定位关节点位置的问题,取得了不错的效果,但其一定程度上会因遮挡等环境因素而导致检测性能下降。使用二维姿态恢复的方法则是寻找由二维关节点向三维空间的映射[16-17],这种方法相比其他方法更为直接,且最终的检测结果往往依赖于二维关节点坐标是否精确。

以上研究算法大多建立在对单帧图像进行分析的基础上,而现实生活中更多的数据源来自于视频输入,视频作为多帧连续图像的组合,包含了更为复杂的时序信息。而基于单帧图像进行估计一定程度上会导致相邻帧的检测结果存在巨大差异,因此,基于视频的三维人体姿态估计比单帧图

收稿日期: 2019-07-09; **录用日期**: 2019-08-19; **网络出版时间**: 2019-08-22 14:43
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190822.1308.005.html
基金项目: 国家自然科学基金(61973029); 中央高校基本科研业务费专项资金(FRF-BD-17-002A)
* **通信作者**. E-mail: hzeng@ustb.edu.cn

引用格式: 杨彬, 李和平, 曾慧. 基于视频的三维人体姿态估计[J]. 北京航空航天大学学报, 2019, 45(12): 2463-2469.
YANG B, LI H P, ZENG H. Three-dimensional human pose estimation based on video[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2463-2469 (in Chinese).

像检测具有更大的挑战。在时序分析领域,循环神经网络(Recurrent Neural Network,RNN)一直因其善于处理序列化数据而有着广泛地应用,英国著名的人工智能公司 DeepMind 于 2016 年提出的 WaveNet[18] 通用模型证明一维的卷积神经网络(Convolutional Neural Network,CNN)同样对序列化数据特征有着良好的提取能力,另外与 RNN 相比不容易受到梯度消失和爆炸的影响而且有着更为简单的网络结构。因此以一维卷积为基础设计深层网络来挖掘分析视频中的时序信息可能会具有更加突出的作用。

本文受到上述启发,构建了一种以视频中人体二维关节点坐标作为输入恢复得到三维人体姿态的算法,主要贡献概括如下:基于一维卷积对时序信息的提取能力,设计了一种高效网络,对视频中的三维人体姿态实现了准确的估计。深入研究视频相邻帧之间视觉信息的连续性,提出了一种新的损失函数,改进姿态估计结果的平滑性和有效性。最后在特定数据集上进行试验并对比分析,充分验证了本文方法对视频中的三维人体姿态估计的有效性,研究成果也为一些实际应用提供了技术支持。

1 三维姿态估计

1.1 方法概述

直觉上,二维关节点坐标向三维空间的映射可能会因缺少深度信息而导致错误姿态,不过 Martinez 等[16] 提出的基准方法证明了使用网络实现二维关节点恢复三维姿态是完全可行的,网络能够很好地依据关节相对位置来预测深度信息和连接关系。因此本文设计了一种以连续二维关节点坐标序列为输入恢复视频相关三维人体姿态的方法,如图 1 所示。二维关节点坐标直接由数据集的标注得到,除此之外,还可以通过将单帧图像送入二维姿态检测器得到人体二维关节点坐标,本文方法可以与目前许多高精度二维姿态检测器相结合,实现对于任意图像或视频输入,都能够准确恢复人体的三维姿态。之后对得到的二维关节点坐标序列进行归一化处理,加快网络收敛速度。最后将处理过的序列数据送入三维姿态估计网络,训练时网络会生成与序列数据相同数目的姿态,测试时本文只取中间一帧的姿态作为输出,因此输入二维关节点坐标序列的数目应为奇数。

1.2 三维姿态估计网络

三维姿态估计网络主要由 4 个具有相同结构的残差网络模块进行串联组成,除输入输出外,第

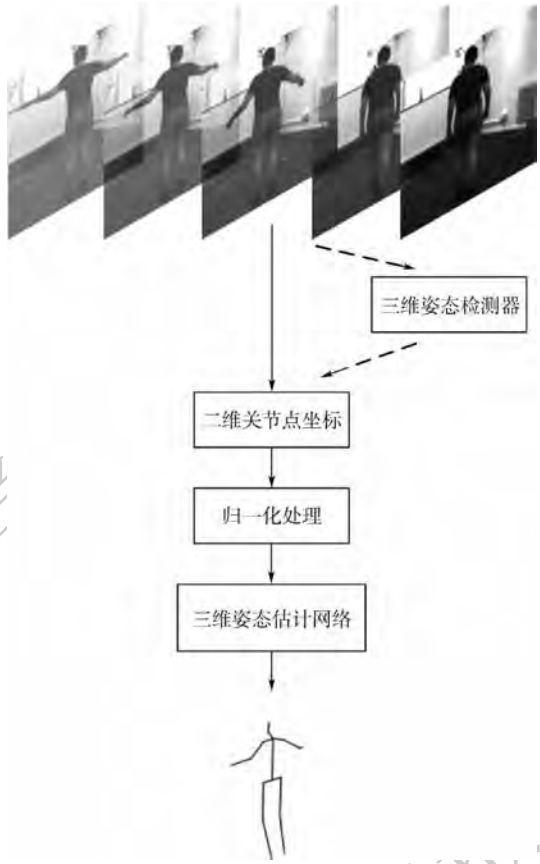


图 1 三维姿态生成过程

Fig. 1 Generation process of three-dimensional pose

一层的 3×1 卷积和最后一层的 1×1 卷积分别用于将输入维度进行扩展以增加网络宽度和将维度降至输出维度。残差网络模块由卷积层、Batch-Normal(BN)层、ReLU 层以及残差连接组成。

1)卷积层。残差网络模块在卷积层主要使用 3×1 的一维卷积和卷积核大小为 1×1 的点卷积,一维卷积通过卷积核的滑动来提取时间序列上的信息,点卷积主要用于改变特征的维度以此来对同维度的特征进行信息融合。

2)BN 层。神经网络各层的输出由于经过层内操作,其数据分布显然会与对应层的输入不同,并且差异会随着网络层的堆叠而逐渐增大,而 BN 层主要用于对每层的输入进行规范化,用于解决数据分布不均而导致的训练深层网络模型困难的问题。BN 层一定程度上起到了正则化的作用,使得训练过程中能够使用较高的学习速率,更加随意的对参数进行初始化,加快训练速度,提高网络的泛化性能。

3)ReLU 层。ReLU 层是一个非线性的激活单元,主要用于增加网络的非线性特征,其单侧抑制特性使得一部分神经元的输出为 0,增加稀疏性,减少了参数间的相互依存关系,缓解了过拟合问题的发生。

4）网络还借鉴了 ResNet[19] 网络结构中残差连接的思想，将输出表述为输入和输入的一个非线性变换的线性叠加，使得各个层级提取到的特征可以随意进行组合，保证特征在网络中的传递，三维姿态估计网络结构如图 2 所示。

对网络的设计不仅要求模型结构有着良好的性能，还要考虑实际应用中网络运行所需要的存储空间以及计算资源。网络模型的空间复杂度主要指的是参数的个数，其中 ReLU 层作为激活单元并没有需要学习的参数，单个 BatchNormal 层也仅有 2 个可以学习的参数，因此网络模型占用的空间大小近似等于所有卷积层的参数量之和，网络模型的时间复杂度主要通过浮点运算次数（FLoating-Point Operations，FLOPs）来衡量。使用连续 9 帧图像中人体关节点二维坐标为输入，计算不同数目的残差模块对于参数个数以及计算资源的消耗，并比较最终的测试误差。

由表 1 可得，在 4 个残差模块的使用下得到了最优结果，此后随着网络的进一步加深，出现了过拟合现象，平均测试误差开始增加，后续实验也采用 4 个残差模块的网络结构与其他方法进行对比分析。本文设计的轻量级网络模型实现了对三维人体姿态准确高效的估计，在有效减少参数的同时也具有极快的处理速度，能够更好地应用在各种硬件设备中。

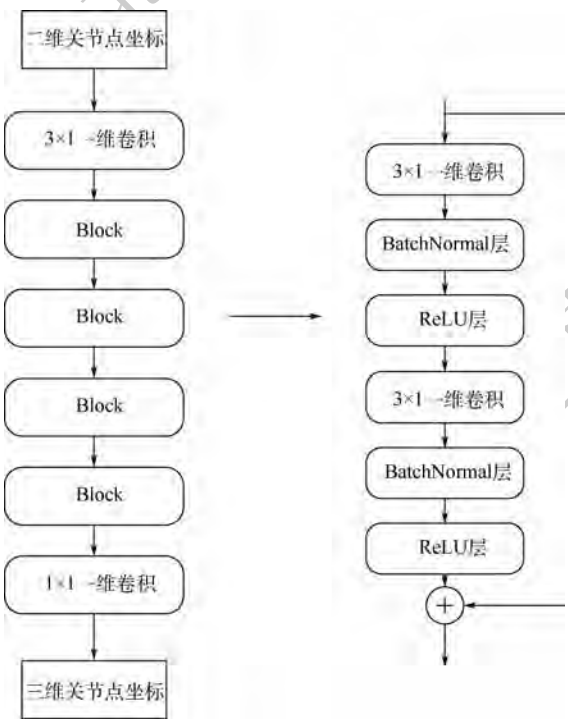


图 2　三维姿态重建网络结构

Fig. 2　Three-dimensional pose reconstruction network structure

表 1　网络模型参数量

Table 1　Parameter number of network model

残差模块数	浮点运算次数/百万	参数个数/百万	平均误差/mm
2	76.9	8.5	46.6
3	114.6	12.7	45.8
4	152.4	16.9	44.7
5	190.2	21.1	45.5

2　相似姿态位移约束损失函数

本文网络主要是利用已有的数据，取连续帧的二维关节点坐标作为输入，对人体关节点坐标从二维到三维空间的映射进行有监督学习并最终输出人体三维关节点的坐标，其本质上是一个回归问题。网络优化的目标是使得预测得到的三维关节点的坐标与真值之间的差值最小，因此首先定义姿态距离（Pose Distance，PD）的损失函数：

$$L_{\mathrm{PD}}=\frac{1}{T}\cdot\frac{1}{N}\sum_{t=1}^{T}\sum_{i=1}^{N}\left\|\boldsymbol{p}_i^t(\mathrm{pred})-\boldsymbol{p}_i^t(\mathrm{gt})\right\|_2^2 \quad (1)$$

式中：T 为同时输入网络连续帧关节点的数目；N 为人体关节点的数目，在实验中 $N=17$；$\|\cdot\|_2^2$ 表示 Euclidean 范数，通过使用预测值与真值的欧氏距离作为衡量关节点之间差异的标准；$\boldsymbol{p}_i^t(\mathrm{pred})$ 和 $\boldsymbol{p}_i^t(\mathrm{gt})$ 分别表示输入第 t 帧图像中第 i 个关节点的三维坐标预测值和真实值。

视频数据承载的信息不仅仅存在于单帧图像中，其更多的语义信息会通过连续帧来表达，而传统的视频姿态估计算法大多基于单帧图像，然后将结果整合为视频输出，无法充分利用视频的时空结构特性，往往存在输出不连续等问题。本文随机选取任意视频序列进行分析，并通过计算两个姿态间各个关节点之间的欧氏距离之和作为姿态差异，将实验结果取平均，根据图 3 可以得出同一视频段中姿态差异随序列增加近似呈线性增长，且相邻帧保持着微小的差异，通过网络训练来学习这一特性，可以使网络能够依据当前时刻的

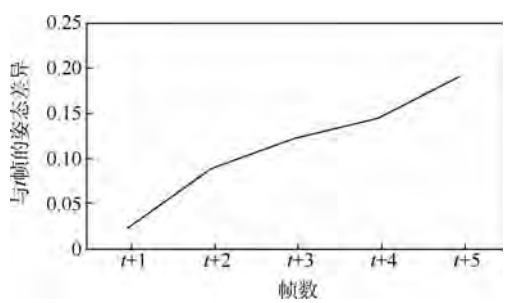


图 3　相邻帧姿态差异

Fig. 3　Pose difference between adjacent frames

输入预测下一时刻的输出,同时也保证后一帧的姿态预测结果与前一帧相比能够近似一致,以此来增加视频中姿态估计的准确性和平滑性。

基于上述分析,本文设计一种名为相似姿态位移约束(Similar Pose Displacement Constraint, SPDC)的损失函数来学习视频中的人体姿态在时间维度上的连续性,计算公式为

$$L_{\mathrm{SPDC}} = \frac{1}{T-1}\sum_{t=2}^{T}\sum_{i=1}^{N}\left(\left\|\boldsymbol{p}_i^t(\mathrm{pred}) - \boldsymbol{p}_i^{t-1}(\mathrm{pred})\right\|_2^2 - \left\|\boldsymbol{p}_i^t(\mathrm{gt}) - \boldsymbol{p}_i^{t-1}(\mathrm{gt})\right\|_2^2\right) \quad (2)$$

式中:$\boldsymbol{p}_i^t(\mathrm{pred}) - \boldsymbol{p}_i^{t-1}(\mathrm{pred})$表示预测得到的第 t 和 $t-1$ 帧间的姿态差异;$\boldsymbol{p}_i^t(\mathrm{gt}) - \boldsymbol{p}_i^{t-1}(\mathrm{gt})$则表示相邻帧间的真实姿态差异。最终的损失函数定义为上述两种损失的加权和,即

$$\mathrm{LOSS} = \alpha L_{\mathrm{PD}} + \lambda L_{\mathrm{SPDC}} \quad (3)$$

式中:α 和 λ 分别为姿态距离损失函数以及相似姿态位移约束损失函数的权重比,本文以姿态距离损失函数为主要的优化目标,使输出的每个关节点坐标值都尽可能地回归到真值附近,并辅以相似姿态位移约束损失函数来充分学习相邻帧的近似一致性,增加检测结果的平滑性。对 α 和 λ 的选取规则应该是 α 大于 λ,经过多次实验对比,本文选取 $\alpha=5$ 以及 $\lambda=1$ 作为最优的权重比,最终的损失函数为两种不同损失函数的加权和。

3 实验结果分析

3.1 实验数据集

为了对本文方法的性能进行评价,在三维人体姿态数据集 Human3.6M[20] 上进行了实验,Human3.6M 是目前为止最大也是使用最为广泛的三维人体姿态估计数据集,其主要由 7 位实验者在 4 个不同视角下使用高清摄像机精确捕捉的 360 万个三维人体姿态组成,视频的帧率为50 Hz,分辨率大小为 1 000 × 1 000。数据集被分割为 11 个子类别,其中 7 个类别包含了三维关节点标注,而且还使用相机参数对三维姿态的关节点数据进行投影,并获得准确的二维姿态信息,每个类别中都包括走路、打招呼等 15 个生活中常见动作。

3.2 实验细节和评价标准

实验过程中,使用 Human3.6M 提供的二维关节点坐标,选取某帧前后数目相等的二维关节点坐标序列作为输入,训练时为了保证视频起始端和末端完整性,对输入数据采取边缘填充操作,根据输入连续帧数目对起始帧和结束帧的二维关节点数据进行复制并填充。此外,本文还对输入的二维关节点坐标根据图像大小进行归一化处理。训练时采用 Adam 优化算法,初始学习率设置为 0.001,批处理大小为 1 024,权重衰减参数设为 0.000 65,对整个数据集迭代 50 次。

实验使用 NVIDIA GTX1060 显卡,64 位 Ubuntu 系统,Intel i7-6700 型号 CPU,并 Python3.5 环境配置下使用开源深度学习框架 Pytorch 对网络模型进行训练。使用平均关节位置误差(Mean PerJoint Position Error, MPJPE)作为评价标准,即计算网络预测得到的关节点坐标与真实标签 17 个人体关节点坐标之间欧氏距离的平均值。为了与其他实验方法进行公平比较,根据协议使用 Human3.6M 中的 S1、S5、S6、S7、S8 子数据集用于训练,S9、S11 数据集用于测试。

3.3 实验结果与分析

三维人体姿态估计结果如图 4 所示,每 25 帧连续图像的二维关节点坐标作为输入,采用 4 个残差模块网络结构的条件下,得到了最佳的实验结果。

为了充分评价本文方法在视频三维姿态估计上的有效性,首先将本文方法与目前的流行方法进行比较,在单个和平均动作误差上本文方法相比其他方法均有不同程度的提升,证明了基于一

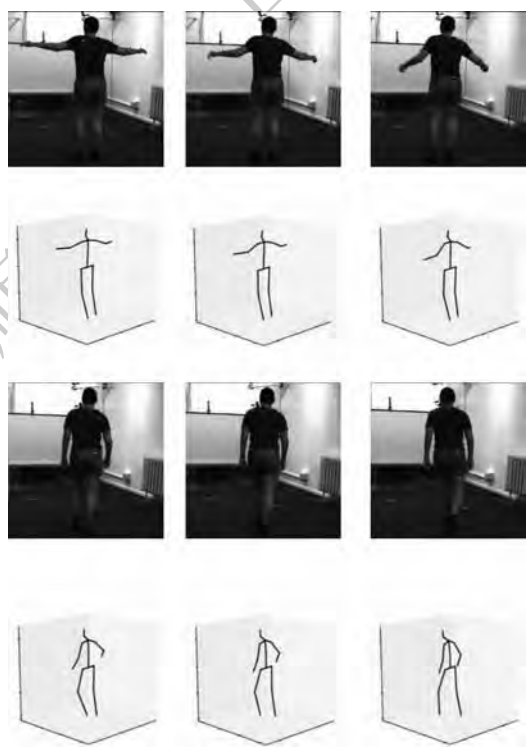


图 4 三维姿态估计结果

Fig. 4 Three-dimensional pose estimation results

维卷积网络结构预测三维人体姿态相比其他方法具有一定的竞争力。为了验证本文方法的鲁棒性,还使用当前较为先进的二维姿态检测器 CPN (Cascaded Pyramid Network)[21] 来对视频中的人体二维关节点坐标进行检测,然后利用所设计的网络得到最终的三维姿态,结果表明本文方法在未提供二维关节点标注的情况下与二维姿态检测器结合同样能得到较为准确的输出,实验结果如表 2 所示。另外本文还对两个不同方面进行了实验,一部分是采用不同数目连续帧的二维关节点坐标作为输入来探索序列长度对于视频三维姿态估计的影响;另一部分是验证本文的网络结构以及损失函数的改进对模型性能带来的提升。

对于不同数目连续帧输入的实验分析如图 5 所示,当输入序列长度大于 25 以后,模型的性能开始下降,平均误差开始增加,推测原因可能因为当前帧的检测结果只与相邻几帧呈高度相关性,其余帧的存在会带来更多的冗余信息。而且由于输入维度的增加,网络前向传播所需的时间也会成倍增加。

表 2　各种三维姿态误差

Table 2　Various three-dimensional pose errors

方法		姿态误差/mm						
		指路	讨论	吃饭	问候	打电话	照相	摆姿势
几何约束方法	文献[6]	54.8	60.7	58.2	71.4	62.0	65.5	53.9
单幅图像方法	文献[7]	58.6	64.6	63.7	62.4	66.9	70.7	57.7
	文献[15]	48.5	54.4	54.4	52.0	59.4	65.3	49.9
二维姿态	文献[12]	53.3	46.8	58.6	61.2	56.0	76.1	58.1
推断方法	文献[16]	52.8	54.8	54.2	54.3	61.8	67.2	53.1
	文献[1]	37.7	44.4	40.3	42.1	48.2	54.9	44.4
本文方法(真实值输入)		**35.4**	**43.0**	**37.9**	**40.0**	**44.4**	**52.1**	**41.7**
本文方法(CPN 检测器输入)		50.2	52.7	53.3	54.9	56.7	69.4	50.7

方法		姿态误差/mm							平均误差/mm
		购买	坐	坐下	抽烟	等待	遛狗	走路	散步	
几何约束方法	文献[6]	55.6	75.2	115.6	64.2	66.0	51.4	63.2	55.3	64.9
单幅图像方法	文献[7]	62.5	76.8	103.5	65.7	61.6	69.0	56.4	59.5	66.9
	文献[15]	52.9	65.8	71.1	56.6	52.9	60.9	44.7	47.8	56.2
二维姿态	文献[12]	48.9	55.6	73.4	60.3	62.2	61.9	35.8	51.1	57.5
推断方法	文献[16]	53.6	71.7	86.7	61.5	53.4	61.6	47.1	53.4	48.3
	文献[1]	42.1	54.6	58.0	45.1	46.4	47.6	36.4	40.4	45.5
本文方法(真实值输入)		**40.4**	**51.8**	**68.4**	**42.0**	**46.0**	**47.4**	**36.4**	**38.4**	**44.3**
本文方法(CPN 检测器输入)		51.2	66.6	83.2	56.4	53.9	61.3	44.9	49.2	57.0

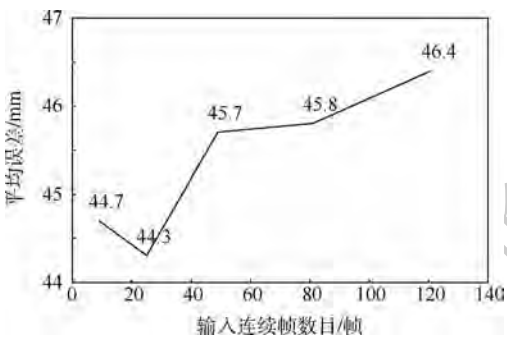


图 5　不同输入序列下的平均误差

Fig. 5　Average errors in different input sequences

接下来对本文设计的网络各个部分进行深入分析,表 3 给出了不同的网络设计对最终测试误差的影响。Dropout[22] 正则化是最简单的网络正则化方法:通过任意丢弃网络层中的参数来减少神经元之间复杂的共适应关系,迫使网络去学习更加鲁棒的特征,缓解过拟合的发生,起到正则化的作用。然而加入 Dropout 反而增加了大约 10 mm的误差,分析原因,可能由于 Dropout 随机删除卷积层参数,破坏了一维卷积提取时序信息的连续特征过程。与此同时,BN 层的加入减少了 14.9 mm 的测试误差,大幅提高了网络的泛化性能。另外,残差连接的设计也为本文的网络带来了 0.6 mm 误差的减小。

表 3　不同网络结构测试误差

Table 3　Testing errors of different network structures

网络结构	平均误差/mm	误差变化/mm
原始网络	**44.3**	
加入 Dropout(0.1)	54.8	+10.5
删除 BN 层	59.2	+14.9
删除残差连接	44.9	+0.6

最后分析本文所提出的损失函数对于网络性能的影响,具体方法为同时训练加入和不加入 SPDC 损失函数的网络,损失函数曲线如图 6 所

示。由图 6 可见,在训练初期,随着三维点坐标回归的逐渐精确,两个网络的相似姿态位移差异同时减小,但加入 SPDC 损失函数的网络下降幅度更大。在继续迭代的过程中,加入 SPDC 损失函数网络的相似姿态位移差异进一步减小且具有更小的震荡幅度,这说明 SPDC 损失函数的加入使得网络很好地学习了视频帧间的连续性,增加了视频三维姿态估计输出的平滑性,另外,SPDC 损失函数的加入最终减少了网络 0.8 mm 的误差,进一步提高了估计结果的准确性。

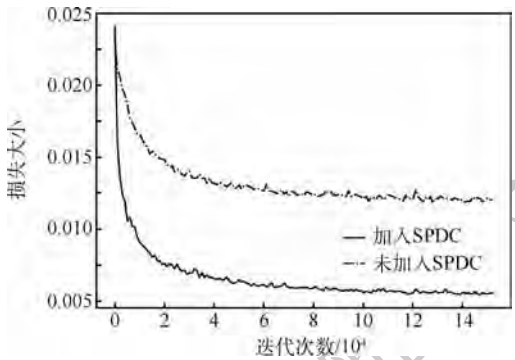


图 6 损失曲线对比

Fig. 6 Loss curves comparison

4 结 论

本文结合用于提取时序信息的一维卷积神经网络,提出了一种基于视频的三维人体姿态估计方法。研究结论如下:

1)本文方法能够以连续帧图像中人体二维关键点坐标作为输入,将已有的二维姿态准确地映射到三维空间中。

2)针对帧间信息缺失的情况,本文又设计了一种新的损失函数,对帧间的近似一致性进行学习,充分利用视频时间维度上的相关性来估计视频中的三维人体姿态。

3)实验表明,基于连续帧输入的姿态重建网络具有一定的合理性,并且本文方法可以与二维姿态检测器相结合,具有一定的鲁棒性。

下一步的主要研究工作是将本文方法与二维姿态估计任务相结合,设计通用的框架同时对二维和三维的人体姿态进行估计,并利用三维的姿态估计结果对二维的输出进行优化。

参考文献(References)

[1] BO L, SMINCHISESCU C. Twin gaussian processes for structured prediction[J]. International Journal of Computer Vision, 2010, 87(1-2): 28.

[2] RADWAN I, DHALL A, GOECKE R. Monocular image 3D human pose estimation under self-occlusion[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2013: 1888-1895.

[3] ZHOU X, HUANG Q, SUN X, et al. Towards 3D human pose estimation in the wild: A weakly supervised approach[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 398-407.

[4] PAVLAKOS G, ZHOU X, DERPANIS K G, et al. Coarse-to-fine volumetric prediction for single-image 3D human pose[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 7025-7034.

[5] NEWELL A, YANG K, DENG J. Stacked hourglass networks for human pose estimation[C]// Proceedings of the European Conference on Computer Vision. Berlin: Springer, 2016: 483-499.

[6] PAVLAKOS G, HU L, ZHOU X, et al. Learning to estimate 3D human pose and shape from a single color image[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 459-468.

[7] ROGEZ G, SCHMID C. Mocap-guided data augmentation for 3D pose estimation in the wild[C]// Advances in Neural Information Processing Systems, 2016: 3108-3116.

[8] VAROL G, ROMERO J, MARTIN X, et al. Learning from synthetic humans[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 109-117.

[9] CHEN C H, RAMANAN D. 3D human pose estimation = 2D pose estimation + matching[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 7035-7043.

[10] BOGO F, KANAZAWA A, LASSNER C, et al. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image[C]// European Conference on Computer Vision. Berlin: Springer, 2016: 561-578.

[11] PISHCHULIN L, INSAFUTDINOV E, TANG S, et al. Deepcut: Joint subset partition and labeling for multiperson pose estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 4929-4937.

[12] LOPER M, MAHMOOD N, ROMERO J, et al. SMPL: A skinned multi-person linear model[J]. ACM Transactions on Graphics (TOG), 2015, 34(6): 248.

[13] LUVIZON D C, PICARD D, TABIA H. 2D/3D pose estimation and action recognition using multitask deep learning[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 5137-5146.

[14] MORENO-NOGUER F. 3D human pose estimation from a single image via distance matrix regression[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 2823-2832.

[15] ZHOU X, ZHU M, LEONARDOS S, et al. Sparseness meets deepness: 3D human pose estimation from monocular video [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016:

4966-4975.

[16] MARTINEZ J, HOSSAIN R, ROMERO J, et al. A simple yet effective baseline for 3D human pose estimation[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 2640-2649.

[17] DROVER D, MV R, CHEN C H, et al. Can 3D pose be learned from 2D projections alone? [C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 78-94.

[18] OORD A, DIELEMAN S, ZEN H, et al. WaveNet: A generative model for raw audio[EB/OL]. (2016-09-19)[2019-06-13]. https://arxiv.org/abs/1609.03499.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 770-778.

[20] IONESCU C, PAPAVA D, OLARU V, et al. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 36(7): 1325-1339.

[21] CHEN Y, WANG Z, PENG Y, et al. Cascaded pyramid network for multi-person pose estimation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 7103-7112.

[22] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: A simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

作者简介:

杨彬　男,硕士研究生。主要研究方向:计算机视觉。

李和平　男,博士,副研究员。主要研究方向:模式识别、机器学习。

曾慧　女,博士,副教授,硕士生导师。主要研究方向:模式识别、图像处理。

Three-dimensional human pose estimation based on video

YANG Bin[1,2], LI Heping[3], ZENG Hui[1,2,*]

(1. School of Automation & Electrical Engineering, University of Science and Technology Beijing, Beijing 100083, China;
2. Beijing Engineering Research Center of Industrial Spectrum Imaging, Beijing 100083, China;
3. Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China)

Abstract: The existing 3D human pose estimation method focuses on estimating the 3D pose of the human body through a single frame image, while ignoring the correlation between the front and back frames in the video. Therefore, by investigating the information of the video in the time dimension, the accuracy of the 3D human pose estimation can be further improved. Based on this, the convolutional neural network structure that can fully extract the temporal information in the video is designed. It has the advantage of low computational resources and high precision. The complete 3D human pose can be restored only by using the coordinates of the 2D articulation point as input. Furthermore, a new loss function is proposed, which uses the continuity of human pose between adjacent frames to improve the smoothness of 3D pose estimation in video sequences, and also solves the problem of accuracy degradation due to lack of inter-frame information. By testing on the Human 3.6M dataset, the experimental results indicate that the average test error of the proposed method is 1.2 mm lower than that of the current standard 3D pose estimation algorithm, and the proposed method has a high accuracy for the 3D human pose estimation of video sequences.

Keywords: three-dimensional human pose; convolutional neural network; video sequences; loss function; smoothness

Received: 2019-07-09; **Accepted**: 2019-08-19; **Published online**: 2019-08-22 14:43

URL: kns.cnki.net/kcms/detail/11.2625.V.20190822.1308.005.html

Foundation items: National Natural Science Foundation of China (61973029); the Fundamental Research Funds for the Central Universities (FRF-BD-17-002A)

* **Corresponding author**. E-mail: hzeng@ustb.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0393

一种低成本的机器人室内可通行区域建模方法

张釜恺[1], 芮挺[2,*], 何雷[2], 杨成松[2]
(1. 中国人民解放军陆军工程大学 研究生院, 南京 210000; 2. 中国人民解放军陆军工程大学 野战工程学院, 南京 210000)

摘 要: 基于单目视觉的同步定位与建图(SLAM)是机器人领域中的一项热门技术。然而,在场景建图方面,由于其计算量较大,各主流方法还无法在低运算能力的平台上实现实时的场景建模。针对室内环境与小型机器人的特定情况,提出了一种新的可通行区域建模方法。该方法建立在单目特征点 SLAM 的基础上,通过 HSV 色彩空间内的图像自适应阈值分割获取地面分割图像,并与 SLAM 生成的稀疏点云进行交叉比对,进而获取地平面与准确的地面分割区域,再将地面分割区域反投影到地平面上,获取地面的稠密建模。在室内场景的实验中,所提方法的平均运算速度能达到 21 帧/s,速度约为 ORB-SLAM 的 70%,能够满足移动平台的实时性要求。对于地平面位置的还原平均误差为 5.8%,地面上道路宽度的建模误差在 3.5% ~12.8%。

关 键 词: 单目同步定位与建图; 室内场景; 可通行区域; 地面建模; 图像分割
中图分类号: TP242; TP37
文献标识码: A **文章编号**: 1001-5965(2019)12-2470-09

近年来,随着计算机、传感器以及图像处理技术[1]的进步,机器人技术也得到了蓬勃的发展。基于视觉的同步定位与建图(Simultaneous Localization and Mapping, SLAM)[2]技术是机器人领域的重要技术之一。SLAM 可以利用输入的连续图像来实时解算摄像机的运动轨迹,并且构建场景的地图用以指导机器人的行进。

室内环境是目前小型机器人的主要应用领域。在包括仓储物流、工厂物料搬运、家居服务等应用场景中,小型机器人都有很大的应用潜力。对于以室内为主要工作环境的小型机器人来说,单目摄像头视觉传感器由于较低的成本、较小的体积以及较高的布置灵活性,是机器人获取周围信息的主要来源。利用单目视觉信息获取场景信息以指导机器人的行动,是相关研究的重点方向之一。其中,地面可通行区域的信息,包含了地面的纹理与可通行地面的范围,可以直接用于对机器人的导航。地面可通行区域的建模可以为指导机器人的行动提供很大的便利。

然而,由于室内小型机器人的体积小、视角低、运算能力弱、视场有限,应用视觉 SLAM 方法来实时理解场景较为困难。能够构建稠密地图的 SLAM 方法对于计算资源有着较高的要求,不适合运算能力较弱的小型机器人使用。而且,由于其视野内的主要部分为地面且纹理特征较少,基于特征与梯度的 SLAM 方法只能获得极为稀疏的地图,对于场景的表现能力较差,无法指导机器人行动。但另一方面,绝大多数室内地面均为平面,而且纹理分布较为均匀,这又为地面的建模提供了方便。

针对此问题,本文提出了一种融合图像分割与单目 SLAM 的室内地面快速建模方法来获取机器人的可通行区域。该方法基于单目特征点 SLAM,并结合图像分割,利用 SLAM 生成的点云

收稿日期: 2019-07-16; **录用日期**: 2019-08-18; **网络出版时间**: 2019-08-22 14:43
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190822.1027.001.html
* **通信作者**. E-mail: rtinguu@sohu.com

引用格式: 张釜恺, 芮挺, 何雷, 等. 一种低成本的机器人室内可通行区域建模方法[J]. 北京航空航天大学学报, 2019, 45(12): 2470-2478. ZHANG F K, RUI T, HE L, et al. A low-cost indoor passable area modeling method for robots[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2470-2478 (in Chinese).

与图像分割之间的相互印证来提高地面构建的准确性。最终,生成地面可通行区域的稠密建模,用以指导机器人的行动。

通过实验与对比,在小型移动机器人平台上,相对于主流 SLAM 方案,本文方法可以显著提高对于地面建模的效果,并且满足在移动平台上实时处理的要求。相比较于经典的单目 SLAM 方法,本文方法的运行速度接近特征点 SLAM,而对地面可通行区域的建模效果则接近稠密,对于机器人的行动有较大的参考价值。

1　相关工作

SLAM 问题自 1988 年被提出,经过近 30 年的发展,取得了丰硕的成果。单目 SLAM 目前主要有 2 条技术路线。直接法 SLAM 直接利用像素的亮度或梯度来匹配图像间的像素,并可以在解算相机运动的同时生成稠密[3-4]或者半稠密[5-6]的场景地图。特征法 SLAM 利用点[7-8]或者线特征[9]来匹配图像间的点或者线,进而解算相机运动并生成只包括特征点或线的场景地图。

在建模方面,直接法生成的稠密或半稠密地图相对于特征法生成的稀疏地图有着更多的点,对于场景的描述也更加具有表现力。其中,代表性的系统有 2011 年的 DTAM(Dense Tracking and Mapping)[10]、2014 年的 LSD-SLAM(Large-Scale Direct monocular SLAM)[11]等。然而,虽然基于特征的 SLAM 系统只能获得稀疏的地图,但是一般来说具有更快的速度。而且由于便于使用束集调整(Bundle Adjustment,BA[12])进行优化,并进行回环检测[13],特征法获取的轨迹一般来说更加准确与鲁棒。在特征法 SLAM 中,2007 年的 MonoSLAM 系统[14]与同一年的 PTAM(Parallel Tracking and Mapping)[15]是这一领域具有开创性的成果。2015 年的 ORB-SLAM(ORiented Brief SLAM)[16]系统各项功能较为完备,已经被广大研究者用作后续二次开发研究的基础系统。

为了在保持稳定跟踪的同时,获得更加稠密的地图,有一些方法在特征法的基础上进行场景重建。2010 年,Newcombe 和 Davison[17]提出的方法是在获取了相机位姿与稀疏点云的基础上,利用光流来重建场景表面。2015 年,Murartal 和 Tardos 基于 ORB-SLAM 系统,利用逆深度滤波像素匹配,构建了基于概率的半稠密场景地图[18]。一般来说,此类方法的基本思想为基于 SLAM 获得的相机轨迹对图像像素进行匹配,并计算其空间位置。此类思路虽然可以结合直接法和特征点法的优点,但是运算量也大大提升,并不符合小型移动平台的运算能力。

为了提高运算的速度,以适应小型移动平台较低的运算能力,近年来,基于视觉的低成本建模方法也成为了一个有意义的研究方向。2016 年,Hinzmann 等[19]提出了一种针对无人机平台的低成本地面建模方法,利用无人机自身的传感器感知姿态,并结合图像处理方法获得对于地面的稠密建模。但是,该方法基于的图像内容全部为地面,不适用于小型地面无人车。2017 年,蒙山和唐文名将直线引入点特征单目 SLAM,并通过点线特征以及直线增强的 J-Linkage 算法构建包含特征平面的场景地图[20],但是该地图相对于纯特征点地图,对场景描述效果的提升有限。2017 年,von Stumberg 等[21]针对小型旋翼无人机,基于 LSD-SLAM 提出了一种本地探索避障策略与体素模型构建方法,但是对于地面的建模效果较差。2018 年,蒋林等[22]对基于嵌入式平台三维重建算法进行了研究,针对差动式机器人的运动特点对算法进行了针对性的简化,但是所利用的视觉传感器为深度相机而非单目相机。

总体来说,现有的主流单目 SLAM 系统在小型室内机器人上运行时,由于平台自身的特点,或是地图构建过于稀疏;或是虽地图稠密,但不能实现实时运行。一般并不能在构建场景方面取得令人满意的效果。而现有的低成本建模方法中,针对小型地面机器人特点的研究不多,尚未出现较为合适的地面区域低成本建模方法。

2　室内地面快速建模方法

2.1　整体流程

本文提出的融合图像分割与单目 SLAM 的室内地面快速建模方法构建在单目特征点 SLAM 的基础上,利用 SLAM 系统获得的相机位姿与稀疏点云,结合地面区域的分割图像获得地面的稠密建模。

系统的基本流程如下:首先,利用 SLAM 系统获得的相机位姿与稀疏点云。其次,利用地面区域的分割图像,将稀疏点云中投影在地面区域的点提取出来,获取粗略的地面稀疏点云。然后,进行滤波后,利用优化模型对这些地面点拟合平面,获取地平面的数学表达。之后,通过计算稀疏点云中的点与该地平面之间的位置关系,将它们相对精确地分为地面与非地面点云。再利用该分类来筛选地面分割图像。最后,通过相机的位姿,将筛选过的地面分割图像投影到地面平面上,并进

行滤波与降噪,获得稠密的地面建模。整个系统的流程示意图如图 1 所示。

后文将对本文方法流程中的各主要功能模块逐一进行介绍。

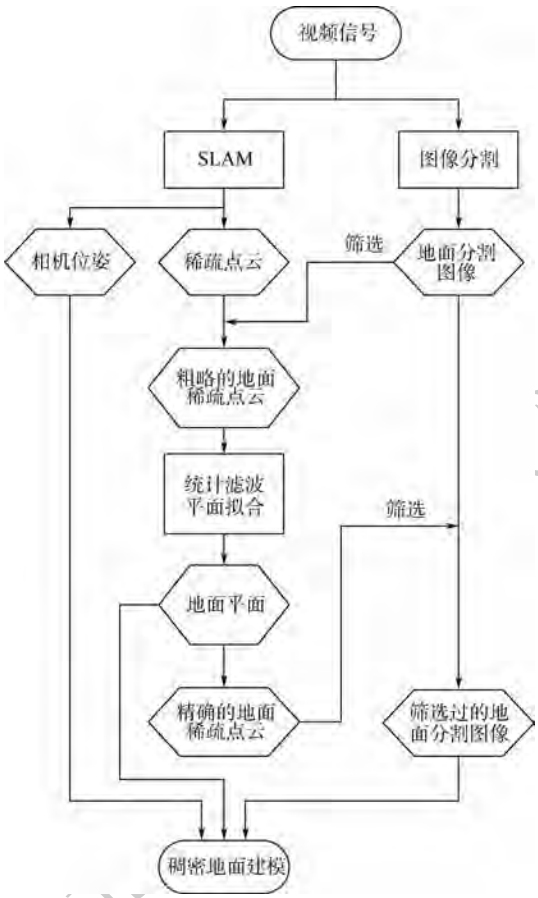


图 1 本文方法整体流程图

Fig. 1 Flowchart of proposed method

2.2 单目 SLAM 模块

本文使用 ORB-SLAM[16] 作为系统的单目 SLAM 模块。ORB-SLAM 是一个功能完善,接口便捷的单目 SLAM 系统。整个系统围绕 ORB 点特征建立,在效率与精度之间达到了较好的平衡,自发表以来就被研究者们所广泛使用。该系统包含 3 个平行的线程:跟踪线程(tracking thread)、局部建图线程(local mapping thread)和回环检测线程(loop closing thread)。其中,跟踪线程负责计算相机的运动,以及指定关键帧,并在跟踪丢失时负责重定位;局部建图线程主要负责用局部束集调整(local BA)来对局部地图进行优化,以及地图点的插入和删除;回环检测线程主要负责进行回环检测,在有回环出现时进行闭环全局优化。

2.3 地面区域的分割

在 SLAM 进程对相机位姿进行跟踪的同时,系统对每一帧输入图像进行分割,以获得地面区域的分割图像。由于系统获得的是一个地面分割的图像序列,并且还需要和点云数据进行比对,所以对于单张地面分割图像的精度要求不高,但是对于图像分割的速度有着较高的要求,这使得采用目前主流的深度学习语义分割的方法[23]较为困难。基于以上分析,本文采用基于 HSV(Hue, Saturation, Value)色彩空间的自适应阈值分割方法来对地面区域进行分割。

计算机存储和处理图片时一般采用的颜色编码方式是 RGB 颜色空间,其将颜色按红、绿、蓝 3 个维度上的分量进行编码,而 HSV 颜色空间是按照色度、饱和度和亮度来进行编码,分别对应 H、S、V 3 个分量,更贴近人眼的视觉感知,受光照的影响也更小[24]。在本文的分割方法中,先将待分割的图像转化到 HSV 色彩空间。

由于室内小型机器人在运动过程中会与障碍物保持一定距离,所以其摄像头拍摄图像的中间底部一般为地面区域。基于此假设,对于输入图像 I,本系统设定图像的底部中间区域为默认地面区域 M_G,其位置与尺寸如图 2(a)所示。首先,统计得到默认地面区域 M_G 内 H、S、V 3 个分量的最值。然后对全图像的每个像素 p_i 进行判断:

$$\begin{cases} \min[H_G] \leqslant h_i \leqslant \max[H_G] \\ \min[S_G] \leqslant s_i \leqslant \max[S_G] \\ \min[V_G] \leqslant v_i \leqslant \max[V_G] \end{cases} \quad p_i \in G_1 \tag{1}$$

式中:H_G、S_G、V_G 分别为默认地面区域 M_G 中所有像素 H、S、V 分量的集合;h_i、s_i、v_i 分别为像素 p_i 的 H、S、V 分量。若像素 p_i 的 3 个分量均在默认地面区域 M_G 的分量范围内,该像素就被加入候选地面区域 G_1,如图 2(b)所示。

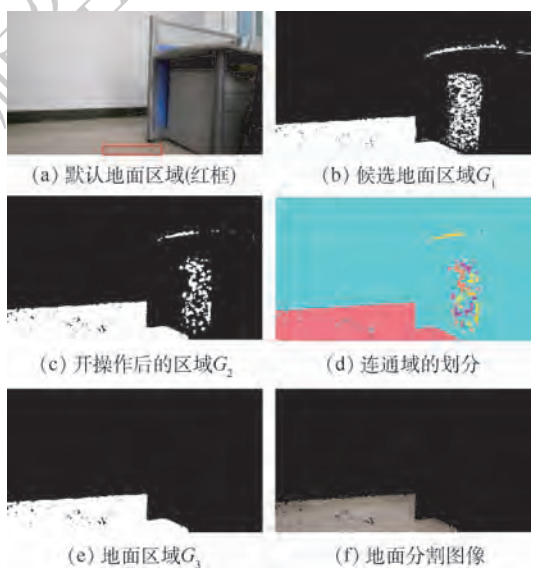


图 2 地面区域的分割示例

Fig. 2 Example of ground area segmentation

在获取候选地面区域之后,还需要进行后续的滤波与处理。首先进行开操作(采用运算符号“○”表示)。开操作用于使对象的轮廓变得光滑,断开狭窄的间断和消除细的突出物。开操作的流程是先用结构元素 S 对 G_1 腐蚀(采用运算符号“⊕”表示),然后再用 S 对结果进行膨胀(采用运算符号“⊖”表示)[25]。将候选地面区域 G_1 用开操作处理后,得到区域 G_2:

$$G_2 = G_1 \circ S = (G_1 \ominus S) \oplus S \tag{2}$$

区域 G_2 如图 2(c)所示。其中,本文方法采用的结构元素 S 为半径为 2 像素的圆形结构。

最后对区域进行连通域划分,划分结果如图 2(d)所示。选取默认地面区域 M_C 所在的连通域作为分割得到的地面区域 G_3,如图 2(e)所示。获得的地面分割图像如图 2(f)所示。

2.4　地面点云的获取、滤波与地平面的计算

在 SLAM 模块向稀疏点云地图中添加每一个空间地图点 Q_i 时,都会判断该点在其对应的地面分割图像上是否位于地面区域。如果位于地面区域,则将该点加入地面点云 C_g。

然后,利用该地面点云来获得地平面的数学表达。由于点云位置计算误差以及地面分割误差的存在,在后续处理之前,需要先去除地面点云中明显的误差离群点。本文采用统计滤波的方法进行离群点的去除[26]:

1) 计算点 Q_i 与其邻域范围内 k 个邻域点的平均距离。

2) 分析所有点 k 邻域平均距离,计算其均值 m 和标准差 σ。

3) 设置标准差倍数阈值 std,确定距离阈值为

$$d = m + \sigma \cdot \text{std} \tag{3}$$

4) 若点 Q_i 的 k 邻域平均距离大于距离阈值 d,则为噪声点,去除。

本文方法的应用环境是地面为平面的室内环境。建立优化模型对经过统计滤波的地面点云 C_g' 进行平面拟合,使点到平面的总距离最小。设地平面 F 的表达式为

$$Ax + By + Cz + D = 0 \tag{4}$$

优化模型为

$$\min_{A,B,C,D} \sum_{i=0} \frac{|Ax_i + By_i + Cz_i + D|}{\sqrt{A^2 + B^2 + C^2}}$$
$$\text{s.t.} \quad Q_i \in C_g', Q_i = (x_i, y_i, z_i) \tag{5}$$

2.5　对地面分割图像的筛选

在获取了地平面 F 的表达式后,系统对地面分割图像进行筛选。对于每一个关键帧,系统计算其观察到的所有地图点与地平面 F 的位置关系,并将所有位于地平面 F 上方 μ(单位为 cm)以上的点视为非地面点,其余视为地面点。

若某张地面分割图像的地面区域上有超过 n 个非地面点的投影,则略去这张分割。在本文的实验中,略去的分割如图 3 所示(其中红色为非地面点,绿色为地面点)。

从图 3 可以看出,将墙体下沿的部分瓷砖也划入了地面分割,从而包括了大量非地面点,并因此被略去。

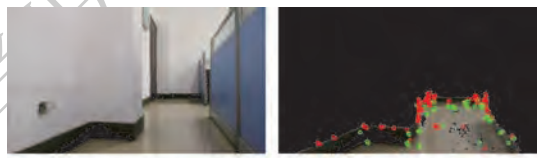


图 3　被略去的地面分割图像示例

Fig. 3　An example of excluded ground segmentation image

2.6　稠密地面建模

至此,系统通过对稀疏点云与地面分割图像的交叉比对,已经获得了相对精确的地平面与分割图像。利用关键帧的位姿,将地面分割图像投影到地平面上即可得到稠密的地面点云模型,如图 4 所示。

首先,通过相机内参矩阵 $\boldsymbol{K}$,将地面分割图像上的像素 p_i 从图像坐标 $[u,v]^{\mathrm{T}}$ 反投影到相机坐标系中,获得相机坐标系下的坐标 $\boldsymbol{P}_C = [X_c, Y_c, Z_c]^{\mathrm{T}}$:

$$\boldsymbol{P}_C = \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} = \boldsymbol{Z}_c \boldsymbol{K}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \tag{6}$$

式中:深度值 Z_C 为未知量。

然后,通过该帧的相机位姿变换矩阵 $\boldsymbol{T}_{wc}$,将相机坐标系下的坐标转换到世界坐标系坐标 $\boldsymbol{P}_w$,即

$$\boldsymbol{P}_w = \boldsymbol{T}_{wc} \boldsymbol{P}_c \tag{7}$$

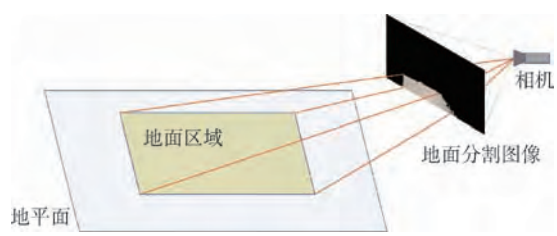


图 4　将分割图像投影到地平面对地面建模

Fig. 4　Projecting segmented images to ground plane for ground modeling

该世界坐标(x_w,y_w,z_w)应该在平面 F 上,即

$$Ax_w + By_w + Cz_w + D = 0 \tag{8}$$

利用式(8)解出未知深度 Z_C,即可以得到该像素点对应的世界坐标。对所有地面分割图像如此操作后,再进行体素滤波,设定滤波半径为 0.01,即半径 0.01 的范围内只保留一个点。最终,得到地面的稠密点云模型。

3 实验与分析

为了验证本文方法的可行性与实用性,使用搭载了单目摄像头的小型机器人获取数据。数据的处理使用笔记本电脑,搭载 Ubuntu 16.04 系统,处理器(CPU)为 Intel 4 200 H,内存 4 G。并同样的环境中用同样的平台运行了 ORB-SLAM 系统,与本文的系统相对比。

分别以办公室与教室作为实验环境。其中,办公室的地面为瓷砖,颜色与纹理相对连续,但是如图 5(a)所示,地面存在很严重的倒影现象。教室场景较为简单,光照较为均匀,但是如图 5(b)所示,教室的地面上存在明显的黑色地砖缝,地面色彩与纹理不连续。2 个实验环境对于地面区域的建模均具有一定的挑战。

经过预实验,在实验中将点云的统计滤波参数设置为 $k = 10$,std = 1,将地面分割图像筛选过程中的参数设置为 $\mu = 3$ cm,$n = 50$。

(a) 办公室 (b) 教室

图 5 实验环境

Fig. 5 Experimental environment

3.1 场景与建模效果

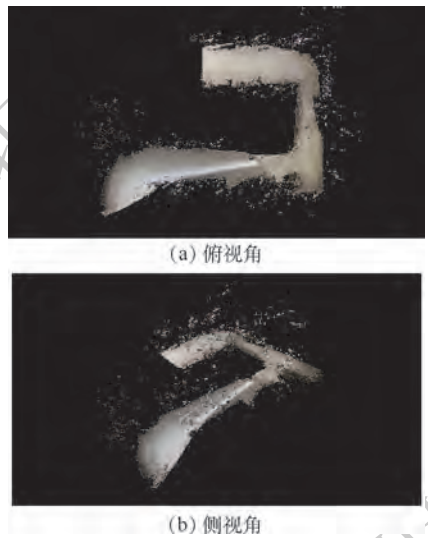
3.1.1 办公室场景

办公室环境和机器人的运动轨迹如图 6 所示。本文方法得到的稠密地面建模与非地面稀疏点云如图 7 所示。ORB-SLAM 得到的场景稀疏点云如图 8 所示。

图 6 办公室环境与机器人运动轨迹

Fig. 6 Office environment and robot movement track

办公室场景的地面可通行区域边界本身就并不明晰,所以建模的边缘并不规则。但是本文方法得到的模型能够基本准确地反映通路的宽窄与通过情况,例如中间一段通路中最细的部分反映了因墙面突起造成的障碍。

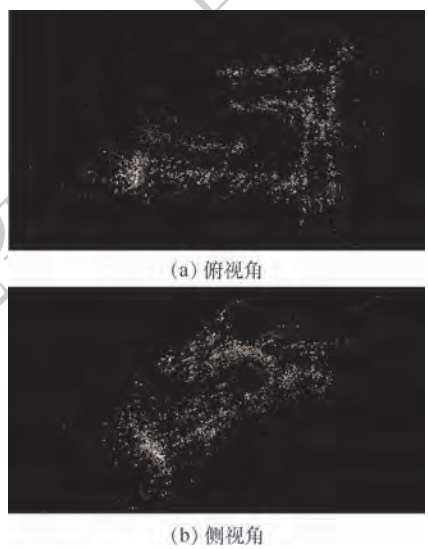


(a) 俯视角

(b) 侧视角

图 7 本文方法得到的办公室稠密地面建模与非地面稀疏点云

Fig. 7 Office dense ground modeling and non-ground sparse point cloud obtained by proposed method



(a) 俯视角

(b) 侧视角

图 8 ORB-SLAM 得到的办公室稀疏点云

Fig. 8 Office sparse point cloud obtained by ORB-SLAM

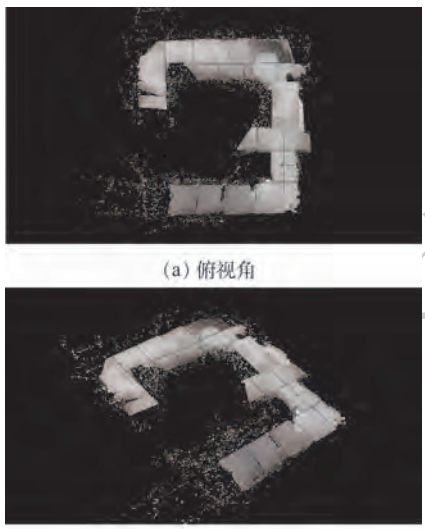
3.1.2 教室场景

教室环境和机器人的运动轨迹如图 9 所示。本文方法得到的稠密地面建模与非地面稀疏点云如图 10 所示。ORB-SLAM 得到的场景稀疏点云如图 11 所示。

教室场景的地面比较规则,本文方法所建的

图 9　教室环境与机器人运动轨迹

Fig. 9　Classroom environment and robot movement track

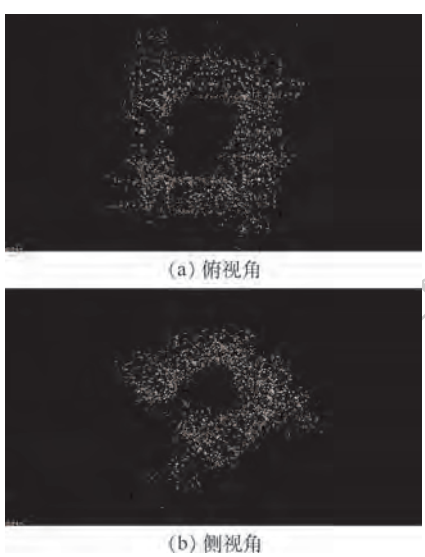


(a) 俯视角

(b) 侧视角

图 10　本文方法得到的教室稠密地面建模与非地面稀疏点云

Fig. 10　Classroom dense ground modeling and non-ground sparse point cloud obtained by proposed method



(a) 俯视角

(b) 侧视角

图 11　ORB-SLAM 得到的教室稀疏点云

Fig. 11　Classroom sparse point cloud obtained by ORB-SLAM

模型较好地还原了通行区域的地面情况。但是由于通路狭窄,在转弯的位置相机面对旁边的课桌,地面的图像区域不足,造成转弯部分建模效果稍差。

3.1.3　建模效果

从图 11 中可以看出,原 ORB-SLAM 生成的点云十分稀疏,只是标记了特征点的位置,根本无法从中获得地面的位置以及场景的有效信息,对场景的还原效果较差,基本不能用于对场景的描述,指导机器人行动。而本文方法获得了对于地平面的稠密建模,基本涵盖了机器人视野范围内的可通行区域。相比较原 ORB-SLAM 算法,本文方法生成的地图更具有实用性。

3.2　运算速度

在实际的运行中,本文方法与 ORB-SLAM 算法平均每一帧各阶段的运算时间如表 1 所示。

从表 1 数据可以看出,相比较于 ORB-SLAM,本文方法中的 SLAM 运算时间基本不变。增加的计算时间中,占比最大的部分是图像的分割与地面稠密建模。由于本文方法添加的操作全部是基于关键帧的,所以实际平均到每一帧的时间都很短。总体来看,本文方法能够达到 21 帧/s 的平均运行速度,ORB-SLAM 的平均运行速度约为 30 帧/s。本文方法约为 ORB-SLAM 的 70%,能够满足预期的应用要求。

表 1　平均每帧运算时间

Table 1　Average operation time per frame

模块	运算时间/ms	
	本文方法	ORB-SLAM
SLAM 进程	31.21	32.30
图像分割	5.16	—
地面点云获取	1.54	—
拟合平面	0.13	—
分割图像筛选	1.84	—
地面稠密建模	7.80	—
总计	47.68	32.30

3.3　精度分析

由于单目 SLAM 方法生成的空间数据具有尺度不确定性,所以为分析本文方法的建模精度,首先将生成的相机轨迹与真实的轨迹进行配准,得到数据的正确尺度后,再与实际数据相对比,进行精度分析。而 ORB-SLAM 得到的稀疏点云地图则过于稀疏,基本上无法从中提取通路的位置与具体范围,精度也无从谈起,故不对其进行精度分析。

3.3.1　地平面的位置还原精度

本文方法对地面进行还原的重要步骤是拟合地平面。由于搭载的相机高度已知,通过对比相机真实高度与相机和拟合地平面之间的平均距离,可以得到地平面的位置还原精度,如表 2 所示。可

见,地平面位置的还原平均绝对误差为 5.8%。

3.3.2 场景建模精度

在 2 个场景中,分别在系统生成的稠密地面模型上选取 3 处通路的宽度进行测量,与实际地面上通路的宽度进行对比,以分析方法的场景建模精度。由于转弯处存在遮挡,相机拍不到通路的全貌,故选取直线前行的部分通路作为实验分析的对象。

1) 办公室

对地面上 3 段通路的宽度进行测量(见图 12)。

与真实的宽度进行对比,得到的结果如表 3 所示。

通过对地面 3 处通路的宽度进行测量后,我们发现 a 处和 b 处精度较高,而 c 处误差较大。经分析,这部分的误差较大主要是由于地面倒影的影响,分析详见后文 3.4.1 节。

2) 教室

对系统生成的稠密地面模型进行测量,对地面上 3 段通路的宽度进行测量(见图 13)。

表 2 地平面位置还原精度

Table 2 Accuracy of ground plane position restoration

场景	实际距离/cm	拟合地平面距离/cm	误差/%
办公室	42.0	43.6	3.81
教室	42.0	38.7	-7.86

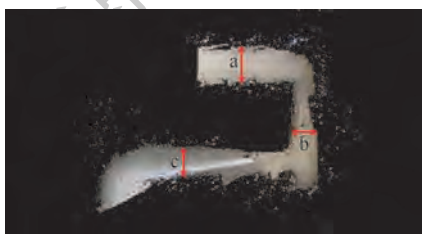


图 12 办公室通路宽度测量

Fig. 12 Path width measurement of office

表 3 办公室地面建模精度

Table 3 Ground modeling accuracy of office

位置	实际距离/cm	模型距离/cm	误差/%
a	135	129	4.4
b	112	108	3.5
c	128	116	10.3

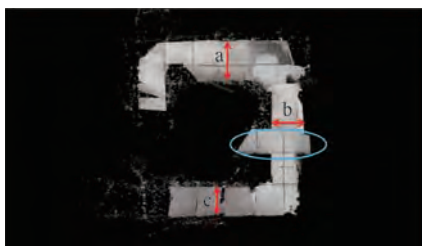


图 13 教室通路宽度测量

Fig. 13 Path width measurement of classroom

与真实的宽度进行对比,得到的结果如表 4 所示(b 处下方蓝圈内为两条通路交叉的位置,并非通路宽度)。

通过对地面 3 处通路的宽度进行测量后,发现 a 处和 b 处精度较高,而 c 处相对误差较大。经分析,该部分误差主要由地面纹理不连续造成,分析详见后文 3.4.2 节。

表 4 教室地面建模精度

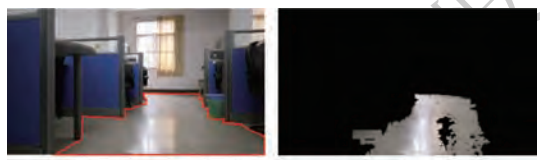
Table 4 Ground modeling accuracy of classroom

位置	实际距离/cm	模型距离/cm	误差/%
a	82	85	3.6
b	77	81	5.2
c	70	61	12.8

3.4 外在因素的影响分析

3.4.1 地面倒影

在图 12 办公室建模的 c 处,这一区域摄像机面向窗户,由于室外亮度较大,在地面瓷砖上产生了倒影,使得地面中间区域亮度较高,地面的颜色出现了较大区别。对于地面划分出现了一定的误差,如图 14 所示。



(a) 原图与真实地面区域 (b) 分割图

图 14 地面倒影对图像分割的影响

Fig. 14 Effect of ground reflection on image segmentation

3.4.2 地面不均匀纹理

地面的不均匀纹理对于本文方法存在一定的影响。在教室中,地面存在明显的地砖缝,导致有少部分图像对于地面区域的分割不连续,经过连通域计算后只留下一块地砖内的地面,如图 15 所示。

根据对于重建效果的观察和重建精度的分析,虽然存在倒影与地面纹理不均匀造成的图像分割误差,但是它们对于整体的地面重建效果影响较为有限。由于本文方法利用了图像的序列整体信息,对于部分图像的信息缺失,可以通过序

(a) 原图与真实地面区域 (b) 分割图

图 15 地面不均匀纹理对图像分割的影响

Fig. 15 Effect of uneven texture of ground on image segmentation

列中其他相对正确的分割投影所掩盖。

4　结　论

关于场景的建模,从场景中获取有意义的信息。速度与效果间的矛盾一直是研究中一个绕不过去的问题。近年来,国内外有许多研究试图在二者之间达到合适的平衡。本文结论如下:

1)对于室内小型机器人这一特定应用平台,以及室内环境这一特定场景,很多问题可以得到简化。

2)本文方法最大限度地利用了特定环境与特定需求,将对全场景的建模简化为地面的建模,将图片深度的求取简化为直接投影在地平面上,可以用尽可能小的计算量来为机器人实时提供尽可能多的信息。

3)实验证明,本文方法可以以接近特征点 SLAM 的速度,生成较为准确的地面可通行区域稠密建模。点云与分割的交叉验证以及图像序列的使用也可以在相当程度上抵消地面倒影、地面分割困难等不利情况的影响。

参考文献(References)

[1] LI Z, TANG J. Weakly supervised deep metric learning for community-contributed image retrieval[J]. IEEE Transactions on Multimedia, 2015, 17(11): 1989-1999.

[2] LIU H, ZHANG G, BAO H. A survey of monocular simultaneous localization and mapping[J]. Journal of Computer-Aided Design & Computer Graphics, 2016, 28(6): 855-868.

[3] TURAN M, ALMALIOGLU Y, ARAUJO H, et al. A non-rigid map fusion-based direct SLAM method for endoscopic capsule robots[J]. International Journal of Intelligent Robotics and Applications, 2017, 1(4): 399-409.

[4] JAN S, GUMHOLD S, CREMERS D. Real-time dense geometry from a handheld camera[C]// Proceedings of the Pattern Recognition-32nd DAGM Symposium. Berlin: Springer, 2010: 22-24.

[5] ENGEL J, KOLTUN V, CREMERS D. Direct sparse odometry [J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2016, 40(3): 611-625.

[6] KIM J H, CADENA C, REID I. Direct semi-dense SLAM for rolling shutter cameras[C]// Proceedings of the IEEE International Conference on Robotics and Automation. Piscataway, NJ: IEEE Press, 2016: 1308-1315.

[7] BAILEY T, NIETO J, GUIVANT J, et al. Consistency of the EKF-SLAM algorithm[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE Press, 2006: 9-15.

[8] PIRKER K, MATTHIAS R, BISCHOF H. CD-SLAM-Continuous localization and mapping in a dynamic world[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway, NJ: IEEE Press, 2011: 25-30.

[9] GEE A P, MAYOL-CUEVAS W. Real-time model-based SLAM using line segments[C]// Proceedings of the International Symposium on Visual Computing. Berlin: Springer, 2006: 354-363.

[10] NEWCOMBE R A, LOVEGROVE S J, DAVISON A J. DTAM: Dense tracking and mapping in real-time[C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE Press, 2011: 6-13.

[11] ENGEL J, SCHÖPS T, CREMERS D. LSD-SLAM: Large-scale direct monocular SLAM[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2014: 834-849.

[12] TRIGGS B, MCLAUCHLAN P F, HARTLEY R I, et al. Bundle adjustment a modern synthesis[C]// Proceedings of the Vision Algorithms: Theory and Practice. Berlin: Springer, 2000: 298-372.

[13] ULRICH I, NOURBAKHSH I R. Appearance-based place recognition for topological localization[C]// Proceedings of the International Conference on Robotics and Automation. Piscataway, NJ: IEEE Press, 2000: 1023-1029.

[14] DAVISON A J, REID I D, MOLTON N D, et al. MonoSLAM: Real-time single camera SLAM[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1052-1067.

[15] KLEIN G, MURRAY D. Parallel tracking and mapping for small AR workspaces[C]// Proceedings of the IEEE and ACM International Symposium on Mixed and Augmented Reality. Piscataway, NJ: IEEE Press, 2007: 225-234.

[16] MUR-ARTAL R, MONTIEL J M M, TARDÓS J D. ORB-SLAM: A versatile and accurate monocular SLAM System[J]. IEEE Transactions Robotics, 2015, 31(5): 1147-1163.

[17] NEWCOMBE R A, DAVISON A J. Live dense reconstruction with a single moving camera[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE Press, 2010: 1498-1505.

[18] MURARTAL R, TARDOS J D. Probabilistic semi-dense mapping from highly accurate feature-based monocular SLAM [C]// Proceedings of the Robotics Science and Systems, 2015.

[19] HINZMANN T, SCHNEIDER T, DYMCZYK M, et al. Robust map generation for fixed-wing UAVs with low-cost highly-oblique monocular cameras[C]// Proceedings of the IEEE/RSJ International Conference on Intelligent Robots & Systems. Piscataway, NJ: IEEE Press, 2016: 3261-3268.

[20] 蒙山,唐文名. 单目 SLAM 直线匹配增强平面发现方法[J]. 北京航空航天大学学报, 2017, 43(4): 660-666.
MENG S, TANG W M. Monocular SLAM linear matching enhanced plane discovery method[J]. Journal of Beijing University of Aeronautics and Astronautics, 2017, 43(4): 660-666 (in Chinese).

[21] VON STUMBERG L, USENKO V, ENGEL J, et al. Autonomous exploration with a low-cost quadrocopter using semi-dense monocular SLAM[C]// Processing of the European Conference on Mobile Robots, 2017.

[22] 蒋林,郭晨,朱志超. 嵌入式平台上的三维重建算法研究

[J]. 机械设计与制造,2018,330(8):264-266.
JIANG L,GUO C,ZHU Z C. Research on 3D reconstruction algorithm based on embedded platform[J]. Machinery Design & Manufacture,2018,330(8):264-266(in Chinese).

[23] BADRINARAYANAN V,KENDALL A,CIPOLLA R. SegNet:A deep convolutional encoder-decoder architecture for scene segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(12):2481-2495.

[24] 黄寅. 基于 HSV 颜色空间与形态学的车辆目标分割算法[J]. 大庆师范学院学报,2017,37(6):11-15.
HUANG Y, Vehicle target segmentation algorithm based on HSV color space and morphology[J]. Journal of Daqing Normal University,2017,37(6):11-15(in Chinese).

[25] 闫敬文. 数字图像处理:MATLAB 版[M]. 北京:国防工业出版社,2011:125-127.
YAN J W. Digital image processing by MATLAB[M]. Beijing: National Defence Industry Press,2011:125-127(in Chinese).

[26] 李鹏飞,吴海娥,景军锋,等. 点云模型的噪声分类去噪算法[J]. 计算机工程与应用,2016,52(20):188-192.
LI P F,WU H E,JING J F,et al. Noise classification and denoising algorithm for point cloud model[J]. Computer Engineering and Applications,2016,52(20):188-192(in Chinese).

作者简介:

张釜恺 男,硕士研究生。主要研究方向:计算机视觉、VSLAM。

芮挺 男,博士,教授,硕士生导师。主要研究方向:图像处理、模式识别、人工智能。

何雷 男,博士,讲师,硕士生导师。主要研究方向:系统科学,智能算法。

杨成松 男,博士,讲师。主要研究方向:图像处理、虚拟现实。

A low-cost indoor passable area modeling method for robots

ZHANG Fukai[1], RUI Ting[2,*], HE Lei[2], YANG Chengsong[2]

(1. School of Graduate, Army Engineering University of PLA, Nanjing 210000, China;
2. School of Field Engineering, Army Engineering University of PLA, Nanjing 210000, China)

Abstract: Monocular vision-based simultaneous localization and mapping (SLAM) is a popular technology in the field of robotics in recent years. However, due to the huge computation resource required by reconstruction, mainstream methods are not able to generate meaningful reconstruction of scene in real time on platforms with low computing power. This paper proposes a new fast passable area modeling method for the specific situation of indoor environment and small robots. The method is based on the monocular feature-based SLAM. Firstly, it obtains the road segmentation image through segmentation in the HSV color space with adaptive threshold. Then, the system cross-matches the segmentation with the sparse point cloud generated by SLAM, to obtain the ground plane and accurate ground segmentation area. Finally, it projects the ground segmentation area to the ground plane for dense modeling of the floor. In the experiment of indoor scene, the average calculation speed of the proposed method can reach 21 frames per second, and the speed is about 70% of ORB-SLAM, which can meet the real-time requirements of mobile platforms. The position error for the floor plane is 5.8% on average, and the modeling error of the road width is between 3.5% and 12.8%.

Keywords: monocular simultaneous localization and mapping; indoor scene; passable area; floor modeling; image segmentation

Received: 2019-07-16; **Accepted**: 2019-08-18; **Published online**: 2019-08-22 14:43
URL: kns.cnki.net/kcms/detail/11.2625.V.20190822.1027.001.html

* **Corresponding author**. E-mail: rtinguu@sohu.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0415

基于深度视觉语义嵌入的视频缩略图推荐

张梦琴，孟权令，张维刚*
（哈尔滨工业大学（威海）计算机科学与技术学院，威海 264209）

摘 要：视频缩略图作为视频内容最直观的表现形式，在视频共享网站中发挥很重要的作用，是吸引用户是否会点击观看该视频的关键要素之一。一句与视频内容相关的描述性语句，再搭配一幅与语句内容相关的视频缩略图，往往对用户更有吸引力，因此提出一种深度视觉语义嵌入模型来构建完整的视频缩略图推荐框架。该模型首先使用卷积神经网络（CNN）来提取视频关键帧的视觉特征，并使用循环神经网络（RNN）来提取描述语句的语义特征，再将视觉特征与语义特征嵌入到维度相同的视觉语义潜在空间；然后通过比较视觉特征与语义特征之间的相关性来推荐与特定的描述语句内容密切相关的视频关键帧作为视频缩略图推荐结果。在不同类型的网络视频数据上的实验表明，所提方法能够有效地从网络视频中推荐出与给定描述性语句内容较相关的视频缩略图序列，提升视频的用户浏览体验。

关 键 词：视频缩略图；关键帧；卷积神经网络（CNN）；循环神经网络（RNN）；视觉语义嵌入
中图分类号：TP391.4；TP183
文献标识码：A **文章编号**：1001-5965(2019)12-2479-08

随着移动互联网的快速普及和智能电子设备的高速发展，人们获取信息的方式已不仅仅满足于图像，而直接选择使用信息更丰富、画面感更强的视频，因此也带动了很多视频分享网站和 APP 的快速崛起，比如 YouTube、优酷，以及最近两年快速流行的抖音、快手等短视频分享 APP。人们越来越热衷于将自己生活中发生的事情用短视频的形式记录下来，并且分享到主流的视频共享网站上。再加上一些新闻、体育、电视、电影等相关的视频，使得互联网上的网络视频每天爆炸式增长，充斥在网络生活中。

为了能在海量的视频资源中快速地提高某段视频的点击率，以及快速高效地为用户找到需要的视频资源，现在较大型的视频分享网站都会对每段视频添加一个视频缩略图，并配上合适的标注文本，以将视频的内容“直截了当”的呈现给用户。可见，对于一段视频，用户初步看到的是视频缩略图的内容及对应的标注文本，这也是决定其是否点击并观看该视频的关键要素之一。一个好的视频缩略图会让这段视频更有吸引力，所以研究一种能够自动提取有意义且有较好代表性的视频缩略图的方法就显得尤为重要。

对于标注文本，一般由视频上传者手动输入，也可通过视频字幕（video captioning）技术来自动生成。既然有了标注文本，缩略图自然要和文本内容相匹配，因此，如何根据已有标注文本为视频选择一个合适的视频缩略图就成为一个值得研究的问题。本文提出一种基于深度视觉语义嵌入的网络视频缩略图自动生成框架，主要针对给定的一段视频及描述视频内容的标注文本，从视频中选取出既与标注文本内容相符又满足用户浏览体验需求的视频帧作为该视频的缩略图。该方法首

收稿日期：2019-07-26；**录用日期**：2019-08-14；**网络出版时间**：2019-08-21 16:46
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190821.1435.002.html
基金项目：国家自然科学基金（61672497）；山东省自然科学基金（ZR2017MF001）
* **通信作者**. E-mail：wgzhang@hit.edu.cn

引用格式：张梦琴，孟权令，张维刚. 基于深度视觉语义嵌入的视频缩略图推荐[J]. 北京航空航天大学学报，2019，45(12)：2479-2486. ZHANG M Q，MENG Q L，ZHANG W G. Video thumbnail recommendation based on deep visual-semantic embedding[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2479-2486 (in Chinese).

先使用卷积神经网络(CNN)来提取视频帧的视觉特征,并使用循环神经网络(RNN)来提取标注文本的语义特征;然后将视觉特征与语义特征嵌入到视觉语义潜在空间,视觉语义潜在空间是指视觉特征与语义特征具有相同维度与表示方式的空间,以便对视觉特征与语义特征进行相似度匹配;最后按照相似度得分对视频帧排序,选出分数最高的一个视频帧作为该视频的缩略图。同时,该方法还将选出多个与文本语义内容相关联的视频帧来作为推荐缩略图呈现给用户,以提高用户的可选择性。为了将选出的得分最高的缩略图与推荐的缩略图序列进行区分,本文将前者称为关键缩略图,后者称为推荐缩略图序列。

本文的贡献在于:①提出一个完整的视频缩略图推荐框架,该框架能够根据描述语句推荐相关联的视频缩略图序列;②提出一种深度视觉语义嵌入模型,模型将整个语句的语义特征与图像的视觉特征嵌入到共同的潜在空间中以获得两者的相关性;③在已有相关数据集的基础上,创建适用于本文任务的数据集,并取得较好的结果。

1 相关工作

大部分视频分享网站都用到了视频缩略图自动生成技术,但一些视频分享网站的视频缩略图通常是来自于视频的固定某个时序位置(第一帧、最后一帧或者中间帧),或者借助相关的图像捕获工具随机从视频中捕获一张图片作为视频缩略图。很显然,这种方法获取的视频缩略图不具有代表性并且选取的图片质量也得不到保证。

为了让自动生成的视频缩略图更具有代表性,Gao 等[1]提出了一种反映视频内容主题的视频缩略图提取算法,他们注意到基于视频帧颜色和运动信息等底层特征的所选帧可能不具有语义代表性,所以使用主题标准对生成缩略图的关键帧进行排序。Lian 和 Zhang[2]提出使用包含正面人脸信息、侧面人脸信息的高级视频特征与包含灰度直方图、像素值标准方差等低级视觉信息特征进行融合来选择最后的缩略图的方法。Jiang 和 Zhang[3]提出一种矢量量化方法来生成视频缩略图,利用视频时间密度函数(VIDF)来研究视频数据的时间特性,使用独立分量分析(ICA)构建空间特征。Liu 等[4]提出了一种查询敏感的动态网络视频缩略图生成方法,所选的缩略图不仅在视频内容上具有代表性还满足了用户的需求。而 Zhang 等[5]结合其之前在文献[6]中所提出的基于图像质量评估和视觉显著性分析的视频缩略图提取方法以及在文献[4]中发表的方法,提出一种综合考虑图像质量评估、图像的可访问性、图像的内容代表性以及缩略图与用户查询的关系等因素的方法,推荐出同时满足视频用户与浏览器需求的缩略图。Zhao 等[7]在基于视觉美学的自动缩略图选择系统[8]的基础上提出一种利用视觉元数据和文本元数据来自动合成类似杂志封面形式的视频缩略图方法;所推荐的缩略图不是取自于原视频,而是通过自动合成来得到的。

上述的部分视频缩略图选择方法[2-3,6]都集中于单纯从视频内容来学习视觉代表性。文献[1,4-5]研究了如何将查询与视频内容相结合来为不同查询提供不同的缩略图,但是他们都是使用基于搜索的方法。而 Liu 等[9]首次引入基于学习的方法,将深度视觉语义嵌入模型应用到视频缩略图生成任务中,开发一种多任务深度视觉语义嵌入模型,将查询和视频缩略图映射到一个共同的潜在语义空间,直接计算查询与视频缩略图之间的相似度,使得应用可以根据视觉和边缘信息自动选择依赖于查询的缩略图,但该方法的局限在于所用查询是以 Word Embedding(词嵌入向量,是指将一个单词转换成固定长度的向量表示形式)嵌入到潜在空间中,且一次只能获取单个单词的 Word Embedding 表示。对于查询语句或者多个查询关键词需要将查询以单词的向量形式依次与视频帧特征进行相似度匹配,这种匹配方式忽略了查询单词之间的关联信息。此外,文献[9]中的视频帧特征只使用简单几层的 CNN 进行提取,所提取的视频帧特征也不够丰富。本文提出的基于深度视觉语义嵌入的视频缩略图提取方法框架如图 1 所示。首先使用预训练的深度 CNN 有效地依次提取各关键帧的视觉特征,同时使用基于RNN的神经语言模型将整个语句的

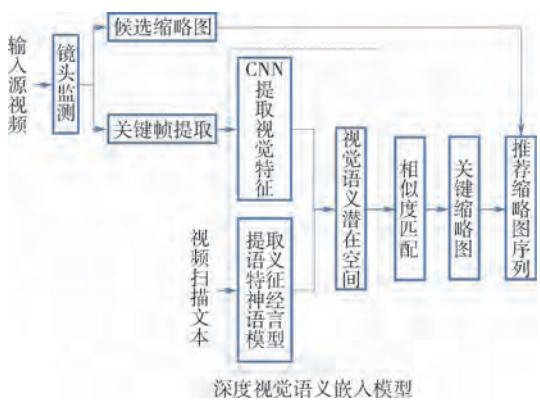


图 1 基于深度视觉语义嵌入模型的视频缩略图推荐框架

Fig. 1 Video thumbnail recommendation framework based on deep visual-semantic embedding model

语义特征嵌入到一个固定的向量,使得语义特征不但包含了单词之间的关联信息,也更易与视觉特征进行相关性比较。

2　深度视觉语义嵌入学习方法

基于深度视觉语义嵌入的缩略图推荐学习方法首先需要训练深度视觉语义嵌入模型,然后利用该模型实现缩略图推荐。因此本节首先介绍深度视觉语义嵌入模型,然后介绍整个视频缩略图推荐框架。

2.1　深度视觉语义嵌入模型

深度视觉语义嵌入模型[10],实现将从文本域中学习的语义信息和图像中的视觉信息共同嵌入到一个潜在的空间中,以便直接计算文本与图像之间的相关性。并且对于与图像内容无关的语义信息也能够根据与图像之间的相关性,返回相关性较高的图像[9]。而本文的任务就是从视频中选择与给定的语句语义相关的视频帧作为推荐的视频缩略图,因此可利用深度视觉语义嵌入模型来计算给定的语句信息与视频帧之间的相关性。所提出的深度视觉语义嵌入模型框架如图 2 所示。首先使用预训练的 CNN 依次提取视频关键帧序列的视觉特征,再使用基于 RNN 的神经语言模型来提取文本的语义特征,并将视觉特征与语义特征嵌入到视觉语义潜在空间。

近年来,RNN 由于具有特定的记忆功能已经在处理序列问题和自然语言处理等领域取得了很大的成功,所以本文使用 RNN 的变式即擅长解决中长文本序列间依赖问题的门控循环单元(GRU)[11]来提取单词之间的依赖信息。对于给定的描述语句,本文方法不再是简单的提取一个单词的 Word Embedding 来作为其特征向量,而是将一句话中的所有单词都用 Word Embedding 表示,然后将整句话作为序列输入GRU单元,最终输出一个表达这句话语义特征的向量,这也是典型的多对一关联模型,其主要结构如图 3 所示,本文方法使用 GRU 的最后一层隐层状态 $\boldsymbol{h}_n$,作为语义特征 $\boldsymbol{V}_s$:

$$\boldsymbol{V}_s = \boldsymbol{h}_n = \mathrm{GRU}(\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n) \tag{1}$$

式中:$\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n$ 为单词序列 $x_1, x_2, \cdots, x_n$ 的 Word Embedding 形式;n 为输入的单词个数。

本文方法使用预训练好的 ResNet152 模型[12]来提取视频帧的视觉特征,由于所提方法旨在将视觉特征嵌入到固定的潜在空间,因此需去除 ResNet 网络的最后一层全连接层,并提取最后一个卷积层的特征得到视觉特征 $\boldsymbol{V}_i$。

将语义特征和视觉特征分别嵌入到一个 N 维的潜在空间得到潜在语义特征 $\boldsymbol{V}_s'$ 和潜在视觉特征 $\boldsymbol{V}_i'$。为使语义特征向量尽可能地拟合提取到的视觉特征,本文采用均方误差损失函数:

$$\mathrm{Loss}(\boldsymbol{V}_s', \boldsymbol{V}_i') = \mathrm{MSE}(\boldsymbol{V}_s', \boldsymbol{V}_i') = \frac{\sum_{j=0}^{N-1} (\boldsymbol{V}_{sj}' - \boldsymbol{V}_{ij}')^2}{N} \tag{2}$$

式中:MSE 函数用于计算两个向量之间的均方误差;N 为潜在空间的维度(本文中是 2048 维)。

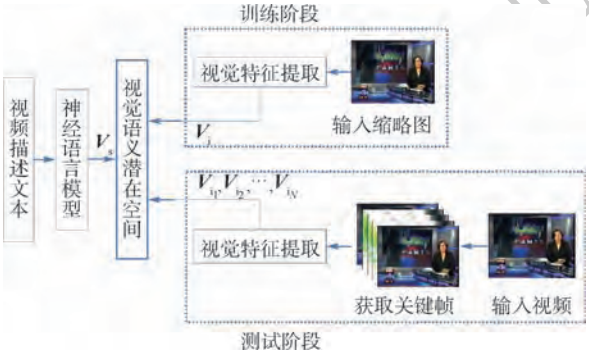


图 2　深度视觉语义嵌入模型示意图

Fig. 2　Schematic diagram of deep visual-semantic embedding model

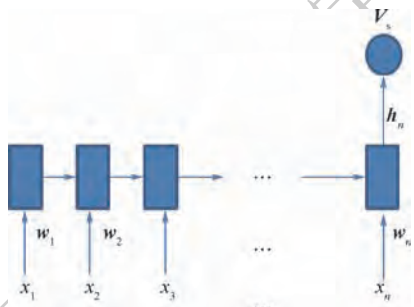


图 3　神经语言模型

Fig. 3　Neural language model

2.2　基于深度视觉语义嵌入模型的缩略图推荐

本节将介绍如何利用上述深度视觉语义嵌入模型来为一段视频选择缩略图。本过程分为 2 个阶段,分别是关键帧提取和缩略图推荐。

2.2.1　关键帧提取

本文方法使用基于顺序聚类与 K-means 聚类相结合的关键帧提取算法对视频进行镜头分割。参考文献[13]中的视频镜头分割算法,对视频进行初步聚类。首先将视频帧映射到 HSV(Hue,Saturation,Value)颜色空间,将 3 个颜色空间分别分成 12、5 和 5 三个量级,生成对应的归一化颜色直方图。将 3 个颜色空间的直方图结果组合到一起为每个视频帧生成一个 22 维的颜色空间向量。再用顺序聚类方法对转换过的视频帧进行镜头分割[13]。

对顺序聚类后的每个类的视频帧依次进行清晰度、亮度、色偏检测,并将最后检测的得分进行加权融合,得出图像质量评价得分D_f,检测方法如下:

清晰度检测:使用 Tenengrad 梯度函数来计算每个视频帧的清晰度得分f_d,得分越高,图像越清晰[14]。

色偏检测:将 RGB 图像转变到 CIE $L^*a^*b^*$ 空间(L^*表示明暗度,a^*表示红-绿轴,b^*表示黄-蓝轴),通常存在色偏的图像,在 $a*$ 分量和 $b*$ 分量上的均值会偏离原点很远,方差也会偏小;计算衡量图像色偏程度的 K 因子得到色偏检测得分f_c,分数越高,图像色偏越严重[15]。

亮度检测:与色偏检测相似,计算图片在灰度图上的均值和方差,当存在亮度异常时,均值会偏离开均值点(假设为 128);同样根据计算衡量图像亮度程度的 K 因子得到亮度检测得分f_b,分数越高,图像亮度异常越严重。

最后的图像质量评价得分为 3 种评价属性的得分的加权融合。本文设置清晰度检测的权重为 0.5,亮度检测的权重为 0.3,色偏检测的权重为 0.2,由于图像质量与亮度检测得分 f_c 以及色偏检测得分 f_b 成反比,所以最后的得分 D_f 为

$$D_f = 0.5f_d - 0.2f_c - 0.3f_b \tag{3}$$

得到每个视频帧的图像质量得分 D_f 后,通过分数排序将每个聚类中分数较低的一半视频帧过滤掉(对只包含一个视频帧的聚类不进行过滤)。获得过滤后的视频帧和聚类数 K_0。接下来对过滤后的视频帧再进行 K-means 聚类,K 值设置为 K_0,得到最后的镜头分割结果。

如图 4 所示,将视频帧进行镜头分割后,需要从每个镜头中提取能够充分代表视频镜头的视频帧添加到关键帧序列以及为每个视频镜头挑选候选缩略图序列。

关键帧序列是从每个视频镜头中挑选出最具代表性的一个视频帧组成的序列,其将作为视觉语义嵌入模型的输入以获得关键帧的视觉特征序列。模型输出关键帧的视觉特征序列中与输入的文本语义特征相关性最高的视觉特征所对应的关键帧作为关键缩略图。

候选缩略图序列是对关键帧序列的扩充,从每个视频镜头中挑选最具代表性的某几个视频帧,组成该视频镜头的候选缩略图序列。每个视频镜头都有一个候选缩略图序列,它们之间是独立的。如果关键缩略图属于这个视频镜头,那么这个镜头的候选缩略图序列就会成为最后的推荐缩略图序列。

为了从每个镜头中提取合适的视频关键帧,

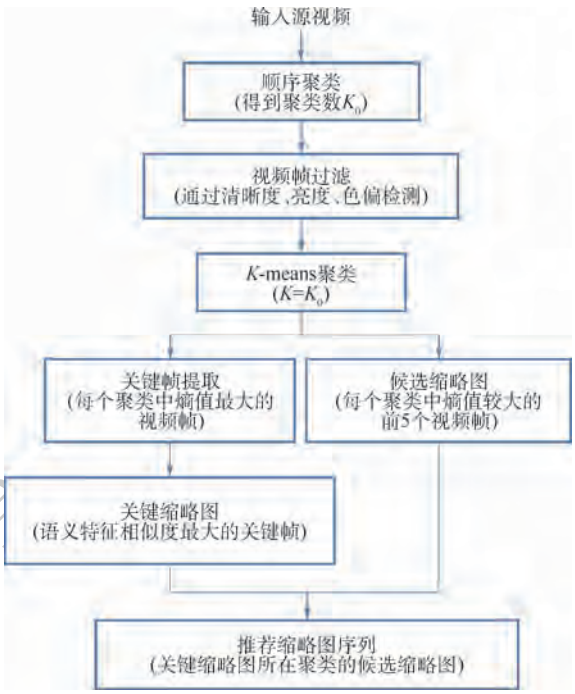


图 4 基于已训练模型的视频缩略图推荐框架

Fig. 4 Video thumbnail recommendation framework based on trained model

本文使用熵值来作为视频信息量的度量,熵值越大,表明图像的信息越丰富,越具有代表性,在本文中,分别计算 HSV 三个颜色空间的熵值,其计算式为

$$E(f_i) = -\sum_{k=1}^{m} P_k \log_2 P_k \tag{4}$$

式中:P_k 为每个视频帧生成归一化颜色直方图的值。然后分别计算出 H 分量的熵值 $E(f_h)$,$m=12$;S 分量的熵值 $E(f_s)$,$m=5$;以及 V 分量的熵值 $E(f_v)$,$m=5$。由于人眼对 Hue 的敏感性比对 Saturation 和 Value 高[11],所以本文中对 H、S、V 三个分量的熵值分别赋 0.5、0.3、0.2 的权重,最终的熵值 $E(f)$ 为

$$E(f) = 0.5E(f_h) + 0.3E(f_s) + 0.2E(f_v) \tag{5}$$

得到视频镜头中每个视频帧的熵值 $E(f)$ 后,本文提取每个镜头中熵值最大的视频帧作为该镜头的关键帧添加到关键帧序列。提取熵值最大的前 5 个视频帧(不足 5 帧的镜头取全部视频帧)作为该镜头的候选缩略图序列。所以每个镜头的关键帧也包含在该镜头的候选缩略图序列中。由于聚类数是由顺序聚类结果而定,所以聚类数(视频镜头数)是不固定的,取得的关键帧序列的数量也是不固定的。

2.2.2 缩略图推荐

缩略图推荐,如图 4 所示,获取关键帧序列之后需要对所有的关键帧进行视觉特征提取得到视

觉特征序列 $\boldsymbol{V}_{i_1}, \boldsymbol{V}_{i_2} \cdots, \boldsymbol{V}_{i_N}$,并将这些视觉特征都映射到视觉语义潜在空间中。对于输入文本,使用 2.1 节中训练好的神经语言模型提取语义特征 $\boldsymbol{V}_s$,并将其也映射到视觉语义潜在空间中。分别计算潜在语义特征 $\boldsymbol{V}'_s$ 与潜在视觉特征序列 $\boldsymbol{V}'_{i_1}$, $\boldsymbol{V}'_{i_2} \cdots, \boldsymbol{V}'_{i_N}$ 之间的余弦相似度,相似度最大的视觉特征所对应的关键帧将作为关键缩略图。为了增加推荐缩略图的多样性与可选择性,如果某个关键帧被选取作为关键缩略图,由于与关键缩略图在同一个视频镜头中的视频帧可能具有相同的视觉特征,因此将关键缩略图所在镜头的候选缩略图序列作为最后的推荐缩略图序列。显然,推荐的缩略图序列,必然包含关键缩略图,推荐的缩略图序列最少包含 1 帧,最多包含 5 帧。

3　算法实现

3.1　数据集

本文的关键是训练深度视觉语义嵌入模型,在训练阶段需要采用深度学习中的有监督学习方法,所以首先要解决的就是数据集问题。已存在的缩略图生成方法都是在图像质量评价或建立关联模型等图像处理的层面来获取对应的缩略图,并且都是依赖于选择算法的缩略图生成方法。所以目前还没有专门针对缩略图推荐任务的数据集可供训练,本文只能自行收集或者对已有的相关数据集进行改动使其满足对本任务的训练。

首先考虑的是微软的 MS COCO 2014[16] 数据集,该数据集中包含 8 万多张训练图片,4 万多张测试图片,由于该数据集是做图像字幕(image captioning)任务常用的数据集,所以每张图片都有 5 个左右的人工标注文本语句来对图像内容进行描述。图像字幕任务是输入一张图片得到一个描述该图像内容的句子,本文提出的基于深度视觉语义嵌入的视频缩略图推荐任务,虽然是基于视频的任务,但实质上是根据输入一句视频描述文本,从视频中获取与文本内容相关联的一张图像。相比于图像字幕任务,本文的任务是从文本到图像的一个相反的过程,所以可以将 MS COCO 2014 数据集中的标注语句作为模型的训练数据,将图片作为目标值来进行模型的训练,获得用于训练的{语句,缩略图}对。

另一个数据集是微软的 MSR-VTT(MSR Video-to-Text)[17],该数据集是用来做视频相关任务,比如视频字幕。数据集中每段视频大约有 20 条描述性语句,每段视频时长在 20 s 以内。由于本文的训练目标是获得图像,所以需要对此数据集进行重新标注,生成适合本任务的数据集。本文根据 MSR-VTT 数据集提供的描述性语句从对应的视频中选出与语句内容最相关并且视觉效果较好的视频帧,组成训练需要的{语句,缩略图}对。由于该数据集中有很多存在偏义以及无法从视频中获得相关视频帧的描述语句,所以本文对原数据集提供的描述语句进行了筛选,每段视频只留取 10 句左右的描述语句。之所以要在 MS COCO 2014 数据集训练后还用这个视频数据集来训练是因为 MS COCO 2014 数据集图片之间的差异较大,图片之间以及标注文本之间的关联性较小,不利于细节处的学习。而从 MSR-VTT 数据集中选出的句子有些来源于同一段视频,所以语句之间的关联性较大,而对应的图片之间也会存在一定的联系,更有利于模型的学习。最后从 MSR-VTT 数据集的前 400 段视频中收集了 4 000 多个描述语句以及对应的 800 多个视频帧,其中不同的标注文本可能会对应同一个视频帧,最后得到包含 4 000 多个{语句,缩略图}对的数据集。

3.2　模型的训练

3.2.1　文本预处理

所使用的语料库来源于 MS COCO 2014 数据集中的标注文本。首先将所有标注语句中的单词都转化为小写形式,然后使用 NLTK 工具包中的 word_tokenize 函数对句子进行分词处理,并将标点符号都移除。对每个单词出现的次数进行统计,将出现频率小于 3 次的单词移除。再加上 < pad > < unk > 这 2 个补齐和补缺的标注符,一共得到8 576 个单词的语料库。对于输入的文本语句,需要将语句中单词映射成词汇表中对应的单词序列号,并且将每条语句的长度都设置为 30,语句长度不够就用 < pad > 对应的序列号补齐,在语料库中找不到对应单词则用 < unk > 对应的序列号代替。

3.2.2　图像预处理

训练阶段的缩略图图像以及测试阶段的视频关键帧在进行视觉特征提取之前,需要将图像缩放到 224 × 224 大小。然后使用预训练的 ResNet152 网络模型对图像进行特征提取,并将最后一个卷积层的特征保存下来,得到一个 2 048 维的特征向量。

3.2.3　参数设置

训练模型阶段:数据集形式为{语句,缩略图}对,潜在语义空间的维度是 2 048 维。训练神经语言模型时,设置 Word Embedding 维度为 512,

输出的语义特征维度为 2 048。由于 2 个数据集的规格不同,需要对其进行分开训练,先在 MS COCO 2014 数据集上进行训练,设置 BatchSize 为 128,学习率为 0.001,学习了 2 个 epoch。然后在 MS COCO 2014 训练的基础上再在处理过的 MSR-VTT 数据集上进行训练,设置 BatchSize 为 16,学习率为 0.001,学习了 100 个 epoch。所有的参数都使用 Adam 优化器进行优化。

在生成缩略图阶段:首先对视频每 6 帧提取一帧作为视频的输入,在进行顺序聚类时,经过多次实验将阈值设置为 0.85。顺序聚类所得的聚类数作为 K-means 聚类的 K 值。

4 实验结果

为了有效地对本文提出的框架进行评估,本文分别从 YouTube 和优酷网站下载了不同类别的视频来测试该框架推荐的视频缩略图的效果。用于测试的视频类别主要有:教育、娱乐、电影、游戏与卡通、新闻与政治、生活、体育等。使用击中率 HIT@1[9] 作为评价指标,即推荐的缩略图序列中如果有与描述语句的语义内容相匹配的缩略图则为击中,反之如果推荐的缩略图序列中的所有缩略图与语句描述之间都不相关的话则为不击中。本文为了合理地显示实验结果,共设置了 3 个等级:A 表示推荐的缩略图序列中有与语义内容完全相关的缩略图,称为完全击中;B 表示推荐的缩略图序列中有与语义内容部分相关的缩略图,称为一般击中;C 表示推荐的缩略图序列与所给出的语义内容完全不相关,称为完全不击中。

本文为每个类别各选取 2 ~ 4 段视频,每个视频时长在 3 min 左右,且都包含了多个场景的切换。针对每段视频,结合所给的 5 个描述文本语句,分别进行缩略图推荐。根据每个描述文本语句从对应的视频中挑选出关键缩略图作为默认缩略图,同时返回 1 ~ 5 个视频帧作为推荐的缩略图序列供用户选择。

本文采用主观评价的方式来测定方法的有效性,邀请 10 位了解过视频缩略图任务的用户对本文方法的推荐结果进行了主观打分评价。即针对每段视频的每条描述文本语句所推荐的缩略图序列,结合上述给出的 3 个评价等级来对推荐的结果进行评价,在实验测试集上获得的评价结果如表 1 所示。从击中率可以看出,所提方法对生活、电影与娱乐类视频所推荐的视频缩略图有较好的效果,而对游戏与动画、教育类的视频效果不是很理想。

图 5 给出了部分推荐缩略图示例,每个示例的第一个视频帧即是选出的关键缩略图,剩余为推荐的缩略图序列。其中图 5(a)显示的是完全击中的视频缩略图(生活类视频的推荐结果),图 5(b)显示的是完全没有击中的视频缩略图(体育类视频的推荐结果)。从图 5 (a)中可以看出,针对每个描述文本语句,可得到 1 ~ 5 幅推荐缩略图,推荐的缩略图与给出的语句内容完全相符,而且缩略图序列具有一定的丰富性,增加了用户的可选择性;但从图 5(b)中看出,该方法针对体育类的视频效果并不是很好,一方面是因为体育类的视频包含较多的大幅度运动,不能很好地提取和表征其视觉特征;另一方面是所采用的描述文本语句,如果其本身较复杂不易被理解,则会影响所提取的语义特征表达,最终使得本文所描

表 1 不同类别网络视频的击中率

Table 1 Hit rates of web videos in different categories

视频类别	视频个数	语句条数	完全击中率(A)/%	击中率(A + B)/%
教育	3	15	29.3	56.0
娱乐	3	15	30.7	63.3
电影	2	10	43.0	72.0
游戏与动画	2	10	20.0	51.0
新闻与政治	3	15	27.3	71.3
生活	4	20	53.0	87.0
体育	3	15	33.3	66.7
合计	20	100	35.0	68.3

图 5 本文方法所获得的推荐缩略图序列示例

Fig. 5 Examples of recommended thumbnail sequence obtained by proposed method

述的视觉语义嵌入模型不能很好的拟合。

此外,由于不是在专门的缩略图推荐评测视频集上测试,且训练集中所使用的{语句,缩略图}对中缩略图特征表达受限,所以本文方法目前对于通俗简单易懂的描述文本语句有较好的效果,而对描述复杂理解偏难的语句,并不能较好地推荐出合适的缩略图序列,因此还具有很大的改进空间。

5　结　论

1)本文针对网络视频缩略图的自动推荐问题,提出一种深度视觉语义嵌入模型和缩略图推荐框架,实现了将图片的视觉信息与语句的语义信息嵌入到共同的潜在空间。

2)本文提出的框架能有效地根据给定的描述语句为用户推荐内容相关且具有视觉代表性的视频缩略图序列。20 段 3 min 左右的视频,100 条描述语句,推荐的缩略图序列完全击中率为 35.0%,完全击中与一般击中的总比率为 68.3%。

3)本文提出的框架能够为用户推荐细节多样化的缩略图,推荐的缩略图语义场景相同但具体细节不同,增加了用户的可选择性。

4)对视频表观内容有较好描述的文本语句,能够获得更好的缩略图推荐结果;但对视频中的运动信息表征识别能力偏弱,导致缩略图推荐结果受影响。

由于没有专门针对本文任务的数据集,所训练模型的准确度还有待改进。今后也需要在关键帧提取以及语料库上进行算法调整,以便进一步提高模型的准确度。

参考文献(References)

[1] GAO Y L,ZHANG T,XIAO J. Thematic video thumbnail selection[C]// Proceedings of the 2009 IEEE International Conference on Image Processing (ICIP). Piscataway,NJ:IEEE Press,2009:4333-4336.

[2] LIAN H C,LI X Q,SONG B. Automatic video thumbnail selection[C]// Proceedings of the 2011 IEEE International Conference on Multimedia Technology (ICMT). Piscataway,NJ:IEEE Press,2011:242-245.

[3] JIANG J F,ZHANG X P. Video thumbnail extraction using video time density function and independent component analysis mixture model[C]// Proceedings of the 2011 IEEE International Conference on Acoustic,Speech,and Signal Processing (ICASSP). Piscataway,NJ:IEEE Press,2011:1417-142.

[4] LIU C X,HUANG Q M,JIANG S Q. Query sensitive dynamic web video thumbnail generation[C]// Proceedings of the 2011 IEEE International Conference on Image Processing (ICIP). Piscataway,NJ:IEEE Press,2011:2449-2452.

[5] ZHANG W G,LIU C X,WANG Z J,et al. Web video thumbnail recommendation with content-aware analysis and query-sensitive matching[J]. Multimedia Tools and Applications,2014,73:547-571.

[6] ZHANG W G,LIU C X,HUANG Q M,et al. A novel framework for web video thumbnail generation[C]// Proceedings of the 8 th International Conference on Intelligent Information Hiding and Multimedia Signal Processing. Piscataway,NJ:IEEE Press,2012:343-346.

[7] ZHAO B Q,LIN S J,QI X,et al. Automatic generation of visual-textual web video thumbnail[C]// Siggraph Asia Posters. New York:ACM,2017:41.

[8] SONG Y L,REDI M,VAllMITJANA J,et al. To click or not to click:Automatic selection of beautiful thumbnails from videos[C]// Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York:ACM,2016:659-668.

[9] LIU W,MEI T,ZHANG Y D,et al. Multi-task deep visual-semantic embedding for video thumbnail selection[C]// Proceedings of IEEE Conference on Computer Vision an Pattern Recognition(CVPR). Piscataway,NJ:IEEE Press,2015:3707-3715.

[10] ANDREA F,GREG S C,JON S,et al. Devise:A deep visual-semantic embedding model[C]// Proceedings of Neural Information Processing Systems Conference. Nevada:NIPS,2013:2121-2129.

[11] CHUNG J,GULCEHRE C,CHO K H,et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[EB/OL]. (2014-12-11)[2019-07-25]. https://arxiv.org/abs/1412.3555.

[12] HE K M,ZHANG X Y,REN S Q,et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway,NJ:IEEE Press,2016:770-778.

[13] 金红,周源华,梅承力. 用非监督式聚类进行视频镜头分割[J]. 红外与激光工程,2000,29(5):42-46.
JIN H,ZHOU Y H,MEI C L. Video shot segmentation using unsupervised clustering[J]. Infrared and Laser Engineering,2000,29(5):42-46(in Chinese).

[14] 李祚林,李晓辉,马灵玲,等. 面向无参考图像的清晰度评价方法研究[J]. 遥感技术与应用,2011,26(2):239-246.
LI Z L,LI X H,MA L L,et al. Research of definition assessment based on no-reference digital image quality[J]. Remote Sensing Technology and Application,2011,26(2):239-246(in Chinese).

[15] 徐晓昭,蔡轶珩,刘长江,等. 基于图像分析的偏色检测及颜色校正方法[J]. 测控技术,2008,27(5):10-12.
XU X Z,CAI Y H,LIU C J,et al. Color cast detection and color correction methods based on image analysis[J]. Chinese Journal of Measurement and Control Technology,2008,27(5):10-12(in Chinese).

[16] LIN T Y,MAIRE M,BELONGIE S,et al. Microsoft COCO:Common objects in context[C]// Proceedings of European Conference on Computer Vision. Berlin:Springer,2014:740-755.

[17] XU J,MEI T,YAO T,et al. MSR-VTT:A large video descrip-

tion dataset for bridging video and language[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE Press, 2016: 5288-5296.

作者简介:

张梦琴 女,硕士研究生。主要研究方向:多媒体分析、机器学习等。

孟权令 男,博士研究生。主要研究方向:多媒体分析、机器学习等。

张维刚 男,博士,副教授,硕士生导师。主要研究方向:多媒体分析、计算机视觉、模式识别、机器学习等。

Video thumbnail recommendation based on deep visual-semantic embedding

ZHANG Mengqin, MENG Quanling, ZHANG Weigang*

(School of Computer Science and Technology, Harbin Institute of Technology, Weihai 264209, China)

Abstract: Video thumbnail, as the most intuitive form of video content, plays an important role in video sharing sites and is one of the key elements to attract users to click and watch the video. However, a descriptive statement related to video content with a video thumbnail associated with the content of the statement is often more attractive to user. Therefore, a complete video thumbnail recommendation framework with a deep visual-semantic embedding model is proposed in this paper. This model uses the convolutional neural network to extract the visual features of video keyframes, and uses recurrent neural network to extract the semantic features of description sentences. After embedding the visual features and the semantic features into the visual-semantic potential space of the same dimension, the key frames related to the content of the descriptive sentences are recommended as video thumbnails by comparing the correlation between the visual features and the semantic features. Experiments on different categories of web videos show that the proposed method can effectively recommend contented-related video thumbnail sequence from videos for given descriptive statements and enhance the user experience.

Keywords: video thumbnail; keyframes; convolutional neural network (CNN); recurrent neural network (RNN); visual-semantic embedding

Received: 2019-07-26; **Accepted**: 2019-08-14; **Published online**: 2019-08-21 16:46

URL: kns.cnki.net/kcms/detail/11.2625.V.20190821.1435.002.html

Foundation items: National Natural Science Foundation of China (61672497); Natural Science Foundation of Shandong Province (ZR2017MF001)

* **Corresponding author**. E-mail: wgzhang@hit.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0368

基于多尺度失真感知特征的重定向图像质量评估

吴志山，张帅，牛玉贞*
（福州大学 数学与计算机科学学院，福州 350108）

摘 要：在不同宽高比显示设备上的图像观看体验通常受到图像重定向操作方法的影响。为了提高重定向图像主观感知与客观评估之间的一致性，提出了基于多尺度失真感知特征（MSDA）的客观重定向图像质量评估（RIQA）方法。语义失真和细节失真经常出现在图像的不同尺度上，因此从图像的不同尺度中提取失真感知特征。提出了一个描述原始图像和重定向图像之间的宽高比相似度（ARS）的精确度量。此外，使用视觉注意力融合图来模拟人类视觉系统对图像的主观关注度。在2个基准数据库上的实验结果表明，所提出的MSDA方法的肯德尔排名相关系数（KRCC）、皮尔逊线性相关系数（PLCC）和斯皮尔曼秩次相关系数（SRCC）指标分别比对比方法中最优方法提高4.1%、1.8%和4.5%。

关 键 词：图像重定向；重定向图像质量评估（RIQA）；视觉注意力融合；宽高比相似度（ARS）；多尺度
中图分类号：V221+.3；TB553
文献标识码：A **文章编号**：1001-5965(2019)12-2487-08

人们越来越依赖移动设备进行工作和娱乐，例如可以通过展开或闭合屏幕以改变屏幕尺寸的可折叠屏幕设备。因此，如何调整图像大小以适应具有不同大小和宽高比的显示设备屏幕，成为与用户体验相关的重要问题。

已经有许多成熟的图像重定向方法被提出。为了理解图像重定向方法，图1展示了原始图像以及MIT数据库中应用8个重定向方法的结果图像，包括裁剪（CR）、多操作符（MOP）[1]、缝裁剪（SC）[2]、线性缩放（SCL）、移位映射（SM）[3]、拉伸和放缩（SNS）[4]、视频流（SV）[5]和变形（WARP）[6]。操作方法后的括号内是对应图像的主观投票数，数值越大表示该图像质量越好。每幅重定向图像保持原始图像的高度并将宽度减小到原始图像的一半。从图1可以看出，不同的重定向方法具有不同的关注点，并且重定向图像的主观平均意见分数（Mean Opinion Scores，MOSs）也不同。主观地评估每幅重定向图像的质量成本太高。因此，有必要开发有效的客观重定向图像质量评估（Retargeted Image Quality Assessment，RIQA）方法。为了促进对客观RIQA方法的研究，Rubinstein[7]和Ma[8]等建立了2个包含重定向图像和相应的主观评分的数据库。

目前，许多RIQA方法被研究出来。早期的RIQA方法通常使用相似性度量。Simakov等[9]提出了一种基于优化视觉数据的双向相似性（Bi-directional Similarity，BDS）度量方法。Liu等[10]提出了一种逆序（自上而下）的方法来组织图像从全局到局部的特征，并通过基于颜色度量的相似性（Color metric-based Similarity，CSim）评估重定向图像。

近年来，已经出现了基于原始图像和重定向图像之间配准的RIQA度量。Zhang等[11]提出了局部块的宽高比相似性（Aspect Ratio Similarity，

收稿日期：2019-07-09；**录用日期**：2019-08-03；**网络出版时间**：2019-08-12 17:24
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190812.1544.004.html
基金项目：国家自然科学基金（61672158）；福建省自然科学基金（2019J02006）
*通信作者. E-mail：yuzhenniu@gmail.com

引用格式：吴志山，张帅，牛玉贞. 基于多尺度失真感知特征的重定向图像质量评估[J]. 北京航空航天大学学报，2019，45(12)：2487-2494. WU Z S，ZHANG S，NIU Y Z. Retargeted image quality assessment based on multi-scale distortion-aware feature[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2487-2494 (in Chinese).

图 1 重定向操作方法示例

Fig. 1 Examples of retargeting operation method

ARS)度量来描述图像的结构失真。Zhang 等[12]提出了基于多级特征(Multiple-Level Feature, MLF)的 RIQA 方法。MLF 考虑 ARS、脸部块变形、边缘组相似性特征并使用回归学习来预测重定向图像的感知质量。

尽管现有的 RIQA 方法表现出很好的评估性能,但主观和客观 RIQA 之间的一致性仍有待提高。由于高级语义失真,如成分变化和显著对象丢失,以及初级细节失真,如局部内容丢失和局部形状变化可能出现在图像的不同尺度上,本文从多个尺度的图像块中提取失真感知特征。具体而言,本文提出了原始图像和重定向图像之间的 ARS 的改进度量。此外,本文使用融合的视觉注意力图,其结合了显著性图以及脸部和线条特征,以模拟人类视觉系统对图像的主观关注。

1 多尺度失真感知特征度量

本文提出的多尺度失真感知特征(Multi-Scale Distortion-Aware, MSDA)度量可以分为 3 个阶段。第 1 阶段,计算视觉注意力融合图[13],通过后向配准方法[11]建立原始图像和重定向图像之间的像素级对应关系。第 2 阶段,图像重定向过程中的失真由建立的像素级对应关系进行模拟。改进的 ARS[11]结合视觉注意力融合图可以更好地捕获每个局部块的信息丢失和视觉失真。本文还使用边缘组相似度[12]和脸部块相似性特征[12]来分别表示对边缘组和脸部块的失真度量。边缘组相似性度量边缘空间排列的变化。图 2 展示了边缘组检测结果。脸部块相似性用于描述脸部变形程度。图3展示了脸部块检测结果,脸部块用矩形框标出。最后,机器学习方法用于将特征映射到主观 MOSs 并获得训练模型以预测所测试重定向图像的客观质量分数。

图 2 边缘组检测结果

Fig. 2 Edge group detection results

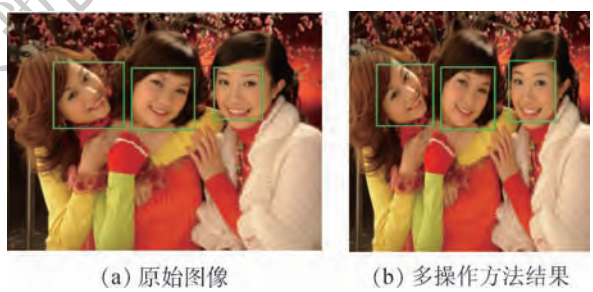


图 3 脸部块检测结果

Fig. 3 Face block detection results

1.1 改进的 ARS 度量

本文采用 ARS[11]作为初级特征,并改进局部

块的 ARS 度量。首先将原始图像划分成规则网格块(例如 16 像素×16 像素)。使用后向配准方法[11]计算重定向图像中与原始图像相对应的变形块。图 4 展示了后向配准示例,(a)为原始图像所划分的块,(b)为由原始图像经过多操作符方法得到的结果图像,为了便于展示后向配准结果,图像块划分为 64×64。如图 4(b)所示,重定向后所对应的块不再是规则的网格块,通过块的形状变化来度量由重定向操作引入的变形。对于重定向前后的每一对图像块,原始 ARS 计算公式为

$$S_{\mathrm{ar}} = \frac{2r_{\mathrm{w}}r_{\mathrm{h}} + C}{r_{\mathrm{w}}^2 + r_{\mathrm{h}}^2 + C}\mathrm{e}^{-\alpha(r_{\mathrm{m}}-1)^2} \tag{1}$$

式中:r_{w} 和 r_{h} 分别为重定向块的宽度和高度的变化率;$r_{\mathrm{m}} = (r_{\mathrm{w}} + r_{\mathrm{h}})/2$ 为绝对尺度变化率;$C = 1 \times 10^{-6}$ 和 $\alpha = 0.3$ 为 2 个参数;$\dfrac{2r_{\mathrm{w}}r_{\mathrm{h}} + C}{r_{\mathrm{w}}^2 + r_{\mathrm{h}}^2 + C}$ 表示度量视觉失真;$\mathrm{e}^{-\alpha(r_{\mathrm{m}}-1)^2}$ 表示度量信息损失。

那么整张图像的 ARS 分数可以表示为 Q_{AR},计算公式为

$$Q_{\mathrm{AR}} = \sum_{i=1}^{N_1} w(i) S_{\mathrm{ar}}(i) \tag{2}$$

式中:N_1 为原始图像当中规则块的数量;$w(i)$ 为第 i 块所对应的权重;$S_{\mathrm{ar}}(i)$ 为第 i 块的 ARS 值。

考虑到式(1)无法在所有情况下正确度量形状畸变。例如,当原始图像的某个块在重定向图像中没有匹配到相应的块时,即 $r_{\mathrm{w}} = r_{\mathrm{h}} = 0$,式(1)计算的视觉失真等于 1,这意味着该情况下,ARS 没有考虑视觉失真。因此,当在重定向期间移除整个块时,本文引入参数 λ 作为视觉失真的惩罚因子。改进的 ARS 度量表示为 S_{iar},计算公式为

$$S_{\mathrm{iar}} = \begin{cases} \dfrac{2r_{\mathrm{w}}r_{\mathrm{h}}}{r_{\mathrm{w}}^2 + r_{\mathrm{h}}^2}\mathrm{e}^{-\alpha(r_{\mathrm{m}}-1)^2} & r_{\mathrm{w}} \neq 0 \text{ 或 } r_{\mathrm{h}} \neq 0 \\ \lambda \mathrm{e}^{-\alpha(r_{\mathrm{m}}-1)^2} & r_{\mathrm{w}} = r_{\mathrm{h}} = 0 \end{cases} \tag{3}$$

此外,式(2)中的权重 $w(i)$ 在 RIQA 方法中起重要作用。通常通过视觉显著性检测方法来计算权重。文献[14]指出,RIQA 方法结合不同的显著性检测算法通常表现出不同的性能。Zhang 等[13]提出了一种视觉注意融合框架(VAF),该框架首先结合了不同显著性检测算法计算得到的显著性图,然后增强了图像中的脸部和线条特征,最终得到更加全面的显著性图。因此,本文将 VAF 计算得到的显著性图作为融合 ARS 块分数的权重。

(a) 原始图像　　(b) 重定向图像

图 4　后向配准示例

Fig. 4　An example of backward registration

对于整张图像,改进的 ARS 度量表示为 Q_{IAR},计算公式为

$$Q_{\mathrm{IAR}} = \sum_{i=1}^{N_1} \tilde{S}(i) S_{\mathrm{iar}}(i) \tag{4}$$

式中:$\tilde{S}$ 为 VAF 计算得到的显著性图。

图 5 为 VAF 示例。图 5(b)所示显著性融合图是通过分别对基于离散余弦变换(Discrete Cosine Transform, DCTS)[15]和基于元胞自动机(Background-based map optimized via Single-layer Cellular Automata, BSCA)[16]方法计算得到的显著性图进行均衡化处理,并将 2 幅结果图进行融合,即在相同位置像素点的显著性值取 2 幅结果图的平均值。图 5(c)所示面线增强图是通过对重定向图像进行面部检测,同时检测大于对图像角线三分之一的直线,并对这些检测结果位置的显著性进行增强。

(a) 原始图像　　(b) 显著性融合图　　(c) 面线增强图

图 5　VAF 示例

Fig. 5　An example of VAF

1.2　MSDA 融合

块是 ARS 度量的最小单位。不同尺度的块包含不同的信息。因此,ARS 在不同尺度块上捕捉到的失真信息是不同的。为了获得最佳评估性能,需要确定最佳的图像块尺度。然而,每幅重定向图像的失真类型不同,与之对应的最佳块尺度也有差异。难以为每幅图像确定最佳尺度,因此使用多尺度方案来解决该问题。由不同尺度捕获的失真弥补单尺度捕获失真的不足,从而取得更好的评估性能。本文使用 8×8 和 16×16 两个尺度的块获取图像 ARS 特征。Q_{IAR}^{8} 和 Q_{IAR}^{16} 分别表示由式(4)计算得到的块尺度为 8 和 16 的 2 个改进后的 ARS 度量。Q_{EGS}[12] 和 Q_{FBS}[12] 分别表示边缘组相似性分数和脸部块相似性分数。对于每幅重定向图像,分别计算该图像的 Q_{IAR}^{8}、Q_{IAR}^{16}、Q_{FBS}、

Q_{EGS}这 4 个相似性分数,并将这 4 个特征作为训练特征,主观 MOSs 作为训练标签,利用支持向量机(SVM)进行回归学习,最后,向学习好的模型输入待评估图像的这 4 个特征,并输出该图像的客观重定向质量评估分数。

本文按照 MLF 方法进行实验,在 CUHK 数据库[8]上采用基于径向基函数(RBF)内核的支持向量回归模型。由于 MIT 数据库[7]提供的是成对比较投票,投票表明来源于同一幅原始图像的重定向图像之间的排名,这与 CUHK 数据库提供的 MOSs 不同。因此,使用基于 RBF 内核的 SVMrank[17]对 MIT 数据库图像进行排名回归而不是直接应用支持向量回归。

2 实验结果

本节将介绍 MIT[7]和 CUHK[8]数据库及其性能评估指标,然后将提出的方法与这 2 个公共数据库上的客观 RIQA 方法进行比较,包括 BDS[9]、SIFT(Scale-Invariant Feature Transform) flow[18]、EMD(Earth-Mover's Distance)[19]、CSim[10]、PGDIL(Perceptual Geometric Distortion and Information Loss)[20]、ARS[11]和 MLF[12]。实验结果表明,与其他方法相比,所提出的 MSDA 方法与人类主观感知更为一致。

2.1 数据库及评估标准

1) MIT 数据库

MIT 数据库包含 37 幅原始图像。每幅图像由 8 种典型的重定向操作方法进行重定向,如图 1所示。经过 8 种重定向操作后,原始图像共生成 296 幅重定向图像。210 名参与者参加了主观测试,该测试成对进行,受试者从 2 个并排的重定向图像中选择他们喜欢的图像。重定向图像被选择的次数用作客观质量评估的主观分数并用于数据库实验评估。本文使用肯德尔排名相关系数(Kendall Rank Correlation Coefficient, KRCC)[21]来衡量客观分数与主观排名之间的相关性。KRCC 计算公式为

$$\mathrm{KRCC}=\frac{n_c-n_d}{0.5n(n-1)} \tag{5}$$

式中:n 为排名序列长度;n_c 和 n_d 分别为与主观排序一致和不一致的图像对数量。

2) CUHK 数据库

该数据库包含 57 幅原始图像和 171 幅重定向图像。除了 MIT 数据库中使用的 8 种图像重定向方法外,CUHK 数据库还包括优化的缝雕刻和缩放(Seam Carving and Scaling, SCSC)[22]和基于能量变形(Energy-based deformation, ENER)[23]方法。CUHK 数据库采用 5 级质量量化策略分别对每幅重定向图像进行评分,以获得主观意见得分。最后,通过平均获得每个重定向图像的 MOSs。4 个常用的性能评估指标,包括皮尔逊线性相关系数(Pearson Linear Correlation Coefficient, PLCC)、斯皮尔曼秩次相关系数(Spearman Rank-order Correlation Coefficient, SRCC)、均方根误差(Root Mean Squared Error, RMSE)和异常值比率(Outlier Ratio, OR)[24],用于评估 RIQA 方法的性能。PLCC 可以通过 Sheikh 等[25]提出的非线性回归映射函数获得,方程式为

$$f(x)=\beta_1\left(\frac{1}{2}-\frac{1}{1+\mathrm{e}^{\beta_2(x-\beta_3)}}\right)+\beta_4 x+\beta_5 \tag{6}$$

式中:$\beta_1,\beta_2,\cdots,\beta_5$ 为需要进行拟合的参数。SRCC 用于测量 RIQA 度量预测的客观分数的单调性。通过计算非线性回归后主观 MOSs 与客观评估得分之间的均方根误差,即 RMSE。OR 为异常值的数量与客观评估得分总数的比率。异常值是在非线性回归之后落在区间[MOS -2σ, MOS $+2\sigma$]之外的分数,其中 σ 为客观评价分数的标准偏差。较大的 PLCC 值和 SRCC 值表明客观评估得分与主观 MOSs 值之间的相关性较高,而较小的 RMSE 和 OR 值表明 RIQA 方法的预测分数与主观分数越接近。

2.2 MIT 数据库性能对比

本文使用基于 RBF 内核的 SVMrank[17]在 MIT 数据库上采用 LOOCV(Left One-Out Cross-Validation)方式进行多尺度失真感知特征融合。对于每幅原始图像,使用原始图像作为查询,并将 8 幅重定向图像的主观投票作为排名顺序。将原始图像和 8 幅重定向图像归为一组,共计 37 组。对于每组,使用剩余的 36 组进行训练,然后对该组进行测试。最后,评估结果。将提出的 MSDA 方法与现有方法进行比较。给出每个 RIQA 度量的平均 KRCC 和标准偏差,以及具有特定属性的图像子集的平均 KRCC。表 1 给出 MSDA 方法在 MIT 各子集的 KRCC 指标,其中 p-val 是指在一个概率模型中,统计摘要(如两组样本均值差)与实际观测数据相同,或甚至更大这一事件发生的概率。实验结果显示,本文提出的 MSDA 方法性能是每个子集中所有对比方法的前两名,并且平均性能比 MLF 方法提高 4.1%。

为进一步研究所提方法的有效性,表 2 给出在 MIT 数据库上的特征分析实验。边缘组相似性和脸部块相似性特征的有效性已在MLF[12]中

表 1　MIT 数据库性能对比

Table 1　Performance comparison on MIT database

方法	KRCC						KRCC 均值	KRCC 标准差	p-val
	线条	人脸	前景物体	纹理	几何结构	对称结构			
BDS	0.040	0.190	0.167	0.060	−0.004	−0.012	0.083	0.268	0.017
SIFTflow	0.097	0.252	0.218	0.161	0.085	0.071	0.145	**0.262**	0.031
EMD	0.220	0.262	0.226	0.107	0.237	0.500	0.251	0.272	1×10^{-5}
CSim	0.097	0.290	0.293	0.161	0.053	0.150	0.164	**0.263**	0.028
PGDIL	0.431	0.390	0.389	0.286	0.438	**0.523**	0.415	0.296	6×10^{-10}
ARS	**0.463**	**0.519**	**0.444**	**0.330**	**0.505**	0.464	**0.452**	0.283	$\mathbf{1\times10^{-11}}$
MLF	**0.486**	**0.605**	**0.544**	**0.384**	**0.536**	**0.536**	**0.512**	**0.251**	$\mathbf{1\times10^{-14}}$
MSDA	**0.511**	**0.633**	**0.539**	**0.455**	**0.571**	**0.535**	**0.533**	0.265	$\mathbf{9\times10^{-21}}$

表 2　MIT 数据库特征分析

Table 2　Feature analysis on MIT database

方法	线条	人脸	前景物体	纹理	几何结构	对称结构	KRCC 均值	KRCC 标准差
$Q_{\mathrm{ARS}}+Q_{\mathrm{EGS}}+Q_{\mathrm{FBS}}$	0.503	0.614	0.532	0.420	0.554	0.500	0.525	**0.260**
$Q^{8}_{\mathrm{IAR}}+Q_{\mathrm{EGS}}+Q_{\mathrm{FBS}}$	0.506	0.610	0.456	0.528	**0.571**	0.512	0.527	0.269
$Q^{16}_{\mathrm{IAR}}+Q_{\mathrm{EGS}}+Q_{\mathrm{FBS}}$	0.503	**0.643**	0.455	**0.552**	0.558	0.512	0.531	0.267
MSDA	**0.511**	0.633	**0.539**	0.455	**0.571**	**0.535**	**0.533**	0.265

得到验证,因此仅对多尺度特征进行实验,该实验使用 VAF 计算的显著性图。表 2 第 1 组实验为不使用改进的 ARS 度量的情况下,仅将 ARS 方法的显著性检测方法更改为 VAF 方法进行实验。第 2、3 组为使用单一尺度的改进宽高比特征 Q^{8}_{IAR} 和 Q^{16}_{IAR} 进行实验,相比于 MLF,性能均有提升,而使用 MSDA 方法,性能进一步提高,说明2 个尺度的 ARS 特征具有互补的作用,可以更好地捕获重定向图像失真。此外,VAF 对 MLF 性能的提高为 2.5%,在 VAF 的基础上,Q^{16}_{IAR} 和多尺度特征分别提高 1.1% 和 1.5%,说明显著性检测算法对评估结果影响较大。

2.3　CUHK 数据库性能对比

本文在 CUHK 数据库上采用 SVR 5 折交叉验证模型。数据库随机分为 2 个子集,80% 作为训练集,20% 作为测试集。训练集和测试集之间没有重叠。重复随机训练测试过程 1 000 次并记录 1 000次迭代的中值。从表3可以看出,尽管 RMSE 和 OR 的指标低于 MLF 方法,但 MSDA 的 PLCC 和 SRCC 指标相对 MLF 方法分别提高 1.8% 和4.5%,总体具有更好的主观感知一致性和预测性能。

表 3　CUHK 数据库性能对比

Table 3　Performance comparison on CUHK database

方法	PLCC	SRCC	RMSE	OR
BDS	0.2896	0.2887	12.922	0.2164
SIFTflow	0.3141	0.2899	12.817	0.1426
EMD	0.2760	0.2904	12.977	0.1696
CSim	0.4374	0.5662	12.141	0.1520
PGDIL	0.5403	0.5409	11.361	0.1520
ARS	**0.6835**	**0.6693**	**9.855**	**0.0702**
MLF	**0.7577**	**0.7383**	**8.525**	**0.0294**
MSDA	**0.7713**	**0.7717**	**8.593**	**0.0350**

与 MIT 数据库实验类似,本文同样给出在验 CUHK 数据库上的特征分析实验,实验采用的特征与 MIT 数据库的对比实验相同,实验过程与表 3相同。表 4 给出 CUHK 数据库上的特征分析结果,可以得到与 MIT 数据库上相同的结论,进一步说明 MSDA 方法的有效性。

表 4　CUHK 数据库特征分析

Table 4　Feature analysis on CUHK database

方法	PLCC	SRCC	RMSE	OR
$Q_{\mathrm{ARS}}+Q_{\mathrm{EGS}}+Q_{\mathrm{FBS}}$	0.7610	0.7585	8.7596	0.0468
$Q^{8}_{\mathrm{IAR}}+Q_{\mathrm{EGS}}+Q_{\mathrm{FBS}}$	0.7639	0.7694	8.7130	0.0409
$Q^{16}_{\mathrm{IAR}}+Q_{\mathrm{EGS}}+Q_{\mathrm{FBS}}$	0.7709	0.7663	8.599	0.0409
MSDA	**0.7713**	**0.7717**	**8.593**	**0.0350**

2.4　讨　论

为了研究不同尺度组合对评估性能的影响,本文对不同尺度块的改进 ARS 特征组合进行实验。表5 和 表6 分别为 MIT 和 CUHK 数据库的

表 5　MIT 数据库不同块尺度特征组合

Table 5　Different block scale feature combinations on MIT database

组合	KRCC 均值	KRCC 标准差
$Q^{8}_{\mathrm{IAR}}+Q^{16}_{\mathrm{IAR}}$	**0.533**	0.265
$Q^{8}_{\mathrm{IAR}}+Q^{32}_{\mathrm{IAR}}$	0.529	**0.260**
$Q^{16}_{\mathrm{IAR}}+Q^{32}_{\mathrm{IAR}}$	0.525	0.267

不同特征组合的对比实验。本文在这个实验中提取的块尺度特征分别为 Q_{IAR}^{8}、Q_{IAR}^{16}和 Q_{IAR}^{32}，由于块尺度过大会导致 ARS 度量忽略细节失真，另一方面，选择的块尺度变化过小，则对实验结果影响不明显。因此，块尺度取 8、16、32。从实验结果可以看出，当块尺度组合为 8 和 16 时，RIQA 方法结果具有最高的主观感知一致性。

表 7 和表 8 给出 MSDA 方法分别在 2 个数据库上结合不同显著性检测算法的对比实验结果，采用当前最新的级联部分解码器（CPD）[26] 以及 ARS 方法使用的 DCTS[15] 显著性检测算法。由表 7 和表 8 实验结果可以得出，VAF 在 3 种不同显著性检测算法当中性能表现最好。并且最新的显著性检测算法并没有表现出好的性能，其原因在于 CPD 方法的检测结果虽然可以很好地检测显著性区域，但该方法并没有对显著性区域和次显著性区域进行区分，即显著性区域都赋予同样的显著性值。而 VAF 方法则对不同显著性区域赋予不同的显著性值，特别是对人脸和线条部分进行显著性增强，使得该显著性检测算法更符合重定向图像在质量评估上的需求。

本文对 ARS 特征进行深入研究，λ 的取值为 0 ~ 1，λ 取值越大表示对重定向图像中被移除块的视觉失真惩罚越小，而不同图像被移除块对重定向图像的影响不同，因此有必要对 λ 进行实验，寻找最佳的参数设置。图 6 给出了不同 λ 取值对实验结果的影响，从柱状图可以看出，λ = 0.66时，MSDA 方法性能最好。

表 6 CUHK 数据库不同块尺度特征组合

Table 6 Different block scale feature combinations on CUHK database

组合	PLCC	SRCC	RMSE	OR
$Q_{\mathrm{IAR}}^{8}+Q_{\mathrm{IAR}}^{16}$	**0.7713**	**0.7717**	**8.5930**	**0.0350**
$Q_{\mathrm{IAR}}^{8}+Q_{\mathrm{IAR}}^{32}$	0.7541	0.7537	8.8658	0.0409
$Q_{\mathrm{IAR}}^{16}+Q_{\mathrm{IAR}}^{32}$	0.7677	0.7666	8.6522	**0.0350**

表 7 MIT 数据库不同显著性检测算法实验结果对比

Table 7 Comparison of experimental results of detection algorithms with different saliency on MIT database

组合	KRCC 均值	KRCC 标准差
VAF[13]	**0.533**	0.265
CPD[26]	0.456	0.314
DCTS[15]	0.523	**0.231**

表 8 CUHK 数据库不同显著性检测算法实验结果对比

Table 8 Comparison of experimental results of detection algorithms with different saliency on CUHK database

组合	PLCC	SRCC	RMSE	OR
VAF[13]	**0.7713**	**0.7717**	**8.5930**	**0.0350**
CPD[26]	0.7469	0.7494	8.9781	0.0468
DCTS[15]	0.7333	0.7256	9.1793	0.0526

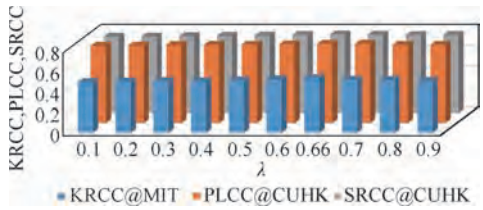


图 6 不同 λ 值对实验结果的影响

Fig. 6 Influence of different λ values on experimental results

3 结 论

本文提出了一个 MSDA 方法来预测客观重定向图像质量，并使用视觉注意力融合图结合 4 个有效特征来捕获图像重定向期间产生的失真。在 2 个公共数据库 MIT 和 CUHK 上，所提出的 MSDA 方法比对比方法具有更高的主观感知一致性。本文深入研究细节失真特征并通过实验可得以下结论：

1）相比于改进的 ARS 度量以及多尺度方法，显著性检测算法对实验结果影响更大。

2）改进的 ARS 度量原始图像中整块像素在重定向过程中全部被移除的情况，该度量与 VAF 计算的显著性图相结合可以更好地捕捉细节失真，从而提高 MSDA 方法性能。

3）多尺度方法为 RIQA 度量提供更全面的失真信息，弥补单尺度捕获失真的不足，从而提高 RIQA 方法与人类主观感知一致性。

虽然所提出的 MSDA 方法表现出良好的性能，但无论是细节失真特征、语义失真特征以及显著性检测算法，都需要进一步研究以更好地进行客观 RIQA。

参考文献（References）

[1] RUBINSTEIN M, SHAMIR A, AVIDAN S. Multi-operator media retargeting [J]. ACM Transactions on Graphics, 2009, 28(3): 23.

[2] AVIDAN S, SHAMIR A. Seam carving for content-aware image resizing[J]. ACM Transactions on Graphics, 2007, 26(3): 10.

[3] PRITCH Y, KAV-VENAKI E, PELEG S. Shift-map image editing[C]// IEEE 12th International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2009: 151-158.

[4] WANG Y S, TAI C L, SORKINE O, et al. Optimized scale-and-stretch for image resizing[J]. ACM Transactions on Graphics, 2008, 27(5): 118.

[5] KRÄHENBÜHL P, LANG M, HORNUNG A, et al. A system for retargeting of streaming video[J]. ACM Transactions on Graphics, 2008, 28(5): 126.

[6] WOLF L, GUTTMANN M, COHEN-OR D. Non-homogeneous content-driven video-retargeting[C]// IEEE 11th International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2007: 1-6.

[7] RUBINSTEIN M, GUTIERREZ D, SORKINE O, et al. A comparative study of image retargeting[J]. ACM Transactions on Graphics, 2010, 29(6): 160.

[8] MA L, LIN W, DENG C, et al. Image retargeting quality assessment: A study of subjective scores and objective metrics[J]. IEEE Journal of Selected Topics in Signal Processing, 2012, 6(6): 626-639.

[9] SIMAKOV D, CASPI Y, SHECHTMAN E, et al. Summarizing visual data using bidirectional similarity[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2008: 1-8.

[10] LIU Y J, LUO X, XUAN Y M, et al. Image retargeting quality assessment[J]. Computer Graphics Forum, 2011, 30(2): 583-592.

[11] ZHANG Y, FANG Y, LIN W, et al. Backward registration based aspect ratio similarity for image retargeting quality assessment[J]. IEEE Transactions on Image Processing, 2016, 25(9): 4286-4297.

[12] ZHANG Y, LIN W, LI Q, et al. Multiple-level feature-based measure for retargeted image quality[J]. IEEE Transactions on Image Processing, 2018, 27(1): 451-463.

[13] ZHANG S, NIU Y Z, LIN J W, et al. Visual attention fusion framework for image retargeting quality assessment[J]. IOP Conference Series: Earth and Environmental Science, 2019, 234(1): 12-64.

[14] CHEN Z, LIN J, LIAO N, et al. Full reference quality assessment for image retargeting based on natural scene statistics modeling and bi-directional saliency similarity[J]. IEEE Transactions on Image Processing, 2017, 26(11): 5138-5148.

[15] FANG Y, CHEN Z, LIN W, et al. Saliency detection in the compressed domain for adaptive image retargeting[J]. IEEE Transactions on Image Processing, 2012, 21(9): 3888-3901.

[16] QIN Y, LU H, XU Y, et al. Saliency detection via cellular automata[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015: 110-119.

[17] JOACHIMS T. Training linear SVMs in linear time[C]// ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM, 2006: 217-226.

[18] LIU C, YUEN J, TORRALBA A. SIFT flow: Dense correspondence across scenes and its applications[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2010, 33(5): 978-994.

[19] PELE O, WERMAN M. Fast and robust earth mover's distances[C]// IEEE 12th International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2009: 460-467.

[20] HSU C C, LIN C W, FANG Y, et al. Objective quality assessment for image retargeting based on perceptual geometric distortion and information loss[J]. IEEE Journal of Selected Topics in Signal Processing, 2014, 8(3): 377-389.

[21] KENDALL M G. A new measure of rank correlation[J]. Biometrika, 1938, 30(1): 81-93.

[22] DONG W, ZHOU N, PAUL J C, et al. Optimized image resizing using seam carving and scaling[J]. ACM Transactions on Graphics, 2009, 28(5): 125.

[23] KARNI Z, FREEDMAN D, GOTSMAN C. Energy-based image deformation[J]. Computer Graphics Forum, 2009, 28(5): 1257-1268.

[24] WANG Z, LU L, BOVIK A C. Video quality assessment based on structural distortion measurement[J]. Signal Processing: Image Communication, 2004, 19(2): 121-132.

[25] SHEIKH H R, SABIR M F, BOVIK A C. A statistical evaluation of recent full reference image quality assessment algorithms[J]. IEEE Transactions on Image Processing, 2006, 15(11): 3440-3451.

[26] WU Z, SU L, HUANG Q M. Cascaded partial decoder for fast and accurate salient object detection[EB/OL]. (2018-04-18)[2019-07-09]. https://arxiv.org/abs/1904.08739? context=cs.CV.

作者简介:

吴志山　男,硕士研究生。主要研究方向:多媒体和计算机图形学。

张帅　男,硕士。主要研究方向:多媒体和计算机图形学。

牛玉贞　女,博士,教授,博士生导师。主要研究方向:计算机视觉、多媒体和计算机图形学。

Retargeted image quality assessment based on multi-scale distortion-aware feature

WU Zhishan, ZHANG Shuai, NIU Yuzhen*

(College of Mathematics and Computer Science, Fuzhou University, Fuzhou 350108, China)

Abstract: The subjective visual experience of viewing images using a variety of display devices is usually affected by image retargeting operation. To improve the consistency between subjective perception and objective assessment for the retargeted images, we present an objective retargeted image quality assessment (RIQA) method based on multi-scale distortion-aware (MSDA) features. Because semantic distortion and detail distortion appear on different scales of the image, we propose to extract distortion-aware features from multiple scales of the image. Specifically, we present an accurate measurement for the aspect ratio similarity (ARS) between the original and retargeted images. Furthermore, we use a fused visual attention map to simulate the subjective attention of the human visual system to the image. The experimental results on the two benchmark databases show that the Kendall rank correlation coefficient (KRCC), Pearson linear correlation coefficient (PLCC), and Spearman rank-order correlation coefficient (SRCC) indicators of the proposed MSDA method are 4.1%, 1.8%, and 4.5% higher than the optimal method in the comparative methods, respectively.

Keywords: image retargeting; retargeted image quality assessment (RIQA); visual attention fusion; aspect ratio similarity (ARS); multi-scale

Received: 2019-07-09; **Accepted**: 2019-08-03; **Published online**: 2019-08-12 17:24

URL: kns.cnki.net/kcms/detail/11.2625.V.20190812.1544.004.html

Foundation items: National Natural Science Foundation of China (61672158); Natural Science Foundation of Fujian Province, China (2019J02006)

* **Corresponding author**. E-mail: yuzhenniu@gmail.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0373

面向行为识别的人体空间协同运动结构特征表示与融合

莫宇剑[1], 侯振杰[1,*], 常兴治[2], 梁久祯[1], 陈宸[3], 宦娟[1]

(1. 常州大学 信息科学与工程学院, 常州 213164; 2. 常州信息职业技术学院 智能制造工业云开放实验室, 常州 213164;
3. 北卡罗来纳大学夏洛特分校 电气与计算机工程系, 夏洛特市 28223)

摘 要: 针对人体执行动作时不同身体部位之间的协同关系,提出了基于人体空间协同运动结构特征的行为识别方法。首先度量人体不同部位对完成动作的贡献度,并将不同部位的贡献度转变为协同运动结构特征模型。然后利用模型无监督、自适应地对不同身体部位的运动特征进行约束。在此基础上借鉴跨媒体检索方法 JFSSL 对不同模态的特征进行特征选择与多模态特征融合。实验表明,所提方法在自建的行为数据库上明显提高了开放测试的识别率,且计算过程简便,易于实现。

关 键 词: 贡献度度量; 协同运动; 结构特征; 特征选择; 多模态融合
中图分类号: TP391
文献标识码: A **文章编号**: 1001-5965(2019)12-2495-11

行为识别实现了计算机对人体行为的理解与描述,是视频监控[1]、智能看护、人机交互[2]等领域的关键技术。人体行为识别的研究具有广泛的应用前景以及可观的经济价值[3]。执行不同动作时,不同身体部位的协同运动结构特征是行为识别的难题。现今行为识别的方法大致分为 3 类:①利用单一模态的行为数据进行人体行为识别;②利用多个不同模态的行为数据进行人体行为识别;③利用单一模态内的不同人体部位间的协同运动结构特征进行行为识别。

利用单一模态的行为数据进行人体行为识别,学者们已经提出了许多成熟的算法。按行为识别算法使用的数据类型大致分为:①基于深度图像的人体行为识别,如 DMM-LBP-FF(Depth Motion Maps-based Local Binary Patterns-Feature Fusion)[4]、HOJ3D(Histograms of 3D Joint locations)[5]等;②基于骨骼数据的人体行为识别,如协方差描述符[6]等;③基于惯性传感器数据的人体行为识别,如函数拟合[7]等。

利用多个不同模态的行为数据进行人体行为识别,也有很多的学者进行了研究。多模态行为数据的融合大致可以分为 2 类:特征级融合和决策级融合[8]。常见的特征级融合方法有 CRC(Collaborative Representation Classifier)[9]、SRC(Sparse Representation Classifier)[10]、CCA(Canonical Correlation Analysis)[11]、DCA(Discriminant Correlation Analysis)[12]等。决策级融合方法如 DS(Dempster-Shafer)证据理论[10]、多学习器协同训练[13]。

利用单一模态内的不同人体部位间的协同运动结构特征进行行为识别,目前已有少量的研究。Si 等[14]提出空间推理网络,其利用 Kinect 采集的

收稿日期: 2019-07-09; **录用日期**: 2019-08-03; **网络出版时间**: 2019-08-19 14:42
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190819.1417.001.html
基金项目: 国家自然科学基金(61063021,61803050); 江苏省物联网移动互联技术工程重点实验室开放课题(JSWLW-2017-013)
* **通信作者**. E-mail: houzj@cczu.edu.cn

引用格式: *莫宇剑, 侯振杰, 常兴治, 等. 面向行为识别的人体空间协同运动结构特征表示与融合[J]. 北京航空航天大学学报, 2019, 45(12): 2495-2505. MO Y J, HOU Z J, CHANG X Z, et al. Structural feature representation and fusion of behavior recognition oriented human spatial cooperative motion [J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2495-2505 (in Chinese).*

人体关节点位置数据来捕捉每个帧内的高级空间结构特征。首先将身体每个部位的连接转换成具有线性层的单独空间特征，然后将身体部位的个体空间特征输入到残差图神经网络（Residual Graph Neutral Network，RGNN）以捕获不同身体部位之间的高级结构特征。Liu 等[15]提出动态姿态图像描述人体行为，其利用姿态估计方法得到人体 14 个联合估计图，每一个联合估计图表示一个关节的运动，然后使用神经网络融合不同关节的运动。邓诗卓等[16]利用人体不同部位的三轴向传感器（三轴加速度、三轴陀螺仪等）的相同轴向数据间隐藏的空间依赖性并结合卷积神经网络（Convolutional Neural Networks，CNN）进行动作识别。

基于单模态、多模态的行为识别方法虽然可以识别不同的行为，但是忽略了人体在执行动作时不同身体部位的空间协同运动结构特征。动作的完成需要多块肌肉共同协调配合，例如走路不仅需要双腿，同时也需要摆动手臂以协调身体平衡[14]。现有计算不同身体部位结构特征的方法大都使用了神经网络，其训练时间复杂度高、计算资源消耗高、可解释性差等。

本文针对使用神经网络计算结构特征的复杂性高等问题，提出利用人体不同部位的三轴加速度数据构建人体空间协同运动的结构特征模型，并无监督、自适应地对不同身体部位的运动特征进行约束。首先对执行不同动作时的人体不同部位的三轴加速度曲线进行曲线特征分析，确定不同身体部位对不同动作的完成具有不同的贡献度；其次对执行某一动作时人体全身三轴加速度曲线进行曲线特征分析，并提出利用曲线的多个统计值度量人体不同部位的贡献度；然后由于多个统计量对构建结构特征模型具有不同的权重，需要进行权重的确定；再使用协同运动结构特征模型对人体不同部位的运动特征进行无监督、自适应地约束；最后使用多模态特征选择与特征融合算法将不同模态的特征进行融合并分类识别。

1 总体框架

本文提出利用人体空间协同运动的结构特征模型，对不同人体部位的运动特征进行约束，然后对多模态特征进行特征选择与特征融合，总体框架如图 1 所示。对骨骼图序列按 Chen 等[17]方法进行特征提取，图 1 中的“骨骼图序列”引自文献[18]。首先，对人体不同部位的三轴加速度数据分别提取特征；然后使用结构特征模型约束特征；最后使用多模态特征选择与特征融合方法训练多模态特征投影矩阵用于融合三轴加速度数据特征与关节点位置数据特征，使用分类器对融合后的特征进行分类识别，得到类别标签。

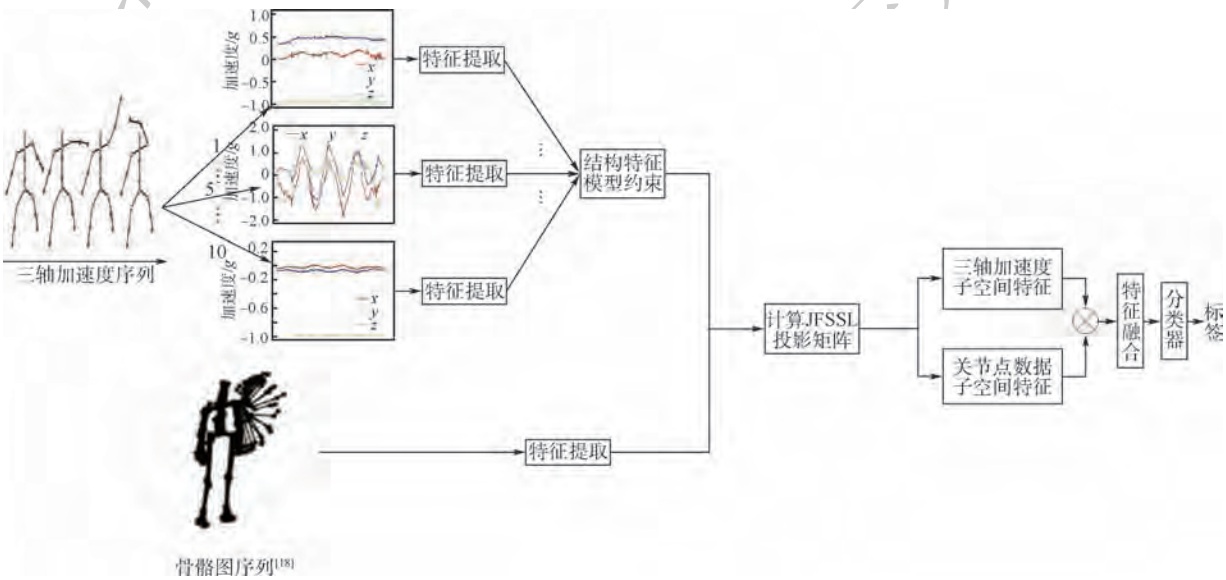


图 1 总体框架

Fig. 1 Overall framework

2 协同运动结构特征表示

设 $\gamma_k=\{a_x,a_y,a_z\}$ 为人体第 k 个惯性传感器采集的三轴加速度数据，人体执行动作时全身的三轴加速度数据为 $\gamma=\{\gamma_1,\cdots,\gamma_k,\cdots,\gamma_K\}$，其中 K 为人体惯性传感器的总数。对人体不同部位的三轴加速度数据进行特征提取，得到人体全身的运动特征 $\boldsymbol{F}^K=[f_{xk},f_{yk},f_{zk}]$，$k=1,2,\cdots,K$，其中 f_{xk}

为人体第 k 个部位 x 轴的特征,f_{yk} 为人体第 k 个部位 y 轴的特征,f_{zk} 为人体第 k 个部位 z 轴的特征。然后使用协同运动结构特征模型对特征进行约束从而使得样本具有更好的分类能力。

将 n 个样本划分为 c 类,n_i 为第 i 类样本的数量 $n=\sum_{i=1}^{c}n_i$。设 $\boldsymbol{x}_{ij}\in\boldsymbol{X}$ 为第 i 类第 j 个样本,第 i 类样本均值 $\bar{\boldsymbol{x}}_i$、所有 c 类样本均值 $\bar{\boldsymbol{x}}$ 的计算式分别为

$$\bar{\boldsymbol{x}}_i=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{x}_{ij}\qquad i=1,2,\cdots,c \tag{1}$$

$$\bar{\boldsymbol{x}}=\frac{1}{n}\sum_{i=1}^{c}n_i\bar{\boldsymbol{x}}_i \tag{2}$$

分别计算类间距离 S_{b} 与类内距离 S_{w}:

$$S_{\mathrm{b}}=\sum_{i=1}^{c}(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_i-\bar{\boldsymbol{x}})^{\mathrm{T}} \tag{3}$$

$$S_{\mathrm{w}}=\sum_{i=1}^{c}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i)^{\mathrm{T}} \tag{4}$$

然后计算类间距离 S_{b} 与类内距离 S_{w} 的比值 J_{w}。为了使得样本具有更好的可分离性,需要增大比值 J_{w}:

$$J_{\mathrm{w}}=\frac{S_{\mathrm{b}}}{S_{\mathrm{w}}} \tag{5}$$

本文利用人体空间协同运动的结构特征模型对特征进行约束,约束后的第 i 类样本均值 $\bar{\boldsymbol{x}}_i^{\,*}$、所有 c 类样本均值 $\bar{\boldsymbol{x}}^{*}$ 的计算式分别为

$$\bar{\boldsymbol{x}}_i^{*}=\frac{1}{n_i}\sum_{j=1}^{n_i}\boldsymbol{W}_{ij}\boldsymbol{x}_{ij}\qquad i=1,2,\cdots,c \tag{6}$$

$$\bar{\boldsymbol{x}}^{*}=\frac{1}{n}\sum_{i=1}^{c}n_i\bar{\boldsymbol{x}}_i^{*} \tag{7}$$

式中:$\boldsymbol{W}_{ij}=[w_{xk},w_{yk},w_{zk}]$,$k=1,2,\cdots,K$ 为第 i 类第 j 个样本对应的结构特征,并用于约束人体全身的运动特征,得到约束后的特征 $(\boldsymbol{F}^{K})^{*}=[f_{xk}^{*},f_{yk}^{*},f_{zk}^{*}]$,$k=1,2,\cdots,K$。以计算 f_{yk}^{*} 为例,$f_{yk}^{*}=w_{yk}f_{yk}$。计算约束后的类间距离 S_{b}^{*} 与类内距离 S_{w}^{*}:

$$S_{\mathrm{b}}^{*}=\sum_{i=1}^{c}(\bar{\boldsymbol{x}}_i^{*}-\bar{\boldsymbol{x}}^{*})(\bar{\boldsymbol{x}}_i^{*}-\bar{\boldsymbol{x}}^{*})^{\mathrm{T}} \tag{8}$$

$$S_{\mathrm{w}}^{*}=\sum_{i=1}^{c}\sum_{j=1}^{n_i}(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i^{*})(\boldsymbol{x}_{ij}-\bar{\boldsymbol{x}}_i^{*})^{\mathrm{T}} \tag{9}$$

其中:S_{b}^{*} 为约束后的样本类间距离;S_{w}^{*} 为约束后的样本类内距离。使 S_{b}^{*} 与 S_{w}^{*} 的比值 J_{w}^{*} 大于 J_{w},即可使样本更好地分离。

以左高挥手、右高挥手动作为例,左高挥手动作的主运动结点为左手肘、左手腕;右高挥手动作的主运动结点为右手肘、右手腕,其他的身体部位为附加运动结点。本文分别对比左高挥手与右高挥手动作的主运动结点以及附加运动结点的三轴加速度曲线。主运动结点曲线的对比如图 2 所示,附加运动结点曲线的对比如图 3 所示。图 2(a)为执行左高挥手动作时左手肘、左手腕、右手肘、右手腕部位对应的三轴加速度曲线图。类似地,图 2(b)为执行右高挥手动作时左手肘、左手腕、右手肘、右手腕部位对应的三轴加速度曲线图。图 3(a)为左高挥手动作附加运动结点的三轴加速度曲线图,图 3(b)为右高挥手动作附加运动结点的三轴加速度曲线图。

首先对比左高挥手、右高挥手动作附加运动结点的三轴加速度曲线,可以发现这 2 个动作附加结点的三轴加速度曲线形状相同均保持平稳即方差差异性小,此外曲线的均值也接近,即这 2 个动作附加结点的类间距离小,分类能力差。然后比对左高挥手、右高挥手动作主运动结点的三轴加速度曲线,对于区分左高挥手与右高挥手动作主要是根据左手肘、左手腕、右手肘、右手腕处三轴加速度曲线的统计特征。为了使样本更好地分离,需要强化每个动作主运动结点的特征、削弱附加运动结点的特征,基于此本文使用基于人体空间协同运动的结构特征模型对不同身体部位的特征进行约束。以右高挥手动作为例,从图 2(b)、图 3(b)中可以观察到,对于动作的完成不同身体部位的贡献度是不同的,其主运动结点为右手肘、右手腕,即右手肘、右手腕对动作的完成具有高的贡献度,附加运动结点的贡献度相对较小。本文使用执行动作时不同身体部位的贡献度构建协同运动的结构特征模型。为了度量不同部位对完成动作的贡献,本文通过式(10)度量人体第 k 个部位 x、y、z 轴的贡献,其他结点贡献度的度量与此类似:

$$\begin{cases}p_{xk}=\lambda_1\sigma(x^k)+\lambda_2|\mu(x^k)|\\ p_{yk}=\lambda_1\sigma(y^k)+\lambda_2|\mu(y^k)|\\ p_{zk}=\lambda_1\sigma(z^k)+\lambda_2|\mu(z^k)|\end{cases} \tag{10}$$

以 p_{xk} 为例,p_{xk} 为人体第 k 个部位 x 轴的贡献度,$\sigma(x^k)$ 为人体第 k 个部位 x 轴运动数据的标准差,$|\mu(x^k)|$ 为人体第 k 个部位 x 轴运动数据均值的绝对值,λ_1、λ_2 为超参数用于平衡标准差和均值的绝对值。然后利用式(11)对贡献度进行归一化得到结构特征模型并用于约束特征:

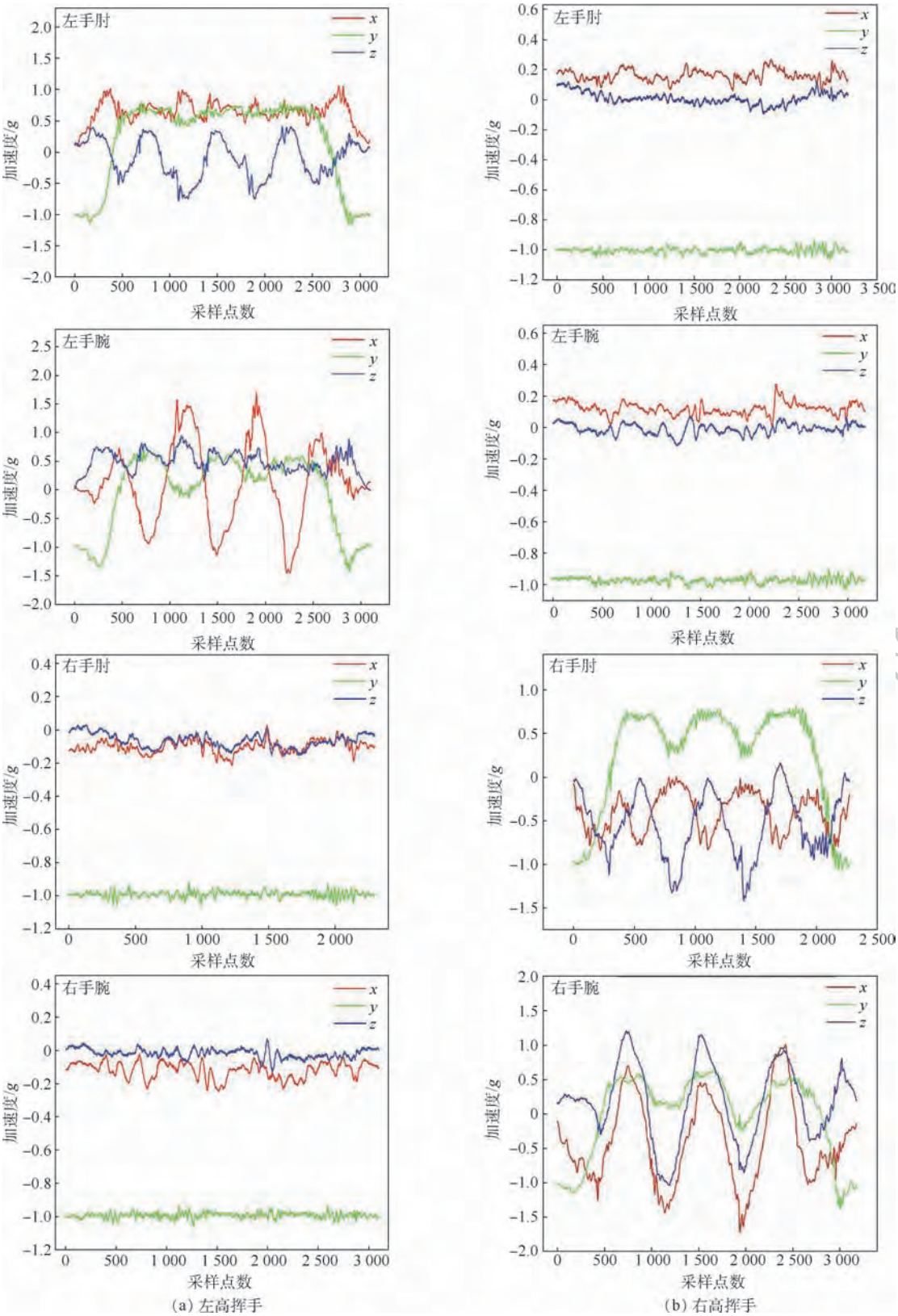


图 2 左高挥手、右高挥手动作主运动结点的三轴加速度曲线

Fig. 2 Triaxial acceleration curves of main motion nodes of left high wave and right high wave

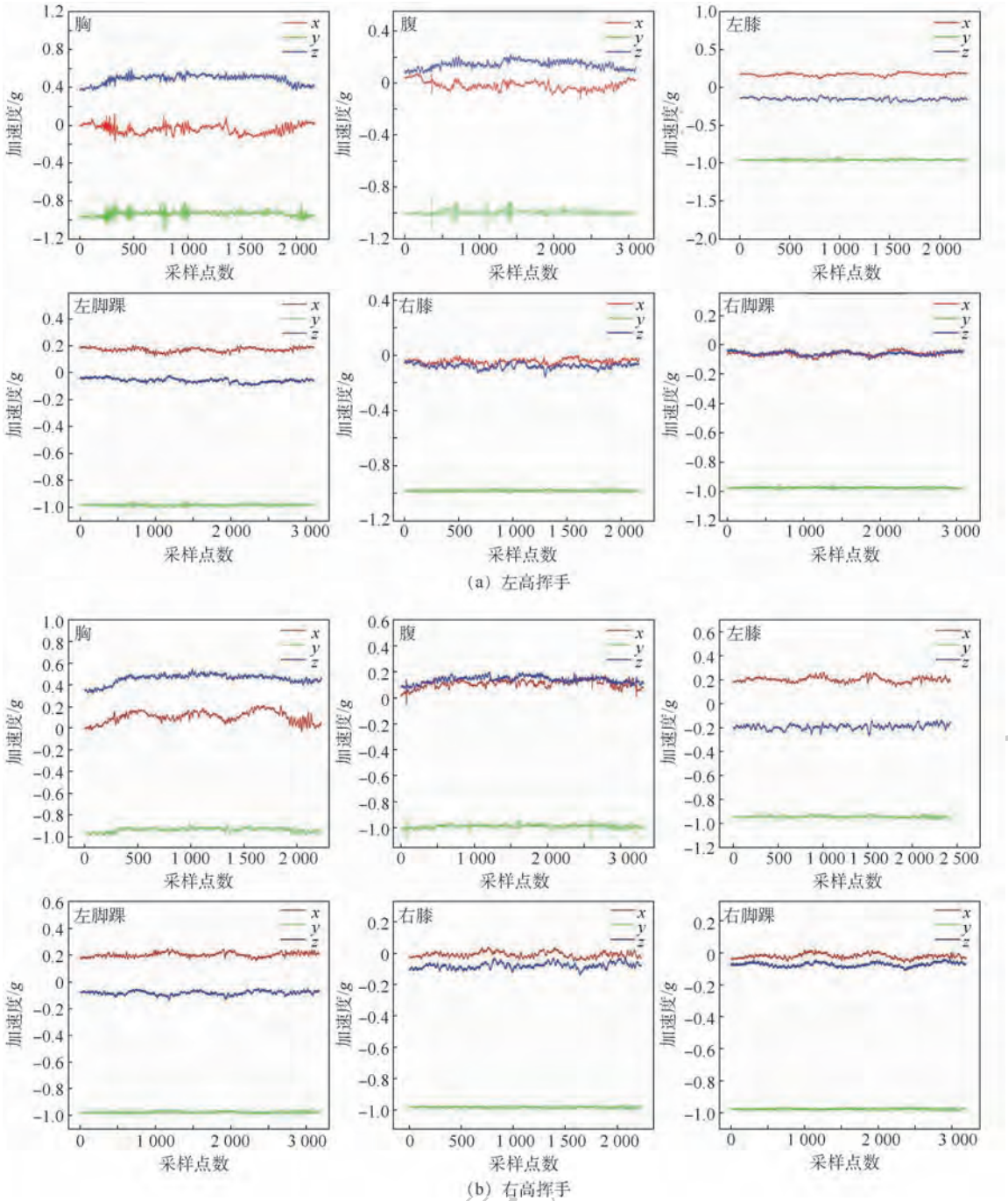


图 3　左高挥手、右高挥手动作附加运动结点的三轴加速度曲线

Fig. 3　Triaxial acceleration curves of additional motion nodes of left high wave and right high wave

$$\begin{cases} w_{xk} = \dfrac{p_{xk}}{\sum_{i=1}^{K} p_{xi}} \\ w_{yk} = \dfrac{p_{yk}}{\sum_{i=1}^{K} p_{yi}} \\ w_{zk} = \dfrac{p_{zk}}{\sum_{i=1}^{K} p_{zi}} \end{cases} \tag{11}$$

为了说明本文提出的结构特征模型的有效性,分别计算类间距离为 2 175、类内距离为10 383 以及 J_w 为 0.209 5。然后再计算使用结构特征模型约束后的特征的类间距离为 575、类内距离为 2 046以及 J_w^* 为 0.281 0。从计算结果可知 $J_w^* > J_w$,即使用结构特征模型约束后的特征具有更佳的分离能力。

3　多模态特征选择与特征融合

在采集人体的行为数据时一般会采集多个模态的数据。为了利用不同模态的数据进行行为识别,需要融合多模态特征。此外每个模态内包含

了大量的冗余特征,需要对特征进行选择。设 $\boldsymbol{\Gamma}=\{x_i^1,x_i^2,\cdots,x_i^M\}_{i=1}^N$ 表示 N 个动作样本,其中每个样本包含 M 个不同模态的特征。虽然第 i 个样本 $\boldsymbol{\Gamma}_i=\{x_i^1,x_i^2,\cdots,x_i^M\}$ 包含了 M 个不同模态的特征,但类别标签相同 $x_i^1,x_i^2,\cdots,x_i^M\rightarrow y_i$。本文为每个模态的特征分别学习投影矩阵。用于将不同模态的特征投影到子空间。并在投影的过程中完成对特征的选择。最后按式(12)进行多模态特征的融合:

$$\boldsymbol{f}=\begin{pmatrix}\boldsymbol{x}_1^*\\ \boldsymbol{x}_2^*\\ \vdots\\ \boldsymbol{x}_M^*\end{pmatrix}=\begin{pmatrix}\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{x}_1\\ \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{x}_2\\ \vdots\\ \boldsymbol{U}_M^{\mathrm{T}}\boldsymbol{x}_M\end{pmatrix} \tag{12}$$

式中:$\boldsymbol{f}$ 为融合后的特征;$\boldsymbol{x}_p$ 为第 p 个模态的特征,$p=1,2,\cdots,M$;$\boldsymbol{U}_p$ 为第 p 个模态的投影矩阵,$p=1,2,\cdots,M$;$\boldsymbol{x}_p^*$ 为第 p 个模态投影后的特征,$p=1,2,\cdots,M$。将融合后的特征称之为多模态融合特征(multi-modal fusion features)并用于分类识别。本文借鉴联合特征选择与子空间学习 JFSSL 方法[19]学习多模态投影矩阵,用于将不同模态的特征投影到子空间,并使用 $\ell_{2,1}$ 范数来实现特征选择[20],其最小化问题如下:

$$\min_{\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_M}\sum_{p=1}^{M}\|\boldsymbol{X}_p^{\mathrm{T}}\boldsymbol{U}_p-\boldsymbol{Y}\|_{\mathrm{F}}^2+\lambda_1\sum_{p=1}^{M}\|\boldsymbol{U}_p\|_{2,1}+\lambda_2\Omega(\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_M) \tag{13}$$

式中:$\boldsymbol{X}_p$ 为第 p 个模态的特征矩阵,$p=1,2,\cdots,M$;$\boldsymbol{Y}$ 为 JFSSL 方法构造的子空间,$\boldsymbol{Y}\in\mathbf{R}^{N\times c}$。式(13)的第 1 项用于学习投影矩阵,第 2 项用于特征选择,第 3 项用于保持模态内和模态间的相似关系。式(13)的第 3 项的推导为

$$\Omega(\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_M)=\frac{1}{2}\sum_{i=1}^{\hat{N}}\sum_{j=1}^{\hat{N}}\boldsymbol{W}_{ij}\|\boldsymbol{f}_i-\boldsymbol{f}_j\|^2=\mathrm{Tr}(\boldsymbol{F}\boldsymbol{L}\boldsymbol{F}^{\mathrm{T}})=\sum_{p=1}^{M}\sum_{q=1}^{M}\mathrm{Tr}(\boldsymbol{U}_p^{\mathrm{T}}\boldsymbol{X}_p\boldsymbol{L}_{pq}\boldsymbol{X}_q^{\mathrm{T}}\boldsymbol{U}_q) \tag{14}$$

其中:$\hat{N}$为所有模态样本总数;$\boldsymbol{W}_{ij}$为第 i 个样本与第 j 个样本的相似度;$\boldsymbol{f}_i$ $\boldsymbol{f}_j$ 分别为投影后的第 i 个样本和第 j 个样本;$\boldsymbol{L}=\boldsymbol{D}-\boldsymbol{W}$ 为拉普拉斯矩阵,其中 $\boldsymbol{D}$ 为度矩阵,$\boldsymbol{W}$ 为邻接矩阵;$\boldsymbol{F}=(\boldsymbol{F}_1^{\mathrm{T}},\boldsymbol{F}_2^{\mathrm{T}},\cdots,\boldsymbol{F}_M^{\mathrm{T}})=(\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{X}_1,\boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{X}_2,\cdots,\boldsymbol{U}_M^{\mathrm{T}}\boldsymbol{X}_M)$为所有模态特征矩阵投影到子空间后的特征矩阵。再将最小化问题式(13)中第 2 项进行推理:根据 Wang 等[21]的引理将第 2 项 $\sum_{p=1}^{M}\|\boldsymbol{U}_p\|_{2,1}$ 推导为 $\sum_{p=1}^{M}\mathrm{Tr}(\boldsymbol{U}_p^{\mathrm{T}}\boldsymbol{R}_p\boldsymbol{U}_p)$。最终将式(13)改写为

$$\min_{\boldsymbol{U}_1,\boldsymbol{U}_2,\cdots,\boldsymbol{U}_M}\sum_{p=1}^{M}\|\boldsymbol{X}_p^{\mathrm{T}}\boldsymbol{U}_p-\boldsymbol{Y}\|_{\mathrm{F}}^2+\lambda_1\sum_{p=1}^{M}\mathrm{Tr}(\boldsymbol{U}_p^{\mathrm{T}}\boldsymbol{R}_p\boldsymbol{U}_p)+\lambda_2\sum_{p=1}^{M}\sum_{q=1}^{M}\mathrm{Tr}(\boldsymbol{U}_p^{\mathrm{T}}\boldsymbol{X}_p\boldsymbol{L}_{pq}\boldsymbol{X}_q^{\mathrm{T}}\boldsymbol{U}_q) \tag{15}$$

其中:$\boldsymbol{R}_p=\mathrm{Diag}(\boldsymbol{r}_p)$,$\boldsymbol{r}_p^i=\frac{1}{2}\|\boldsymbol{u}_p^i\|_2$。对式(15)进行求导,得到投影矩阵迭代计算公式为

$$\boldsymbol{U}_p^{t+1}=\frac{\boldsymbol{X}_p\boldsymbol{Y}-\lambda_2\sum_{q\neq p}\boldsymbol{X}_p\boldsymbol{L}_{pq}\boldsymbol{X}_q^{\mathrm{T}}\boldsymbol{U}_q^t}{\boldsymbol{X}_p\boldsymbol{X}_p^{\mathrm{T}}+\lambda_1\boldsymbol{R}_p+\lambda_2\boldsymbol{X}_p\boldsymbol{L}_{pp}\boldsymbol{X}_p^{\mathrm{T}}} \tag{16}$$

本文使用式(16)迭代计算得到不同模态的投影矩阵,将不同模态的特征投影到子空间,再将子空间的特征按式(12)进行融合用于行为识别。

4 数据采集

本节介绍自建的数据采集系统,并将其用于全身运动数据的获取。

4.1 采样系统搭建

Kinect v2 传感器可以采集 RGB-D 图像以及人体 25 个关节点的位置数据。MPU9250 惯性传感器是一种体积小、功耗小的传感器,采样率为 500 Hz,可同时捕获三轴加速度、三轴角速度和三轴磁场强度数据并输出四元数等。

本文利用 Kinect 采集深度图像和关节点位置数据,利用 MPU9250 采集三轴加速度数据。为了采集人体全身部位的三轴加速度数据,本文搭建了包含 10 个惯性传感器的运动数据采集系统,采样系统架构如图 4 所示,其中 NTP(Network Time Protocol)服务器用于提供标准时间,无线 AP(wireless Access Point)用于组建无线局域网络。利用卡片式计算机树莓派控制 MPU9250 传感器,利用笔记本电脑控制Kinect v2,每次采集数据

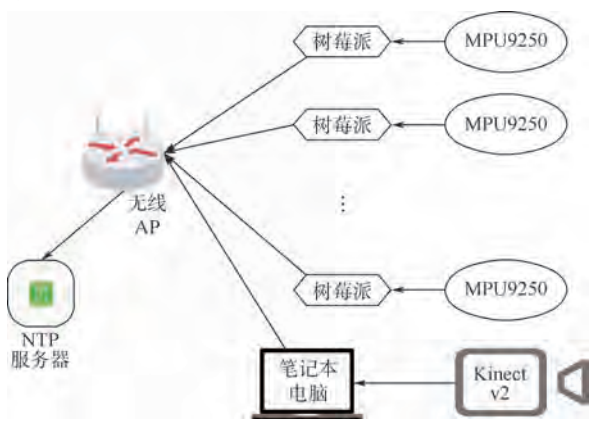


图 4 采样系统架构

Fig. 4 Sampling system architecture

时都与 NTP 服务器进行时间同步。本文综合考虑了 MHAD[22]、UTD-MHAD[23] 等行为数据库的采样方案后,确定惯性传感器的位置如图 5 所示,图中红色标记点为惯性传感器的位置,“左”为人体的左侧,“右”为人体的右侧。图 6 为本文的真实采样场景。

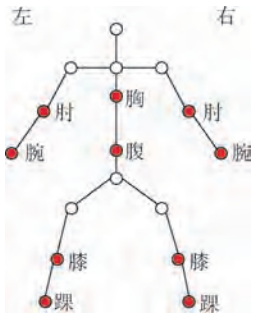


图 5　可穿戴传感器位置示意图

Fig. 5　Position schematic diagram of wearable sensor

图 6　真实采样场景

Fig. 6　Real sampling scene

4.2　数据库描述

本文自建的行为数据库包含 7 位男性受试者,年龄分布为 20 ~ 26 岁,其中包括一个肥胖型受试者、一个瘦弱型受试者。7 个男性受试者分别执行 15 个动作,每个动作重复执行 10 次。15 个动作类别如表 1 所示。

表 1　自建的行为数据库中的 15 个动作

Table 1　Fifteen actions in self-built behavior database

序号	类别
1	右高挥手
2	左高挥手
3	右水平挥手
4	左水平挥手
5	右手锤
6	右手抓
7	右手画叉
8	左手画叉
9	右手画圆
10	左手画圆
11	右脚前踢
12	左脚前踢
13	右脚侧踢
14	左脚侧踢
15	上下挥手

5　实验结果与分析

5.1　特征提取

本文在自建的数据库上进行实验,实验选取了 5 位受试者执行的 15 个行为动作,其中每个动作由每位受试者重复执行 8 次,实验用的行为数据库一共包含 600 个样本。对三轴加速度数据提取经典的时域特征:均值、方差、标准差、峰度、偏度。

5.2　实验设置

1) 设置 1。与 Chen 等[4] 类似,在第 1 组实验中将 3/8 样本作为训练,剩下数据作为测试;在第 2 组实验中将 4/8 样本作为训练;在第 3 组实验中将 5/8 样本作为训练;在第 4 组实验中将 6/8 样本作为训练。本文使用 T1 ~ T4 代表上述 4 组实验。

2) 设置 2。与 Chen 等[24] 类似,在第 1 组实验中将标记为 1、2 对象的数据作为训练样本;在第 2 组实验中将标记为 1、2、5 对象的数据作为训练样本;在第 3 组实验中将标记为 1、2、3、5 对象的数据作为训练样本。本文使用 T5 ~ T7 代表上述 3 组实验。

5.3　参数设置

本文的实验需要对分类器参数、超参数等进行选择,所需参数都通过 CV(Cross Validation)校验的方法进行确定。

1) 分类器参数设置。本文中使用的分类器如 CRC(Collaborative Respresentation Classifier)、KNN (K-Nearest Neighbor)、RandomF (Random Forest)等需要确定最佳参数。最终确定的分类器最优参数为:KNN 的参数设置为 1;RandomF 的参数设置为 65;CRC 的参数设置为 0.000 1。

2) 超参数设置。联合特征选择与子空间学习的参数 λ_1、λ_2 分别设置为 0.001 和 0.001;本文提出的结构特征模型中的超参数 λ_1、λ_2 分别设置为 5 和 0.05。

5.4　三轴加速度数据特征分类识别

本节对全身三轴加速度数据的特征以及使用结构特征模型约束后的特征进行分类识别。表 2 为使用人体全身的三轴加速度数据特征的识别率(即本节的基线实验),表 3 为使用结构特征模型约束后的特征的识别率。

从表 2 可知,封闭测试 T1 ~ T4 在使用判别分析、RandomF 分类器的情况下识别率很高,如表 2 中的 T4 实验使用判别分析分类器的识别率可高

达到 97.40%。然而表 2 中的开放测试 T5 ~ T7 的识别率均低于 T1 ~ T4 的识别率。因此本文的目的是提高开放测试 T5 ~ T7 的识别率。为此，将表 3 中的 T5 ~ T7 与表 2 中的 T5 ~ T7 的识别率进行对比，对比的结果如图 7 所示。

图 7 中每个分类器后的数字为特征约束后的识别率减去基线识别率的结果。从图 7 中可以直观看出，特征约束后 T5 ~ T7 的识别率明显高于基线识别率。说明本文利用基于人体空间协同运动的结构特征模型约束的特征具有更佳的分类

表 2 三轴加速度识别率

Table 2 Recognition rate of triaxial acceleration %

分类器	T1	T2	T3	T4	T5	T6	T7
KNN	39.64	41.85	43.84	45.90	31.11	33.33	40.83
判别分析	92.21	95.80	97.07	97.40	75.28	86.25	85.83
SVM	60.31	65.45	70.13	71.10	47.78	55.42	55.83
朴素贝叶斯	77.45	77.57	78.78	80.00	61.67	59.58	73.33
CRC	81.92	85.07	86.27	89.07	56.94	66.25	77.50
RandomF	94.01	95.15	95.96	96.50	86.39	88.75	86.67

表 3 三轴加速度约束后的识别率

Table 3 Recognition rate of constrained triaxial acceleration %

分类器	T1	T2	T3	T4	T5	T6	T7
KNN	59.95	66.47	65.33	68.80	50.00	51.25	56.67
判别分析	92.59	95.07	96.89	97.20	75.83	88.75	89.17
SVM	71.31	76.33	77.96	78.93	56.39	67.08	74.17
朴素贝叶斯	75.68	78.47	78.58	80.80	68.33	60.83	71.67
CRC	86.56	89.00	90.04	92.93	70.83	80.00	87.50
RandomF	94.83	95.40	96.18	96.53	84.56	85.75	89.17

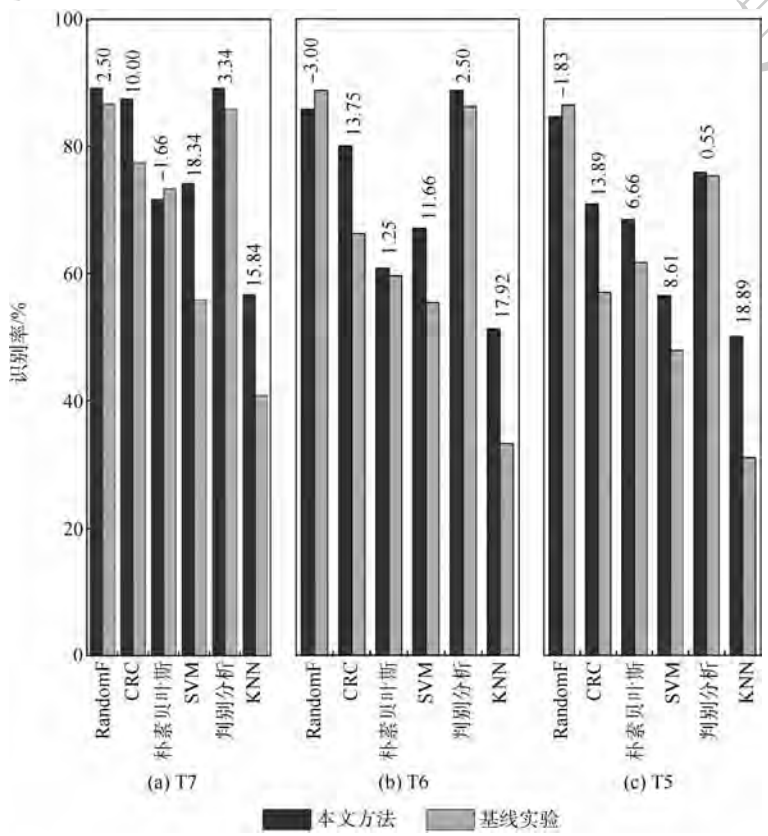


图 7 约束后的识别率与基线识别率的对比

Fig. 7 Comparison of recognition rate after constraints with baseline recognition rate

能力。

5.5　关节点位置数据的识别率

表 4 为关节点位置数据的识别率。由于 5.6 节实验为多模态特征的融合,故本节先计算关节点位置数据的识别率作为多模态特征融合的基线识别率。

5.6　多模态特征选择与特征融合的识别率

本节使用多模态特征选择与特征融合方法对三轴加速度数据特征、关节点位置数据特征进行融合并使用多个分类器进行分类识别,其识别结果如表 5 所示,并且将表 5 作为本节的基线实验,将表 5 与表 4、表 3 的 T1 ~ T7 识别率进行对比可知,多模态特征融合的识别率均优于单个模态特征的识别率。将本文结构特征模型约束后的特征与关节点位置数据特征进行融合,结果如表 6 所示。

表 6 与表 5 的 T1 ~ T4 相比,融合约束后的特征与关节点位置数据特征的识别率均显著高于基线实验。由于本文的重点是提高开放测试的识别率,因此将表 6 与表 5 的 T5 ~ T7 进行对比,对比结果如图 8 所示。

表 4　关节点位置识别率

Table 4　Recognition rate of joint point position　%

分类器	T1	T2	T3	T4	T5	T6	T7
KNN	41.11	42.45	43.33	45.70	37.22	41.25	36.67
判别分析	94.16	91.93	60.40	96.50	86.90	68.33	90.00
SVM	61.95	67.73	71.18	75.43	49.17	60.83	71.67
朴素贝叶斯	84.68	87.43	89.22	90.30	83.33	87.92	85.00
CRC	71.67	80.60	83.11	85.57	73.33	84.17	85.83
RandomF	96.70	97.32	97.33	98.00	91.94	97.92	98.33

表 5　多模态特征选择与特征融合的识别率

Table 5　Recognition rate of multi-modal feature selection and feature fusion　%

分类器	T1	T2	T3	T4	T5	T6	T7
KNN	77.31	84.25	88.27	89.23	89.72	92.08	88.33
判别分析	91.63	96.30	97.80	98.33	91.39	97.50	100.00
SVM	88.43	93.93	96.00	97.07	90.83	97.50	98.33
朴素贝叶斯	78.89	91.47	94.47	96.43	74.72	88.75	97.50
CRC	70.72	83.97	88.71	91.13	89.44	86.25	87.50
RandomF	88.85	93.52	94.84	95.93	82.71	88.40	97.04

表 6　融合约束后的三轴加速度数据特征与关节点位置数据特征识别率

Table 6　Recognition rate of triaxial acceleration feature and joint point position data feature after fusion constraints　%

分类器	T1	T2	T3	T4	T5	T6	T7
KNN	88.79	93.75	96.53	97.10	87.22	97.92	100.00
判别分析	89.84	96.07	97.80	98.17	90.83	98.33	100.00
SVM	89.07	94.88	96.51	97.90	88.33	98.33	100.00
朴素贝叶斯	80.09	90.65	94.96	95.83	79.44	95.42	100.00
CRC	86.85	93.75	96.16	97.50	87.22	97.08	100.00
RandomF	88.48	93.01	94.47	96.00	85.49	94.50	96.29

从图 8 中可以直观看出,在 T6、T7 实验中本文方法明显优于基线实验,在 T5 测试中本文方法略低于基线实验。这是由于 T5 测试选取 2/5 受试者的样本作为训练,选取的受试者数量少,从侧面说明本文提出的算法还可以进一步改良。此外表 6 中 T7 实验的识别率在 5 个分类器中高达 100.00%。上述的对比结果说明本文提出的基于人体空间协同运动的结构特征模型的有效性。

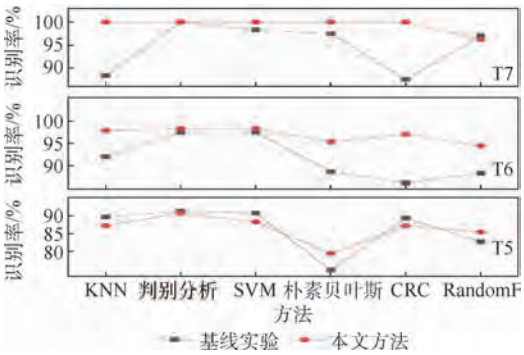


图 8 融合约束后的三轴加速度数据特征和关节点位置数据特征的识别率与基线识别率的对比

Fig. 8 Comparison of recognition rate of triaxial acceleration feature and joint point position data feature after fusion constraints with baseline

6 结 论

本文提出了利用人体不同部位三轴加速度数据的多个统计值用于度量不同部位对完成动作的贡献度,利用不同部位的贡献度构建面向行为识别的人体空间协同运动结构特征模型,并无监督、自适应地对人体不同部位的特征进行约束。在此基础上,借鉴 JFSSL 方法融合多模态的行为特征,并在融合过程中完成了对特征的筛选。实验结果表明:

1)该模型适用于具有全身三轴加速度数据的行为识别,模型的构建不需要复杂的算法计算,具有较好的实时性。

2)在自建的行为数据库的封闭测试(T1 ~ T4)、开放测试(T5 ~ T7)中均有优异的效果。

3)通过实验证明了人体在执行动作时不同部位之间存在协同性。这为进一步探索人体空间协同运动的结构特征提供了实验依据。

参考文献(References)

[1] LUVIZON D C, PICARD D, TABIA H. 2D/3D pose estimation and action recognition using multitask deep learning [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE Press, 2018: 5137-5146.

[2] FABIEN B, CHRISTIAN W, JULIEN M, et al. Glimpse clouds: Human activity recognition from unstructured feature points [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway, NJ: IEEE Press, 2018: 469-478.

[3] HOU R, CHEN C, MUBARAK S. Tube convolutional neural network (T-CNN) for action detection in videos [C]// Proceedings of the IEEE International Conference on Computer Vision (ICCV). Piscataway, NJ: IEEE Press, 2017: 5822-5831.

[4] CHEN C, JAFARI R, KEHTARNAVAZ N. Action recognition from depth sequences using depth motion maps-based local binary patterns [C]// IEEE Winter Conference on Applications of Computer Vision. Piscataway, NJ: IEEE Perss, 2015: 1092-1099.

[5] XIA L, CHEN C, AGGARWAL J K. View invariant human action recognition using histograms of 3d joints [C]// 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway, NJ: IEEE Press, 2012: 20-27.

[6] HUSSEIN M E, TORKI M, GOWAYYED M A. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations [C]// 23rd International Joint Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2013: 2466-2472.

[7] 苏本跃,蒋京,汤庆丰,等. 基于函数型数据分析方法的人体动态行为识别[J]. 自动化学报,2017,43(5):866-876.
SU B Y, JIANG J, TANG Q F, at al. Human dynamic action recognition based on functional data analysis [J]. Acta Automatica Sinica, 2017, 43(5): 866-876 (in Chinese).

[8] GRAVINA R, ALINIA P, GHASEMZADEH H, et al. Multi-sensor fusion in body sensor networks: State-of-the-art and research challenges [J]. Information Fusion, 2017, 35: 68-80.

[9] CHEN C, LIU K, KEHTARNAVAZ N. Real-time human action recogniti on based on depth motion maps [J]. Journal of Real-Time Image Processing, 2016, 12(1): 155-163.

[10] CHEN C, JAFARI R, KEHTARNAVAZ N. Improving human action recognition using fusion of depth camera and inertial sensors [J]. IEEE Transactions on Human-Machine Systems, 2015, 45(1): 51-61.

[11] HAGHIGHAT M, ABDEL-MOTTALEB M, ALHALABI W. Fully automatic face normalization and single sample face recognition in unconstrained environments [J]. Expert Systems with Applications, 2016, 47: 23-34.

[12] HAGHIGHAT M, ABDEL-MOTTALEB M, ALHALABI W. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition [J]. IEEE Transactions on Information Forensics and Security, 2016, 11(9): 1984-1996.

[13] 唐超,王文剑,李伟,等. 基于多学习器协同训练模型的人体行为识别方法[J]. 软件学报,2015,26(11):2939-2950.
TANG C, WANG W J, LI W, et al. Multi-learner co-training model for human actioin recognition [J]. Journal of Software, 2015, 26(11): 2939-2950 (in Chinese).

[14] SI C Y, JING Y, WANG W, et al. Skeleton-based action recognition with spatial reasoning and temporal stack learning [C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 103-118.

[15] LIU M Y, MENG F Y, CHEN C, et al. Joint dynamic pose image and space time reversal for human action recognition from videos [C]// 33rd AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2019: 8762-8769.

[16] 邓诗卓,王波涛,杨传贵,等. CNN 多位置穿戴式传感器人体活动识别[J]. 软件学报,2019,30(3):718-737.
DENG S Z, WANG B T, YANG C G, et al. Convolutional neural networks for human activity recognition using multi-location

wearable sensors[J]. Journal of Software, 2019, 30(3): 718-737(in Chinese).

[17] CHEN C, JAFARI R, KEHTARNAVAZ N. A real-time human action recognition system using depth and inertial sensor fusion [J]. IEEE Sensors Journal, 2016, 16(3): 773-781.

[18] 许艳,侯振杰,梁久祯,等. 权重融合深度图像与骨骼关键帧的行为识别[J]. 计算机辅助设计与图形学学报, 2018, 30(7): 139-146.
XU Y, HOU Z J, LIANG J Z, et al. Action recognition using weighted fusion of depth images and skeleton's key frames [J]. Journal of Computer-Aided Design & Computer Graphics, 2018, 30(7): 139-146(in Chinese).

[19] WANG K, HE R, WANG L, et al. Joint feature selection and subspace learning for cross-modal retrieval[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(10): 2010-2023.

[20] NIE F, HUANG H, CAI X, et al. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization[C]// Advances in Neural Information Processing Systems(NIPS). New York: Curran Associates, 2010: 1813-1821.

[21] HE R, TIAN T N, WANG L, et al. $L_{2,1}$ regularized correntropy for robust feature selection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR). Piscataway, NJ: IEEE Press, 2012: 2504-2511.

[22] OFLI F, CHAUDHRY R, KURILLO G, et al. Berkeley MHAD: A comprehensive multimodal human action database[C]// 2013 IEEE Workshop on Applications of Computer Vision (WACV). Piscataway, NJ: IEEE Press, 2013: 53-60.

[23] CHEN C, JAFARI R, KEHTARNAVAZ N. UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor[C]// 2015 IEEE International Conference on Image Processing (ICIP). Piscataway, NJ: IEEE Press, 2015: 168-172.

[24] CHEN C, ZHANG B G, HOU Z J, et al. Action recognition from depth sequences using weighted fusion of 2D and 3D auto-correlation of gradients features[J]. Multimedia Tools and Applications, 2017, 76(3): 4651-4669.

作者简介:

莫宇剑　男,硕士研究生。主要研究方向:机器学习、行为识别。

侯振杰　男,博士,教授,硕士生导师。主要研究方向:机器学习、人工智能、图像处理。

Structural feature representation and fusion of behavior recognition oriented human spatial cooperative motion

MO Yujian[1], HOU Zhenjie[1,*], CHANG Xingzhi[2], LIANG Jiuzhen[1], CHEN Chen[3], HUAN Juan[1]

(1. School of Information Science and Engineering, Changzhou University, Changzhou 213164, China;
2. Open Lab of Industrial Cloud for Intelligent Manufacturing, Changzhou College of Information Technology, Changzhou 213164, China;
3. Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, Charlotte 28223, USA)

Abstract: In view of the synergistic relationship among different parts of the body when human body performs actions, a behavior recognition method based on human body spatial cooperative motion structural features is proposed. Firstly, the contribution of different parts of the human body to the completion of the action is measured, and the contribution of different parts of the human body is transformed into a structural feature model of cooperative motion. Then, the model is used to constrain the motion characteristics of different parts of the body self-adaptively without supervision. On this basis, feature selection and multi-modal feature fusion are carried out using JFSSL, a cross-media retrieval method. The experiments show that the recognition rate of the open test is obviously improved by the proposed method on the self-built behavior database. At the same time, the calculation process of the method is simple and easy to implement.

Keywords: measure of contribution; cooperative motion; structural features; feature selection; multi-modal fusion

Received: 2019-07-09; **Accepted**: 2019-08-03; **Published online**: 2019-08-19 14:42

URL: kns.cnki.net/kcms/detail/11.2625.V.20190819.1417.001.html

Foundation items: National Natural Science Foundation of China (61063021, 61803050); Jiangsu Province Networking and Mobile Internet Technology Engineering Key Laboratory Open Research Fund Project (JSWLW-2017-013)

* **Corresponding author**. E-mail: houzj@cczu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0374

时空域上下文学习的视频多帧质量增强方法

佟骏超，吴熙林，丁丹丹*
（杭州师范大学 信息科学与工程学院，杭州 311121）

摘 要：卷积神经网络(CNN)在视频增强方向取得了巨大的成功。现有的视频增强方法主要在空域探索图像内像素的相关性，忽略了连续帧之间的时域相似性。针对上述问题，提出一种基于时空域上下文学习的多帧质量增强方法(STMVE)，即利用当前帧以及相邻多帧图像共同增强当前帧的质量。首先根据时域多帧图像直接预测得到当前帧的预测帧，然后利用预测帧对当前帧进行增强。其中，预测帧通过自适应可分离的卷积神经网络(ASCNN)得到；在后续增强中，设计了一种多帧卷积神经网络(MFCNN)，利用早期融合架构来挖掘当前帧及其预测帧的时空域相关性，最终得到增强的当前帧。实验结果表明，所提出的 STMVE 方法在量化参数值 37、32、27、22 上，相对于 H.265/HEVC，分别获得 0.47、0.43、0.38、0.28 dB 的性能增益；与多帧质量增强(MFQE)方法相比，平均获得 0.17 dB 的增益。

关 键 词：时空域上下文学习；多帧质量增强(MFQE)；卷积神经网络(CNN)；残差学习；预测帧
中图分类号：TP391
文献标识码：A **文章编号**：1001-5965(2019)12-2506-08

过去几年，视频逐渐成为互联网的主要流量，根据思科白皮书预测，到 2020 年，互联网有近 80%[1] 流量为视频。未经压缩的视频体积大，给传输和存储都带来巨大挑战。因此，原始视频一般都经过压缩再进行传输或存储。然而，视频压缩会带来压缩噪声，尤其在带宽严重受限的情况下，压缩噪声严重地影响了用户的主观体验。这时，有必要在解码端再次提升压缩视频的质量。

针对图像或视频质量增强，国内外已有不少研究。Dong 等[2] 设计了减少噪声的卷积神经网络(Artifacts Reduction Convolutional Neural Network，ARCNN)，减少了 JPEG 压缩图像所产生的噪声。之后，Zhang 等[3] 设计的去噪神经网络(Denoising Convolutional Neural Network，DnCNN)，具有较深的网络结构，实现了图像去噪，超分辨率和 JPEG 图像质量增强。后来，Yang 等[4] 设计了解码端可伸缩卷积神经网络(Decoder-side Scalable Convolutional Neural Network，DSCNN)，该结构由 2 个子网络组成，分别减少了帧内编码与帧间编码的失真。然而，上述方法都仅利用了图像的空域信息，没有利用相邻帧的时域信息，仍有提升空间。Yang 等[5] 尝试了一种多帧质量增强(Multi-Frame Quality Enhancement，MFQE)方法，利用空域的低质量当前帧与时域上的高质量相邻帧来增强当前帧。同等条件下，MFQE 获得了比空域单帧方法更好的性能。

在视频序列中，尽管帧之间具有相似性，但仍存在一定的运动误差。Yang 等[5] 所提出的 MFQE 做法是首先借助光流网络，得到相邻帧与当前待增强帧之间的光流场；然后根据该光流场

收稿日期：2019-07-09；**录用日期**：2019-08-12；**网络出版时间**：2019-08-22 13:05
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190821.1745.003.html
基金项目：浙江省自然科学基金（LY20F010013）；国家重点研发计划（2017YFB1002803）
* **通信作者**. E-mail：DandanDing@hznu.edu.cn

引用格式：佟骏超，吴熙林，丁丹丹. 时空域上下文学习的视频多帧质量增强方法[J]. 北京航空航天大学学报，2019，45(12)：2506-2513. TONG J C，WU X L，DING D D. Video multi-frame quality enhancement method via spatial-temporal context learning [J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2506-2513 (in Chinese).

对相邻帧进行运动补偿,即相邻帧内的像素点,根据光流信息,向当前帧对齐,得到对齐帧;最后,将对齐帧与当前帧一起送入后续的质量增强网络。上述方法能够取得显著增益,但也有一些不足:

1)视频帧之间的运动位移不一定恰好是整像素,有可能是亚像素位置,一般的做法是通过插值得到亚像素位置的像素值,不可避免地会产生一定误差。也就是说,根据光流信息进行帧间运动补偿的策略存在一定的缺陷。

2)MFQE 利用了当前帧前后各一帧图像对当前帧进行增强,增强网络对应的输入为 3 帧图像,包括当前帧与 2 个对齐帧。视频序列由连贯图像组成,推测,如果在时域采纳更多帧则会达到更好效果,这就意味着需要根据光流运动补偿产生更多对齐帧,神经网络的复杂度与参数量也会急剧上升,并不利于训练与实现。

考虑到上述问题,本文提出一种基于时空域上下文学习的多帧质量增强方法(STMVE),该方法不再从光流补偿,而是从预测的角度出发,根据时域多帧得到当前帧的预测帧,继而通过该预测帧来提升当前帧的质量。在预测时,在不增加网络参数与复杂度的情况下,充分利用了近距离低质量的 2 帧图像、远距离高质量的 2 帧图像,显著提升了性能。

本文的主要贡献如下:

1)在多帧关联性挖掘方面,与传统的基于光流进行运动补偿的方法不同,STMVE 方法根据当前帧的邻近帧,得到当前帧的预测帧。具体地,使用自适应可分离的卷积神经网络(Adaptive Separable Convolutional Neural Network,ASCNN)[6],输入时域邻近图像,通过自适应卷积与运动重采样,得到预测帧。该方法极大地缩短了预处理时间,并且预测图像的质量也得到了明显的改善。

2)在增强策略方面,提出多重预测的方式,充分利用当前帧的邻近 4 帧图像。将该 4 帧图像分为 2 类:近距离低质量的 2 帧图像与远距离高质量的 2 帧图像。这 2 类图像对当前帧的质量提升各有优势,设计神经网络结构并通过学习来结合其优势,获得更佳性能。

3)在多帧联合增强方面,提出了一种时空域上下文联合的多帧卷积神经网络(Multi-Frame CNN,MFCNN),该结构采用早期融合的方式,采用一层卷积层将时空域信息融合,而后通过迭代卷积不断增强。整个网络利用全局与局部残差结构,降低了训练难度。

1　相关工作

1.1　基于单帧的图像质量增强

近年来,在提升压缩图像质量方面涌现出大量工作。比如,Park 和 Kim[7]使用基于卷积神经网络(Convolutional Netural Network,CNN)的方法来替代 H.265/HEVC 的环路滤波。Jung 等[8]使用稀疏编码增强了 JPEG 压缩图像的质量。近年来,深度学习在提高压缩图像质量方面取得巨大成功。Dong 等[2]提出了一个 4 层的 ARCNN,明显提升了 JPEG 压缩图像的质量。利用 JPEG 压缩的先验知识以及基于稀疏的双域方法,Wang 等[9]提出了深度双域卷积神经网络(Deep Dual-domain Convolutional netural Network, DDCN),提高了 JPEG 图像的质量。Li 等[10]设计了一个 20 层的卷积神经网络来提高图像质量。最近,Lu 等[11]提出了深度卡尔曼滤波网络(Deep Kalman Filtering Network, DKFN)来减少压缩视频所产生的噪声。Dai 等[12]设计了一个基于可变滤波器大小的卷积神经网络(Variable-filter-size Residue-learning Convolutional Neural Network, VRCNN),进一步提高了 H.265/HEVC 压缩视频的质量,取得了一定性能。

上述工作设计了不同的方法,在图像内挖掘了像素之间的关联性,完成了空域单帧的质量增强。由于没有利用相邻帧之间的相似性,这些方法还有进一步提升空间。

1.2　基于多帧图像的超分辨率

基于多帧的超分辨与基于多帧的质量增强问题有相似之处。Tsai 等[13]提出了开创性的多帧图像的超分辨率工作,随后在文献[14]中得到了更进一步研究。同时,许多基于多帧的超分辨率方法都采用了基于深度神经网络的方法。例如,Huang 等[15]设计了一个双向递归卷积神经网络(Bidirectional Recurrent Convolution Network, BRCN),由于循环神经网络能够较好地对视频序列的时域相关性进行建模,获取了大量有用的时域信息,相较于单帧超分辨率方法,性能得到显著提升。Li 和 Wang[16]提出了一种运动补偿残差网络(Motion Compensation and Residual Net,MCRes-Net),首先使用光流法进行运动估计和运动补偿,然后设计了一个深度残差卷积神经网络结构进行图像质量增强。上述多帧方法所取得的性能超越了同时期的单帧方法。

Yang 等[5]提出利用时域和空域信息来完成质量增强任务,设计了一种 MFQE 的增强策略。

首先,利用光流网络产生光流信息,相邻帧在光流信息的引导下,得到与当前待增强帧处于同一时刻的对齐帧;然后,该对齐帧与当前待增强帧一同输入质量增强网络。受上述工作的启发,本文提出了一种更加精准和有效的基于多帧的方法,以进一步提升压缩视频的质量。

2 时空域上下文学习多帧质量增强

如图 1 所示,所提方法的整体结构包括预处理部分与质量增强网络。其中,预处理部分采用 ASCNN 网络,其输入相近多帧重建图像,分别生成关于当前帧的 2 个预测帧;然后,这 2 个预测帧与当前帧一起送入质量增强网络,经过卷积神经网络的非线性映射,得到增强后的当前帧。

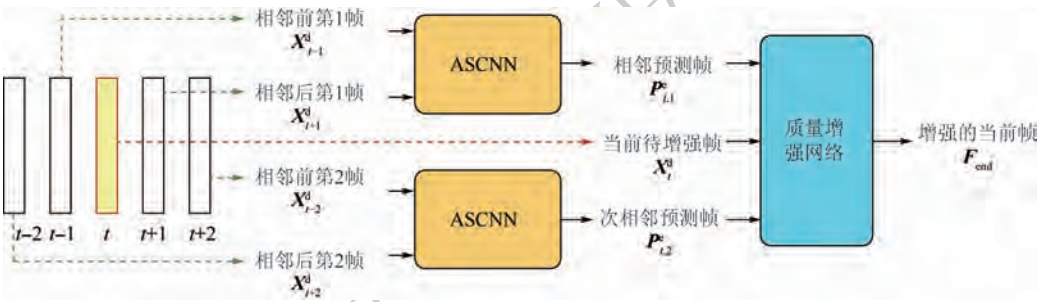


图 1 时空域上下文学习的多帧质量增强方法

Fig. 1 Approach for multi-frame quality enhancement using spatial-temporal context learning

1)使用光流法进行预处理

$\boldsymbol{f}_{(t-1)\to t}$表示 2 帧图像($\boldsymbol{X}^d_{t-1}$、$\boldsymbol{X}^d_t$, $t>1$)。首先,通过光流估计网络 H_{Flow} 得到的光流;然后,对得到光流与 $\boldsymbol{X}^d_{t-1}$ 进行 WARP 操作,得到对齐帧 $\boldsymbol{X}^a_t$;最后,将 $\boldsymbol{X}^a_t$ 和 $\boldsymbol{X}^d_t$ 一起送入卷积神经网络,可得到 t 时刻的质量增强帧。

$$\boldsymbol{f}_{(t-1)\to t} = H_{\text{Flow}}(\boldsymbol{X}^d_{t-1}, \boldsymbol{X}^d_t) \tag{1}$$

$$\boldsymbol{X}^a_t = \text{WARP}(\boldsymbol{f}_{(t-1)\to t}, \boldsymbol{X}^d_{t-1}) \tag{2}$$

2)使用 ASCNN 进行预处理

设 $\boldsymbol{X}^d_{t-1}$ 与 $\boldsymbol{X}^d_{t+1}$ 分别为 $t-1$ 和 $t+1$ 时刻已经解码得到的重建帧,将该 2 帧作为 ASCNN 的输入,得到 t 时刻的预测帧 $\boldsymbol{P}^e_{t,1}$;同理,将 $\boldsymbol{X}^d_{t-2}$ 和 $\boldsymbol{X}^d_{t+2}$ 输入 ASCNN,可以得到 t 时刻的另一个预测帧 $\boldsymbol{P}^e_{t,2}$,即

$$\boldsymbol{P}^e_{t,1} = \text{ASCNN}(\boldsymbol{X}^d_{t-1}, \boldsymbol{X}^d_{t+1}) \quad t>1 \tag{3}$$

$$\boldsymbol{P}^e_{t,2} = \text{ASCNN}(\boldsymbol{X}^d_{t-2}, \boldsymbol{X}^d_{t+2}) \quad t>2 \tag{4}$$

光流法的典型实现是 FlowNet[17]。本文选取了 4 个测试序列,每个序列测试 50 帧,比较了 FlowNet 2.0 与 ASCNN 的时间复杂度,如表 1 所示。对于 2 个网络,分别输入 2 帧图像,并得到各自的预处理时间。经过预处理,光流法得到 2 帧对齐图像,ASCNN 得到 1 帧预测帧。值得注意的是,由于 FlowNet 2.0 网络参数量较大,当显存不足时,每帧图像被分成多块进行处理。从表 1 可以看出,FlowNet 2.0 的预处理耗时约为 ASCNN 的 10 倍。

此外,本文对比了上述 2 种预处理之后分别得到的对齐帧 $\boldsymbol{X}^a_t$ 与预测帧 $\boldsymbol{P}^e_t$ 的主观质量,如图 2所示(POC 表示播放顺序,PSNR 表示峰值信噪比),选取了测试序列 BasketballDrill 的第 4、5、6 帧来进行说明。图 2(a)为 FlowNet 2.0,POC 4 向 POC 5 对齐,得到对齐帧;图 2(b)为 FlowNet 2.0,POC 6 向 POC 5 对齐,得到对齐帧;图 2(c)为 ASCNN,POC 4 和 POC 6 预测,得到预测帧。由图 2(a)、(b)可见,光流法生成的对齐帧中,“篮球”产生了重影,这是由光流信息不精准造成的,而 ASCNN 的预测帧弥补了这一缺陷。从其客观指标来看,ASCNN 得到的预测帧的 PSNR 也较高。

表 1 光流法(FlowNet 2.0)与 ASCNN 预处理时间对比

Table 1 Pre-processing time comparison of optical flow method (FlowNet 2.0) and ASCNN

序列	分辨率	预处理时间/s	
		FlowNet 2.0	ASCNN
BQSquare	416×240	255	108
PartyScene	832×480	1 085	111
BQMall	832×480	1 144	104
Johnny	1 280×720	2 203	116
平均耗时		1 172	110

2.1 光流法与 ASCNN

挖掘多帧之间关联性的关键是对多帧之间的运动误差进行补偿。光流法是一种常见的方法。本文对光流法与基于预测的 ASCNN 的计算复杂度与性能进行了对比。

2.2 多帧质量增强网络

多帧质量增强网络 MFCNN 的结构如图 3 所示,网络结构内部的参数配置如表 2 所示。在 MFCNN 中,H_{CFEN} 为粗特征提取网络(Coarse

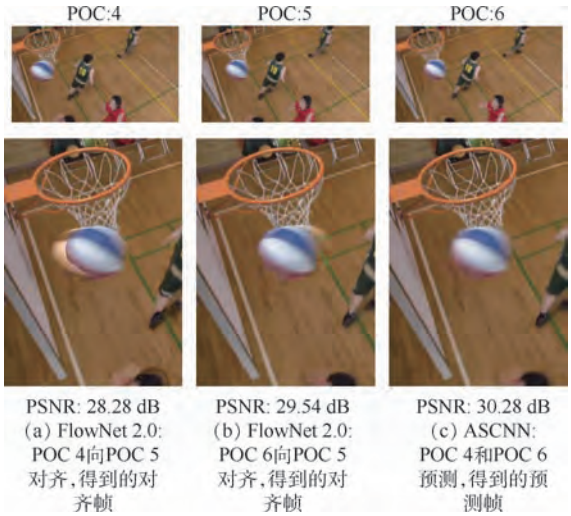


图 2　光流法(FlowNet 2.0)与 ASCNN 预处理得到的输出图像的主观图

Fig. 2　Subjective quality comparison of output image preprocessed by optical flow method (FlowNet 2.0) and ASCNN

Feature Extraction Network,CFEN),分别用于提取 $\boldsymbol{P}^{\mathrm{e}}_{t,1}$、$\boldsymbol{X}^{\mathrm{d}}_{t}$ 和 $\boldsymbol{P}^{\mathrm{e}}_{t,2}$的空间特征:

$$\boldsymbol{F}_{t,-1} = H_{\mathrm{CFEN}}(\boldsymbol{P}^{\mathrm{e}}_{t,1}) \tag{5}$$

$$\boldsymbol{F}_{t,0} = H_{\mathrm{CFEN}}(\boldsymbol{X}^{\mathrm{d}}_{t}) \tag{6}$$

$$\boldsymbol{F}_{t,1} = H_{\mathrm{CFEN}}(\boldsymbol{P}^{\mathrm{e}}_{t,2}) \tag{7}$$

为进一步利用 $\boldsymbol{P}^{\mathrm{e}}_{t,1}$和 $\boldsymbol{P}^{\mathrm{e}}_{t,2}$的时域特征信息,将 $\boldsymbol{F}_{t,-1}$、$\boldsymbol{F}_{t,0}$和 $\boldsymbol{F}_{t,1}$特征级联,得

$$\boldsymbol{F}_{1} = \mathrm{Concat}(\boldsymbol{F}_{t,-1},\boldsymbol{F}_{t,0},\boldsymbol{F}_{t,1}) \tag{8}$$

然后,送入 Conv 4,将所级联的特征进一步融合,同时使用 1 × 1 大小的卷积核来降该层的参数量:

$$\boldsymbol{F}_{4} = \mathrm{Conv}\ 4_{1\times1}(\boldsymbol{F}_{1}) \tag{9}$$

最后,经过 7 个残差网络[18]的残差块,得到特征矩阵为 $\boldsymbol{F}_7$。在 MFCNN 的最后输出层,加入 $\boldsymbol{X}^{\mathrm{d}}_{t}$ 形成全局残差学习结构:

$$\boldsymbol{F}_{\mathrm{end}} = \mathrm{ResB}_7(\boldsymbol{F}_7 + \boldsymbol{X}^{\mathrm{d}}_{t}) \tag{10}$$

在训练该网络时,对于给定的一个训练集 $\{(\boldsymbol{P}^{\mathrm{e},(i)}_{t,1},\boldsymbol{X}^{\mathrm{d},(i)}_{t},\boldsymbol{P}^{\mathrm{e},(i)}_{t,2}),\boldsymbol{Y}^{(i)}\}_{i=1}^{N}$,$N$ 为训练集中训练数据的个数,$\boldsymbol{Y}^{(i)}$为原始图像,定义损失函数:

$$L(\boldsymbol{\theta}) = \frac{1}{2N}\sum_{i=1}^{N}\left\|\boldsymbol{Y}^{(i)} - \mathrm{QE}(\boldsymbol{P}^{\mathrm{e},(i)}_{t,1},\boldsymbol{X}^{\mathrm{d},(i)}_{t},\boldsymbol{P}^{\mathrm{e},(i)}_{t,2})\right\|^2 \tag{11}$$

式中:$\boldsymbol{\theta}$ 为权重参数的集合;QE($\boldsymbol{P}^{\mathrm{e},(i)}_{t,1}$,$\boldsymbol{X}^{\mathrm{d},(i)}_{t}$,$\boldsymbol{P}^{\mathrm{e},(i)}_{t,2}$)为所获得的质量增强之后的图像。训练中,通过小批量梯度下降算法和反向传播来优化目标方程。

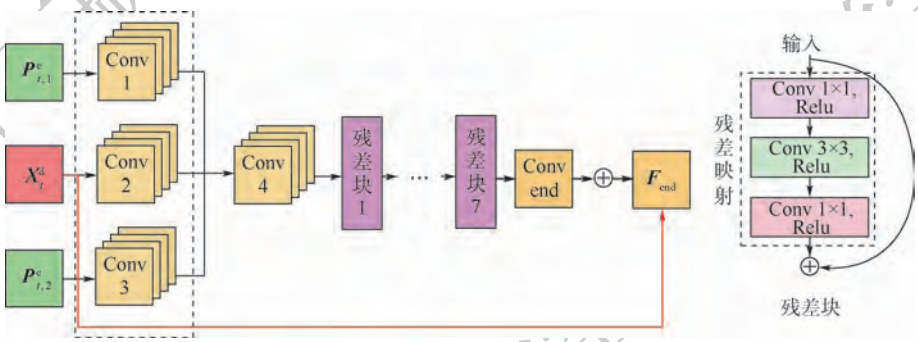


图 3　早期融合网络结构及其内部每个残差块的结构

Fig. 3　Structure of proposed early fusion network and structure of each residual block in it

表 2　多帧质量增强网络结构

Table 2　Structure of proposed quality enhancement network

卷积层	滤波器大小	滤波器数量	步长	激活函数
Conv 1/2/3	3 × 3	64	1	Relu
Conv 4	1 × 1	64	1	Relu
	1 × 1	64	1	Relu
残差块 ×7	3 × 3	64	1	Relu
	1 × 1	64	1	Relu
Conv end	5 × 5	1	1	

3　实验结果与分析

3.1　实验条件

本文的训练与测试环境为 i7-8700K CPU 和 Nvidia GeForce GTX 1080 TI GPU。所有实验都基于 TensorFlow 深度学习框架。本文使用 118 个视频序列来训练神经网络,并在 11 个 H.265/HEVC 标准测试序列进行测试。每个序列都使用 HM16.9 在 Random-Access(RA)配置下图像组(Group of Pictures,GOP)大小设置为 8,量化参数分别设置为 22、27、32、37,并获得重建视频序列。

3.2　实验结果比较与分析

1)与传统光流法的对比

本文采用基于早期融合的质量增强网络结构,对 5 种预处理方法的性能进行了对比,以图 4 所示的第 2 帧图像为例,包括:

① 使用光流法对前后 $t \pm 2$ 帧进行运动补

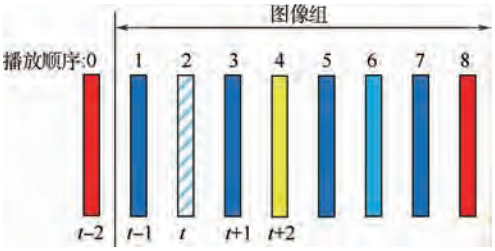


图 4 以图像组为单位对低质量图像进行增强

Fig. 4 Enhancing low-quality images for each GOP

偿,分别得到对齐帧,利用两帧对齐帧对当前帧进行增强。

② 使用 ASCNN 利用前后 $t\pm2$ 帧预测当前帧,利用该预测帧对当前帧进行增强。

③ 结合①与②,利用两帧对齐帧与一帧预测帧对当前帧进行增强。

④ 使用光流法对前后 $t\pm2$ 帧进行运动补偿,分别得到对齐帧,使用 ASCNN 利用前后 $t\pm1$ 帧预测当前帧,利用两帧对齐帧与一帧预测帧对当前帧进行增强。

⑤ 使用 ASCNN,分别利用前后 $t\pm1$ 与 $t\pm2$ 帧预测当前帧,根据所得到的两帧预测帧对当前帧进行增强。

表 3 给出了上述几种方法的性能对比,其中 $t+n$ 代表与 t 时刻相隔 n 帧。可见,使用 ASCNN,将 2 对相邻帧生成当前时刻的 2 个预测帧的方式,取得了最优的性能。

表 3 5 种预处理方式所获得的 PSNR 性能指标对比

Table 3 PSNR performance indicator comparison by five pre-processing strategies dB

序列	H.265/HEVC	FlowNet 2.0 $(t-2,t+2)$	ASCNN $(t-2,t+2)$	FlowNet 2.0$(t-2,t+2)$ + ASCNN $(t-2,t+2)$	FlowNet 2.0$(t-2,t+2)$ + ASCNN $(t-1,t+1)$	ASCNN$(t-2,t+2)$ + ASCNN$(t-1,t+1)$
BQMall	31.00	31.35	31.38	31.20	31.23	**31.46**
BasketballDrill	31.94	32.35	32.32	32.16	32.19	**32.39**
FourPeople	35.59	36.23	36.20	36.00	36.04	**36.32**
BQSquare	29.21	29.59	29.62	29.32	29.35	**29.65**
平均值	31.94	32.38	32.38	32.17	32.20	**32.46**

2)与不同网络结构的对比

本文设计了多种形态的网络结构,并对其性能进行了对比。如图 5 所示,分别设计了直接融合、渐进融合 2 种方式,并与本文的早期融合进行了对比。

3 种结构的区别在于,直接融合方法直接将多帧信息级联作为网络输入,渐进融合方法逐渐地级联卷积特征图。将这 3 种结构设计成具有相近的参数量,实验结果如表 4 所示。可见,渐进融合比直接融合的性能平均提高了 0.03 dB,而早期融合比渐进融合又能够提升 0.04 dB。

该实验证明了对每个输入帧使用更多独立滤波器,可以更好地甄别当前帧与预测帧的重要性。但随着网络深度的增加,渐进融合方法会引入更多参数。因此,相同等参数量的情况下,与早期融合方法相比,渐进融合网络深度较浅,很可能产生欠拟合问题,无法达到同等性能。

3)与单帧质量增强方法的对比

采用 ASCNN,分别对前后帧进行预测,具体地,分别使用 ASCNN 利用前后距离近质量低的两帧图像(如图 4 的第 1 帧与第 3 帧)、前后距离远质量高的两帧图像(如图 4 的第 0 帧与第 4 帧),得到当前帧的两帧预测帧,这两帧预测帧与当前帧一起送入所提出的早期融合网络,得到最终增强的图像。

如表 5 所示,本文所提出的 STMVE 始终优于仅使用空域信息的单帧质量增强方法。具体地,相较于 H.265/HEVC,STMVE 在量化参数为 37、32、27、22 分别取得了 0.47、0.43、0.38、0.28 dB 的增益,相较于单帧质量增强方法,分别获得

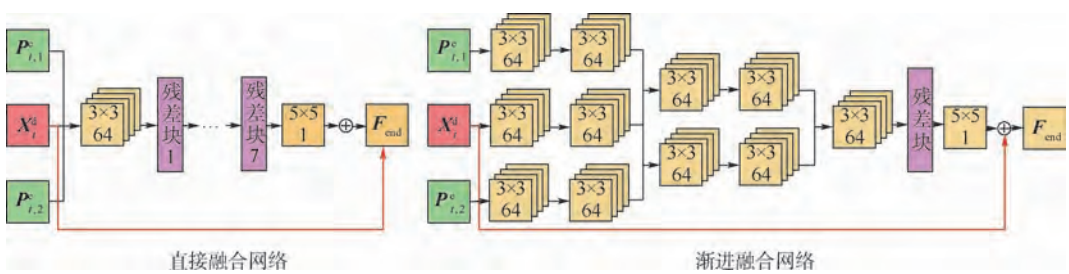


图 5 直接融合网络和渐进融合网络与所提出的早期融合网络的对比

Fig. 5 Comparison of direct fusion networks and slow fusion networks with proposed early fusion networks

表 4　三种网络结构的 PSNR 性能指标对比

Table 4　PSNR performance indicator comparison of three network structures　dB

测试序列	直接融合	渐进融合	早期融合
BQMall	31.39	31.42	**31.46**
BasketballDrill	32.32	32.35	**32.39**
RaceHousesC	29.36	29.38	**29.41**
平均值	31.02	31.05	**31.09**

0.16、0.15、0.15 和 0.11 dB 的性能增益。

4）与多帧质量增强方法的对比

本文也与 MFQE 的结果进行了对比,结果如表 6 所示,其中,ΔPSNR 代表 STMVE 与 MFQE 的 PSNR 之差。随机选取了 4 个测试序列,在量化参数为 37 时,测试其前 36 帧。结果表明,所提出的 STMVE 方法平均比 MFQE 高出 0.17 dB。在参数数量上,MFQE的参数量约为1 715 360,而

表 5　不同方法的 PSNR 性能指标对比

Table 5　Comparison of PSNR performance indicator among different methods　dB

量化参数	类别	测试序列	H.265/HEVC	单帧质量增强	STMVE 方法	相对单帧的提升	相对 H.265/HEVC 的提升
37	C	BasketballDrill	31.94	32.14	**32.39**	0.25	0.45
		BQMall	31.00	31.17	**31.46**	0.29	0.46
		PartyScene	27.73	27.73	**27.94**	0.21	0.21
		RaceHorses	29.08	29.23	**29.41**	0.18	0.33
	D	BasketballPass	31.79	32.02	**32.37**	0.35	0.58
		BlowingBubbles	29.19	29.30	**29.51**	0.21	0.32
		BQSquare	29.21	29.28	**29.65**	0.37	0.44
		RaceHorses	28.69	28.93	**29.18**	0.25	0.49
	E	FourPeople	35.59	36.03	**36.32**	0.29	0.73
		Johnny	37.34	37.61	**37.80**	0.19	0.46
		KristenAndSara	36.77	37.21	**37.43**	0.22	0.66
		平均值	31.67	31.97	**32.13**	0.16	0.47
32		平均值	34.31	34.59	**34.74**	0.15	0.43
27		平均值	37.06	37.28	**37.43**	0.15	0.38
22		平均值	39.89	40.06	**40.17**	0.11	0.28

表 6　STMVE 方法与 MFQE 的 PSNR 性能指标对比

Table 6　PSNR performance indicator comparison between proposed method and MFQE　dB

测试序列(36 帧)	MFQE	STMVE 方法	ΔPSNR
PartyScene	26.95	**27.39**	0.44
BQMall	30.39	**30.64**	0.25
Johnny	36.84	**36.87**	0.03
BlowingBubbles	28.97	**28.93**	0.04
平均值	30.79	**30.96**	0.17

STMVE 的参数量为 362 176,仅为 MFQE 的 21%。可见,所提出的网络虽然具有较少参数,但仍获得了较高性能。

5）主观质量对比

本文还比较了经不同方法处理后得到的图像的主观质量,如图 6 所示。经观察可见,与 H.265/HEVC和单帧质量增强方法相比,所提出的STMVE方法能够明显改善图像的主观质量,

图 6　不同方法获得图像的主观质量对比

Fig. 6　Subjective quality comparison of reconstructed pictures enhanced by different methods

图像的细节被更好地保留下来,主观质量提升明显。

4 结 论

本文提出了一种时空域上下文学习的多帧质量增强方法基于 STMVE。与以往基于单帧质量增强的方法不同,STMVE 方法充分利用了当前帧的邻近 4 帧图像的时域信息。与传统的基于光流法的运动补偿方式不同,本文提出了利用预测帧增强当前帧的质量;为充分挖掘时域信息,提出了多帧增强的早期渐进融合式网络结构。其次,针对所提出的 STMVE 方法,分别就预处理方式、网络组合结构、质量增强方法及主观质量进行了分析,并设计实验与以往方法进行了对比。大量的实验结果表明,与其他方法相比,本文所提出的 STMVE 方法在主观质量与客观质量上都有显著优势。

参考文献(References)

[1] CISCO. Cisco visual networking index: Global mobile data traffic forecast update[EB/OL]. (2019-02-18)[2019-07-08]. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-738429.html.

[2] DONG C, DENG Y, CHANGE LOY C, et al. Compression artifacts reduction by a deep convolutional network[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2015: 576-584.

[3] ZHANG K, ZUO W, CHEN Y, et al. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising[J]. IEEE Transactions on Image Processing, 2017, 26(7): 3142-3155.

[4] YANG R, XU M, WANG Z. Decoder-side HEVC quality enhancement with scalable convolutional neural network[C]// 2017 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ: IEEE Press, 2017: 817-822.

[5] YANG R, XU M, WANG Z, et al. Multi-frame quality enhancement for compressed video[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2018: 6664-6673.

[6] NIKLAUS S, MAI L, LIU F. Video frame interpolation via adaptive separable convolution[C]// Proceedings of the IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2017: 261-270.

[7] PARK W S, KIM M. CNN-based in-loop filtering for coding efficiency improvement[C]// 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), 2016: 1-5.

[8] JUNG C, JIAO L, QI H, et al. Image deblocking via sparse representation[J]. Signal Processing: Image Communication, 2012, 27(6): 663-677.

[9] WANG Z, LIU D, CHANG S, et al. D3: Deep dual-domain based fast restoration of jpeg-compressed images[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 2764-2772.

[10] LI K, BARE B, YAN B. An efficient deep convolutional neural networks model for compressed image deblocking[C]// 2017 IEEE International Conference on Multimedia and Expo (ICME). Piscataway, NJ: IEEE Press, 2017: 1320-1325.

[11] LU G, OUYANG W, XU D, et al. Deep Kalman filtering network for video compression artifact reduction[C]// Proceedings of the European Conference on Computer Vision (ECCV). Berlin: Springer, 2018: 568-584.

[12] DAI Y, LIU D, WU F. A convolutional neural network approach for post-processing in HEVC intra coding[C]// International Conference on Multimedia Modeling. Berlin: Springer, 2017: 28-39.

[13] TSAI R. Multiframe image restoration and registration[J]. Advance Computer Visual and Image Processing, 1984, 11(2): 317-339.

[14] PARK S C, PARK M K, KANG M G. Super-resolution image reconstruction: A technical overview[J]. IEEE Signal Processing Magazine, 2003, 20(3): 21-36.

[15] HUANG Y, WANG W, WANG L. Video super-resolution via bidirectional recurrent convolutional networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 1015-1028.

[16] LI D, WANG Z. Video superresolution via motion compensation and deep residual learning[J]. IEEE Transactions on Computational Imaging, 2017, 3(4): 749-762.

[17] ILG E, MAYER N, SAIKIA T, et al. FlowNet 2.0: Evolution of optical flow estimation with deep networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2017: 2462-2470.

[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2016: 770-778.

作者简介:

佟骏超 男,硕士研究生。主要研究方向:视频图像处理、视频编码。

丁丹丹 女,博士,讲师,硕士生导师。主要研究方向:视频编码、视频图像处理。

Video multi-frame quality enhancement method via spatial-temporal context learning

TONG Junchao, WU Xilin, DING Dandan*

(School of Information Science and Engineering, Hangzhou Normal University, Hangzhou 311121, China)

Abstract: Convolutional neural network (CNN) has achieved great success in the field of video enhancement. The existing video enhancement methods mainly explore the pixel correlations in spatial domain of an image, which ignores the temporal similarity between consecutive frames. To address the above issue, this paper proposes a multi-frame quality enhancement method, namely spatial-temporal multi-frame video enhancement (STMVE), through learning the spatial-temporal context of current frame. The basic idea of STMVE is utilizing the adjacent frames of current frame to help enhance the quality of current frame. To this end, the virtual frames of current frame are first predicted from its neighbouring frames and then current frame is enhanced by its virtual frames. And the adaptive separable convolutional neural network (ASCNN) is employed to generate the virtual frame. In the subsequent enhancement stage, a multi-frame CNN (MFCNN) is designed. An early-fusion CNN structure is developed to extract both temporal and spatial correlation between the current and virtual frames and output the enhanced current frame. The experimental results show that the proposed STMVE method obtains 0.47 dB, 0.43 dB, 0.38 dB and 0.28 dB PSNR gains compared with H.265/HEVC at quantized parameter values 37, 32, 27 and 22 respectively. Compared to the multi-frame quality enhancement (MFQE) method, an average 0.17 dB PSNR gain is obtained.

Keywords: spatial-temporal context learning; multi-frame quality enhancement (MFQE); convolutional neural network (CNN); residual learning; virtual frame

Received: 2019-07-09; **Accepted**: 2019-08-12; **Published online**: 2019-08-22 13:05

URL: kns.cnki.net/kcms/detail/11.2625.V.20190821.1745.003.html

Foundation items: Natural Science Foundation of Zhejiang Province, China (LY20F010013); National Key R & D Program of China (2017YFB1002803)

* **Corresponding author**. E-mail: DandanDing@hznu.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0375

中国国画艺术美感特征分析与分类

湛颖，高妍，谢凌云*
（中国传媒大学 媒介音视频教育部重点实验室，北京 100024）

摘 要：图像艺术美感自动分类是近年的热门研究领域，国画作为中国传统艺术文化的重要体现，其美感也极具研究价值。在5类美感标注的国画数据库基础上，进行了国画艺术美感自动分类研究和相关特征分析。经过特征提取和筛选，得到适用于美感分类的33个图像特征，并基于特征重要性建立了物理特征与艺术美感、美术技法之间的映射关系。同时使用该特征集在多种分类器上进行艺术美感自动识别，验证了国画艺术美感自动分类的可行性。结果表明，国画艺术美感分类的主要相关美术元素按重要性排序为：颜色、笔触、亮度和线条。

关 键 词：美感分类；美感特征；国画；特征选择；图像分类
中图分类号：TP391.41
文献标识码：A **文章编号**：1001-5965(2019)12-2514-09

图像美感自动分析与识别是近年兴起的研究热点之一。较早的图像模式识别研究集中在目标识别任务（如人脸识别）等层面，这些研究关注物体的底层视觉特征，通过边界、形状等美术元素的统计关系构建客观世界与机器学习的桥梁。随着目标识别应用的普及，业界对图像分类的需求不再满足于基本的目标识别，而是转向情感、审美等方面。尤其随着人工智能研究的发展，如何利用计算机分析和学习人类主观审美感知，以之作为人工智能的重要辅助功能，也吸引了众多研究人员的关注。

本文的主要贡献如下：①以中国国画的美感评价与分类为研究对象，建立了一个用于国画视觉美感分析的数据库，包含超过500张各种风格的国画图片，所有画作包含5种美感风格和总体美感强度的量化标注；②基于该数据库，筛选了适合国画美感评价的若干特征，经过对多种主流分类算法的综合评析，初步搭建了国画美感评价与分类框架；③分析了客观特征与美感相关的美术元素之间的映射关系，尝试给出影响国画美感自动分类效果的美学因素解释。

1 相关工作

西方实验美学领域率先对美感展开了一系列定量研究。Joshi 和 Datta 建立了一个美感-情感图像数据库，对图像进行了10个因子8个强度等级的美感标注[1]；Murray 等建立了一个大型美感数据库，对255000张图片进行了美感的语义和强度标注，它的特点是每张图所标注的因子都不同[2]；Luo 等对上万张图片做了二元美感标注（binary aesthetic labels），并按场景将它们分为7类[3]；Li 和 Chen 设计了一系列全局、局部特征，进行美术作品图像美感的二元分类[4]。近两年也有更多基于艺术内容（如油画）的数据库出现，如 WikiArt Emotions[5] 和 The Rijksmuseum Challenge[6] 等。

基于以上美感数据库，也出现了许多对于图

收稿日期：2019-07-09；**录用日期**：2019-08-19；**网络出版时间**：2019-08-27 10:34
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190827.0924.002.html
基金项目：中央高校基本科研业务费专项资金（18CUCTJ086）
* **通信作者**. E-mail：xiely@cuc.edu.cn

引用格式：湛颖，高妍，谢凌云. 中国国画艺术美感特征分析与分类[J]. 北京航空航天大学学报，2019，45(12)：2514-2522.
ZHAN Y，GAO Y，XIE L Y. Aesthetic feature analysis and classification of Chinese traditional painting[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2514-2522 (in Chinese).

像质量和美感的分类研究。Ke 等基于边界空间分布、颜色分布、色调计算和模糊度等高层特征,并加上底层视觉特征,提出了采用朴素贝叶斯分类器(naive Bayes classifier)的摄影美感评价方法[7];Luo 和 Tang 提出了基于视觉主要区域(subject region)评价摄影和视频美感的方法[8];Wu 等采用概率后处理(probabilistic post processing)方法,进行了基于支持向量机(SVM)的审美多元标签评价[9]。

以上相关研究均在西方美感研究体系下进行,而在不同文化背景的影响下,审美倾向会出现差异,因此,针对不同文化背景和艺术内容的审美研究有其必要性。为探讨中国传统文化背景下的审美特点,近年,国内学者就中国国画的特点展开了国画图像分类研究。陈俊杰等提取颜色空间和形状特征,对国画的山水、花鸟内容进行了基于支持向量机的二分类研究[10];刘晓巍等采用调色板冗余、Kolmogorov 有序度、香农熵有序度和作品复杂度 4 个参数,对中国国画与西方油画的艺术特点做了量化统计和比较分析[11];王征等基于 HSV(Hue-饱和度、Saturation-饱和度、Value-明度)颜色直方图、纹理特征、边缘尺寸直方图和 Gabor 小波特征,用稀疏分组套索方法对 6 位画家国画作品进行模式分类[12];盛家川和李玉芝提取小波纹理特征,对 5 位画家的国画作品进行了分类[13];高峰等构建了包含 1 718 幅古代、现代、当代国画大家作品的数据集,采用级联分类策略,融合国画的点、线特征,实现对国画艺术家标签的模式分类[14];李玉芝等使用卷积神经网络(CNN)提取国画视觉特征,采用改进嵌入式学习算法对 10 位国画艺术家作品进行分类[15];张佳婧等采用人工评估与回归分析,对 60 幅齐白石的绘画建立审美模型[16]。

已有的国画分类研究分类标签多为绘画的具体内容(如山水和花鸟)、绘画技法(如工笔和水墨)或作者,欠缺对于国画美感直接评价和分类。本文建立了国画数据库与其美感特征分析方法,并基于特征重要性系数进行了美感自动分类与美术元素对应关联的分析,总体分析框架如图 1 所示。先对图像进行了一定的预处理来均衡国画因保存介质、摄影方式等原因造成的信号失真;然后提取诸多图像特征,采用递归式特征消除来筛选,以获得最有效特征集,并对它们进行相关美术元素的映射标注;用筛选特征进行模式分类,对不同分类模型结果进行比较分析,同时对美感分类与美术元素间关联性进行分析。该方法有望对美感的数量建模提供参考,对计算机视觉美感计算方面有所帮助。

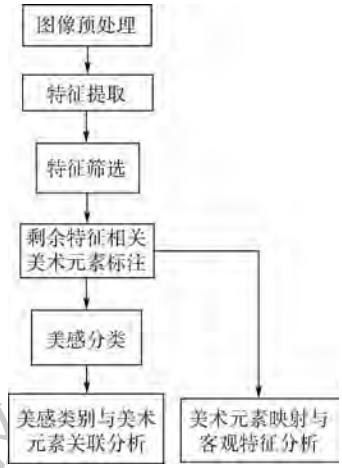


图 1　美感特征分析总体框架

Fig. 1　Framework of aesthetic feature analysis

2　国画美感数据库

数据库的国画数据是通过各种渠道收集的共 511 幅国画图像。其美感类别的定义是从文献[17-27]中收集美感形容词汇后,经过筛选、系列范畴法评价[28]和因子分析[29],获得适合描述国画美感的 5 个类别:气势美、清幽美、生机美、雅致美和萧瑟美,该数据库已经在 Github 上公开,链接地址为 https://github.com/leiyu943/Chinese-painting-aesthetic-database-from-CUC-2019-1,实验过程详见文献[30]。

美感标注实验共有 20 名被试,11 名为女性,9 名为男性。数据库中的每张图画,将先由被试标注美感类别(在上述 5 个美感类别中选择),然后再对该类别下的美感强度进行评分,按 1 ~ 9 的整数等级评定打分。

依照 20 名被试所选人数最多的一类美感作为一幅画作的主要美感。美感分类结果如表 1 所示,5 类美感的实例见图 2。

表 1　图画美感分类数量

Table 1　Number of paintings in each aesthetic classification

美感	山水画	花鸟画	总数目
气势美	110	1	111
清幽美	44	9	53
生机美	54	133	187
雅致美	11	92	103
萧瑟美	30	13	43
无法分类	6	8	14
总数目	255	256	511

图 2 不同美感的绘画内容案例

Fig. 2 Painting content examples of different aesthetics

3 特征提取和筛选

画家的个性化表达和绘画的艺术风格是 2 个影响绘画审美的重要因素。参考以往西方图像分类,与国画研究中画家标签、风格标签的相关分类研究,底层视觉特征(如色彩)、高级视觉特征(如视觉注意)都影响到绘画的美感感知,因此在特征选取上采取底层特征和高层特征相结合的方式。

对于已有特征,出于运算量和便于分析考虑,需要对特征进行筛选,从而用较少维度的特征有效表征图像美感,故对所提图像特征进行了筛选和降维。在特征筛选基础上,进一步进行美术元素和跨文化艺术特征分析。最后进行了中西方美术元素对比分析。

3.1 数据预处理

国画图像的收集渠道不一,在颜色、亮度和尺寸参数上有较大差异性,通过观察和统计库中图像,对这些图像进行以下调整:

1) 黄色校正。颜色偏黄由纸张老化和摄影、扫描时打光不同导致,具体表现为黄色或红色饱和度过高。如图像色相均值落于 0.06 ~ 0.16 之间,则对黄色进行校正,将 H 通道的所有像素点值加 0.02。

2) 亮度调整。亮度偏亮由摄影或扫描的光条件和纸张本身光吸收不同导致。数据库中所有图像亮度均值为 0.705,以此为基准,将过亮、过暗图像向此靠拢。亮度均值大于 0.85 的图片 V 通道所有像素点值减 0.02,小于 0.5 的图片 V 通道所有像素点值加 0.02。

3) 尺寸缩放。将所有图像缩放成 227 × 227 × 3 的尺寸。许多图像尺寸过大,影响特征提取效率,因此做缩小处理;本文所提特征多为统计量,因此不太受尺寸影响,故缩放可行。

预处理算法流程如图 3 所示。

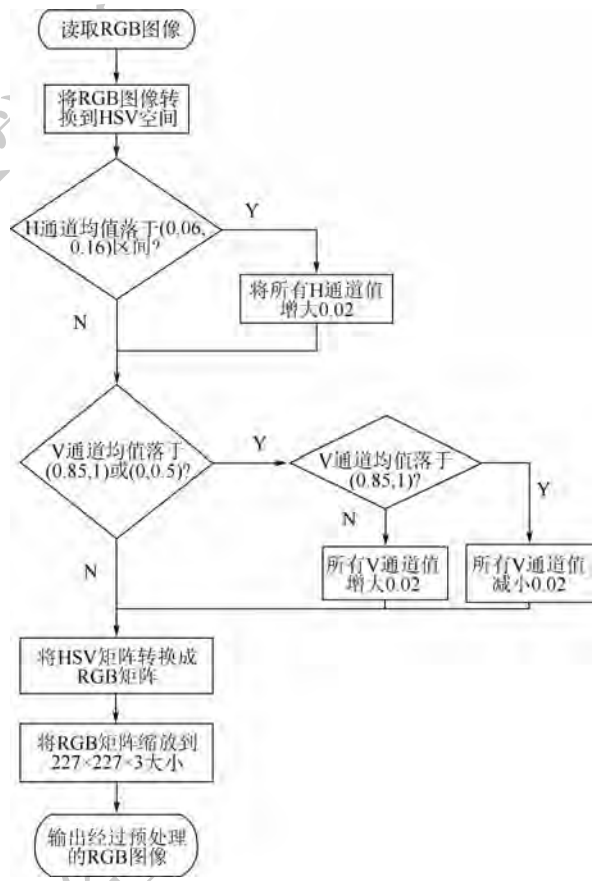


图 3 图像预处理流程

Fig. 3 Image pre-processing flowchart

3.2 适用于国画美感的图像特征

从诸多已有研究中,汇合及筛选有效特征,获得对国画美感分类可能有效的初步特征集。

1) 暗通道直方图

暗通道(dark channel)是一种经典图像去雾算法提到的概念[31]。统计发现,图像中云雾较多部分暗通道呈灰色,并可以此为基准进行图像去雾。暗通道是亮度特征的另一种表达,与传统亮度特征综合使用可以有效提高分类精度。

提取每幅画的暗通道图样,求 10 bin(柄)的直方图,得到 10 维的特征 $D_1 \sim D_{10}$。

2) 边界复杂度及空间分布

边缘是图像不同属性区域的交接处,是区域

属性发生突变的地方,包含着丰富的信息。边缘图像是对原始图像的边缘像素进行判定后,得到的二值图像,是一个二维布尔矩阵。定义边缘像素点占所有像素点的比例为边界复杂度 E。

将 RGB 原图画转化为灰度图像,使用 Canny 算子[32]检测图像边缘。按图 4 区域划分提取边界复杂度 $E_1 \sim E_5$(分区方案出自文献[33])。同时按区域求比值,计算边界的空间分布为

$$\begin{cases} E_6 = (E_1 + E_3)/(E_2 + E_4) \\ E_7 = (E_1 + E_2)/(E_3 + E_4) \\ E_8 = E_1/E_4 \\ E_9 = E_2/E_3 \end{cases} \tag{1}$$

如此构造边界复杂度的对称性特征。加上图像全部区域的边界复杂度 E_{10},构成 10 维边界特征向量。

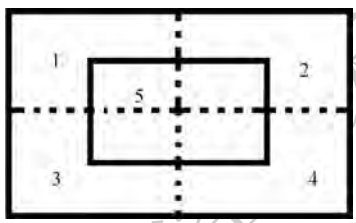


图 4　边界区域划分

Fig. 4　Partition of edges

3) HSV 颜色直方图

颜色直方图是广泛使用的颜色特征[8]。该方法将 RGB 图像转换到 HSV 空间,将 H 通道和 V 通道的值分别均匀量化成 16 个和 8 个,进行相应的直方图统计,获得色相和亮度的直方图统计特征 $H_1 \sim H_{16}$ 与 $V_1 \sim V_8$,共 24 维。

4) 颜色简明度

颜色简明度特征由文献[7]提出。将 H 通道 16 维直方图最大值的 0.01 倍作为阈值,大于此阈值的 bin 的个数作为颜色简明度特征 H_s。

5) RGB 颜色直方图与亮度直方图

将图像 RGB 3 个通道的值分别均匀量化成 10 个,计算直方图 $R_1 \sim R_{10}$、$G_1 \sim G_{10}$ 和 $B_1 \sim B_{10}$,共 30 维;将灰度图像的值均匀量化成 16 个,计算直方图 $\mathrm{Gr}_1 \sim \mathrm{Gr}_{16}$,共 16 维。

6) 直线段个数

对边缘图像使用霍夫变换(Houph transform)进行直线段检测[10],计算直线段的个数作为特征量 L。

7) 邻域差异描述子

邻域差异描述子(neighborhood difference descriptor)计算边界点周边的像素与边界点本身的相似性,以此来描述国画笔法的锋利程度和边界点周围的渐变,类似于图像处理中的锐度。文献[14]用它描述工笔画和写意画在绘画技法上的差别,共 25 维,记为 $N_1 \sim N_{25}$。

8) 显著性图样

显著性图样(saliency map)是由 Koch 和 Ullman 提出的视觉注意力模型[34]。显著性图样表示图像的视觉注意情况。求一幅画的显著性图样,然后求均值,记为 S。该特征表征图中的显著性分布与强度,S 的大小反应显著性强弱。

9) 模糊度

图像模糊度(blur)是 Ke 等提出的度量图像模糊程度的方法[7],它基于高斯模糊假设对图像的模糊程度进行建模,从而计算出图像模糊的程度。图像模糊度值记为 B。

10) 对比度

图像对比度特征的计算方法见文献[7]。分别计算 RGB 3 个通道的 256 bin 直方图 H_r、H_g 和 H_b,将它们相加得到

$$H_\mathrm{a} = H_\mathrm{r} + H_\mathrm{g} + H_\mathrm{b} \tag{2}$$

计算 H_a 包含中间 98% 能量的横轴宽度,如图 5 所示[6]。图中横轴为对比度的分布区间,纵轴为其概率分布 $p(x)$。此宽度值记为对比度 C。

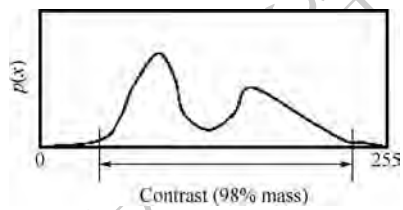


图 5　对比度区间[6]

Fig. 5　Interval of contrast[6]

3.3　特征筛选与分析

3.2 节中提到特征共 120 维,对于美感分类,其中或有冗余。为提高算法运行效率,及更进一步抽象出国画艺术美感所依赖的高级物理特征,需要进行特征筛选。

3.3.1　特征筛选

递归式特征消除(recursive feature elimination)是一种代表性的特征选择方法,它对分类器进行初始特征集的训练[35],并通过相关性等属性来计算特征的重要性。它将依次从每次循环中剔除最不重要的一个或几个特征,然后将这个过程在特征集上递归重复,直到获得能保证较高识别率的最佳特征数。

采用该方法,使用 SVM 作为基础分类器,从 3.2 节提取的 120 个特征中筛选得到 33 个适用于国画美感评价的特征。筛选过程见图 6,其横

轴为选用的特征数,纵轴为分类准确率,可见特征数为 33 时是分类准确率的峰值位置。将筛选出的 33 个特征与美术元素对应,列于表 2 中。

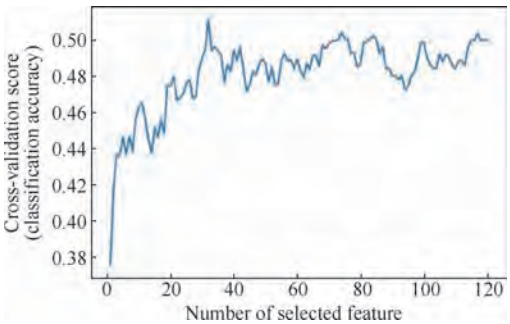


图 6 特征筛选过程

Fig. 6 Process of feature filtering

表 2 有效特征及原理

Table 2 Effective features and theories

特征	具体维度	美术或图像原理
色相直方图	H_4,H_9,H_{14}	颜色
颜色简明度	H_s	颜色
红色直方图	R_7,R_{10}	颜色
绿色直方图	G_6	颜色
蓝色直方图	B_{10}	颜色
灰度直方图	Gr_1,Gr_4,Gr_5,Gr_{12},Gr_{16}	亮度、颜色
对比度	C	亮度和颜色
亮度直方图	V_2,V_3,V_5	亮度
暗通道	D_2	亮度
图像模糊度	B	模糊度、笔触
邻域相似性	N_3,N_4,N_7,N_{10},N_{11},N_{12},N_{16},N_{17},N_{18},N_{24}	笔触
直线段个数	L	线条、笔触
边界复杂度	E_9,E_{10}	线条、构图
显著性图样	S	视觉注意、亮度

3.3.2 特征与美感关联分析

根据表 2,所筛选出的特征相关的美术元素主要为线条、笔触、颜色、亮度和视觉注意等。以下对表 2 中的特征要素与美感的关系进行简要的定性分析:

1) 颜色方面,RGB 直方图有效特征集中在值较高的位置,对应较鲜艳的颜色(R_7、R_{10}、G_6、B_{10});色相直方图显示,橙色(H_4)、黄-绿之间(H_9)、蓝-紫之间(H_{14})颜色的分类效果显著;颜色简明度(H_s)和对比度(C)表明颜色丰富的程度影响国画美感感知。

2) 亮度是重要的美术元素。尽管亮度特征可以直接提取,但特征筛选的结果表明间接特征(灰度、暗通道)与亮度结合可以有效提高分类精度。

3) 线条(L)与笔触(N)的相关特征对视觉感知有直接影响。直线条使得绘画棱角锋利,工笔勾勒与水墨氤氲也导致国画观感有巨大不同。

4) 空间分布(E_9)与视觉注意(S)与构图直接相关,因此构图也会影响国画的美感感知。

表 2 中的分析基于 5 类特征筛选进行。所得结果较为综合、简略,如能对每类美感有关特征进行逐一分析,则能获得更精确详细的结果。为了获得每类美感各自的分类重要特征,采取递归式特征消除方法,对每个美感完成“是否属于该美感”的二分类任务,从而筛选出对每类美感适用的特征,见表 3。

此处的筛选结果与表 2 有较大差异。对比各美感剩余特征发现:①对于区分是否为生机、雅致二类美感,所需的特征相对多且复杂,涉及颜色、亮度、笔触和线条等多个图像原理;②对于区分画作是否处于清幽美,红色直方图特征影响很大;③对于区分萧瑟美,画作的笔触因素微乎其微;④对于区分气势美,线条、笔触因素相对更重要。

表 3 各美感有效特征及原理

Table 3 Effective features and theories for each item of aesthetics

特征	气势美	清幽美	生机美	雅致美	萧瑟美
色相直方图	√		√	√	√
颜色简明度			√	√	
红色直方图		√	√	√	
绿色直方图	√		√	√	
蓝色直方图			√	√	√
灰度直方图			√	√	√
对比度			√		√
亮度直方图	√		√	√	
暗通道			√	√	√
图像模糊度				√	
邻域相似性	√		√	√	√
直线段个数	√		√	√	
边界复杂度			√	√	
显著性图样				√	

4 国画艺术美感分类

4.1 分类器选择与模型检验方法

基于第 2 节国画美感数据库,使用 Extra-Trees、SVM、线性判别分析、随机森林、KNN、朴素贝叶斯、逻辑回归和多元感知机(MLP)8 种具有代表性的分类模型,就本文提出的国画美感特征集进行分类和评析。

采用 k 折交叉检验(k-fold cross validation)[36]作为分类器性能的评价方式。交叉检验将数据平分为 k 等份,依次将每等份作为测试集,剩余部分作为训练集进行参数训练,最后对每一

等份评测结果求平均来评估分类模型的效果。

将查全率(recall)和查准率(precision)[36]作为分类的正确率指标。假设一次分类任务中的真正例数为 TP,假正例数为 FP,真反例数为 TN,对查全率 R 和查准率 P 有

$$\begin{cases} R = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{TN}} \\ P = \dfrac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \end{cases} \tag{3}$$

4.2　分类结果

使用不同分类器就本文提出的国画美感特征进行分类,采用 10 折交叉验证,各算法对于 5 类美感的查全率和查准率结果见表 4。

表 4 分类结果中,各类主流分类器在美感分类任务中的表现相近,其中 Extra-Trees、SVM、线性判别分析、逻辑回归和多元感知机相对来说效果较好。

平均来看,大多数算法的查准率都略高于查全率。样本数量较多的气势、雅致和生机 3 类美感总能保持相对较高的预测精度;一旦出现在数量较少的萧瑟和清幽 2 类美感,查准率几乎都显著高于查全率,且这 2 类美感的预测精度都大幅降低。其中,随机森林、KNN、朴素贝叶斯和多元感知机在数量最少的萧瑟美的识别上,出现了极低的准确率,可见类别数量不均对于分类效果有较大影响。

表 4　不同分类器下的美感自动分类结果

Table 4　Automatic classification results of aesthetics in different classifiers

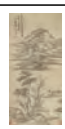
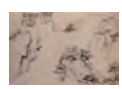
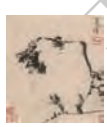
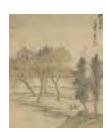
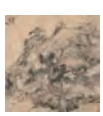
美感	Extra-Trees		SVM		线性判别分析		随机森林		KNN		朴素贝叶斯		逻辑回归		多元感知机	
	P	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*
气势美	0.57	0.63	0.54	0.66	0.57	0.64	0.49	0.61	0.47	0.60	0.34	0.53	0.56	0.65	0.54	0.62
生机美	0.53	0.74	0.58	0.64	0.57	0.67	0.50	0.66	0.55	0.66	0.58	0.18	0.57	0.71	0.57	0.64
雅致美	0.52	0.41	0.51	0.44	0.55	0.48	0.48	0.38	0.51	0.44	0.46	0.44	0.49	0.45	0.44	0.45
萧瑟美	0.20	0.04	0.43	0.23	0.41	0.14	0.20	0.07	0	0	0.18	0.47	0.42	0.14	0.39	0.14
清幽美	0.49	0.21	0.34	0.21	0.34	0.25	0.51	0.21	0.38	0.29	0.16	0.09	0.34	0.16	0.36	0.27
平均	0.46	0.41	0.48	0.44	0.49	0.44	0.44	0.39	0.38	0.40	0.34	0.34	0.48	0.42	0.46	0.42

4.3　分类偏误

以下以 SVM 分类器为例,对各类美感的分类偏误分别进行分析。各美感的误识案例见表 5,表中每个横栏为误识画作原本所属美感,纵栏为画作被误识的美感。观察发现:①在主观评价中结果中,被试间判断较分散的画作更容易被误识;

表 5　美感分类偏误分析

Table 5　Error analysis of aesthetics classification

原美感类别	误识美感类别				
	气势美	生机美	雅致美	萧瑟美	清幽美
气势美					
生机美					
雅致美					
萧瑟美					
清幽美					

②被误识为气势的画作,艺术更有张力;③被误识为清幽美的画作,画面稍显空旷;④被误识为萧瑟美的画作倾向于呈现更有颗粒感的纹理;⑤被误识为雅致美的画作水墨技法占多数。

依照 Extra-Trees 分类器[37]给出特征重要系数(feature importance)。将它们由高至低排列,见表 6。结果显示,颜色简明度、边界复杂度和直线段个数是对美感分类影响最显著的 3 个因素。相对来说,颜色、线条和亮度的有关特征排名更靠前。

按表 2 有效特征分类,将表 6 中同类美术元素特征的重要性系数相加,可得到各类美术元素对美感分类的影响,见图 7。结果显示,对国画美感分类影响最大的美术元素是颜色;其次,笔触、亮度和线条的影响几乎相同;构图对国画美感分类几乎没有影响。

表 6 国画美感特征重要性系数

Table 6 Importance coefficient of aesthetic features in Chinese traditional painting

特征	重要性系数
颜色简明度	0.175 6
边界复杂度 10	0.094 5
直线段个数	0.089 6
亮度直方图 5	0.054 8
红色直方图 10	0.047 3
红色直方图 7	0.040 2
亮度直方图 3	0.035 7
灰度直方图 12	0.026 4
绿色直方图 6	0.025 9
邻域相似性 18	0.024 4
灰度直方图 5	0.023 1
邻域相似性 24	0.021 3
亮度直方图 2	0.020 7
邻域相似性 3	0.020 5
邻域相似性 11	0.020 0
邻域相似性 7	0.020 0
边界复杂度 9	0.019 5
暗通道 2	0.019 4
邻域相似性 12	0.019 1
灰度直方图 1	0.019 1
灰度直方图 16	0.018 6
图像模糊度	0.017 4
显著性区域均值	0.016 9
灰度直方图 4	0.016 4
颜色直方图 4	0.015 9
邻域相似性 4	0.015 4
蓝色直方图 10	0.015 0
对比度	0.014 8
邻域相似性 17	0.014 0
邻域相似性 16	0.012 9
邻域相似性 10	0.010 2
颜色直方图 14	0.009 2
颜色直方图 9	0.006 4

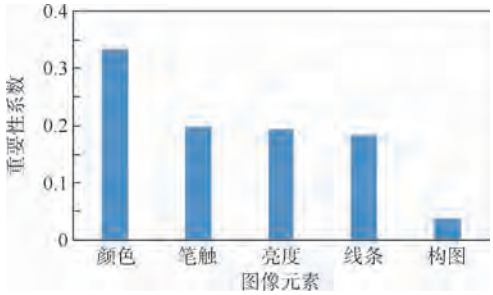


图 7 国画美感分类的美术元素重要性

Fig. 7 Importance of artistic element in aesthetic classification of Chinese traditional painting

5 结 论

为对传统国画的审美进行量化分析和自动分类,本文进行如下工作:

1)建立了一个对国画的艺术美感进行类别与强度标注的图像数据库。

2)基于该数据库,进一步提出一系列适用于国画美感自动分类的图像特征,对特征集与美感之间关联做了定性分析。

3)进行特征筛选后的国画美感自动分类,结果显示,对国画美感进行量化分析与自动分类具有一定可行性。

4)进行了国画的分类误识分析与重要特征分析,通过对特征进行美学标注来建立客观特征与主观美感之间的量化联系。

当然对于国画美感的发掘远不止于此,由于美感本身的模糊性和多义性,基于国画图像的多标签分类也有较高的实用价值,这也是下一步的研究方向。此外,美感数据库还需要进一步地扩充,以解决类别不均衡问题。

参考文献(References)

[1] JOSHI D, DATTA R. Aesthetics and emotions in images[J]. IEEE Signal Processing Magazine, 2011, 28(5): 94-115.

[2] PERRONNIN F, MARCHESOTTI L, MURRAY N. AVA: A large-scale database for aesthetic visual analysis[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2012: 2408-2415.

[3] LUO W, WANG X, TANG X. Content-based photo quality assessment[C]// IEEE International Conference on Computer Vision. Piscataway, NJ: IEEE Press, 2012: 2206-2213.

[4] LI C, CHEN T. Aesthetic visual quality assessment of paintings [J]. IEEE Journal of Selected Topics in Signal Processing, 2009, 3(2): 236-252.

[5] MOHAMMAD S M, TURNEY P D. WikiArt emotions: An anno-

tated dataset of emotions evoked by art[C]// Proceedings of the 11th Edition of the Language Resources and Evaluation Conference,2018.

[6] MENSINK T,VAN GEMERT J C. The Rijksmuseum challenge: Museum-centered visual recognition[C]// Proceedings of International Conference on Multimedia Retrieval. New York:ACM, 2014:451-454.

[7] KE Y,TANG X,JING F,et al. The design of high-level features for photo quality assessment[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway,NJ:IEEE Press,2006:419-426.

[8] LUO Y,TANG X. Photo and video quality evaluation:Focusing on the subject[C]// Proceedings of European Conference on Computer Vision. Berlin:Springer,2008:386-399.

[9] WU Y,BAUCKHAGE C,THURAU C,et al. The good,the bad, and the ugly:Predicting aesthetic image labels[C]// Proceedings of International Conference on Pattern Recognition. Piscataway,NJ:IEEE Press,2010:1586-1589.

[10] 陈俊杰,杜雅娟,李海芳. 中国画的特征提取及分类[J]. 计算机工程与应用,2008,44(15):166-169.
CHEN J J,DU Y J,LI H F. Feature extraction and classification of Chinese painting[J]. Computer Engineering and Applications,2008,44(15):166-169(in Chinese).

[11] 刘晓巍,普园媛,黄亚群,等. 绘画视觉艺术风格的量化统计与分析[J]. 计算机科学与探索,2013,7(10):962-972.
LIU X W,PU Y Y,HUANG Y Q,et al. Quantitative statistics and analysis for painting visual art style[J]. Journal of Frontiers of Computer Science and Technology,2013,7(10):962-972 (in Chinese).

[12] 王征,孙美君,韩亚洪,等. 监督式异构稀疏特征选择的国画分类和预测[J]. 计算机辅助设计与图形学学报,2013,25(12):1848-1855.
WANG Z,SUN M J,HAN Y H,et al. Supervised heterogeneous sparse feature selection for Chinese paintings classification[J]. Journal of Computer-Aided Design & Computer Graphics, 2013,25(12):1848-1855(in Chinese).

[13] 盛家川,李玉芝. 国画的艺术目标分割及深度学习与分类[J]. 中国图象图形学报,2018,23(8): 1193-1206.
SHENG J C,LI Y Z. Learning artistic objects for improved classification of Chinese paintings[J]. Journal of Image and Graphics,2018,23(8): 1193-1206(in Chinese).

[14] 高峰,聂婕,黄磊,等. 基于表现手法的国画分类方法研究[J]. 计算机学报,2017,40(12):2871-2882.
GAO F,NIE J,HUANG L,et al. Traditional Chinese painting classification based on painting technique[J]. Chinese Journal of Computers,2017,40(12):2871-2882(in Chinese).

[15] 李玉芝,盛家川,华斌. 中国画分类的改进嵌入式学习算法[J]. 计算机辅助设计与图形学学报,2018,30(5):893-900.
LI Y Z,SHENG J C,HUA B. Improved embedded learning for classification of Chinese paintings[J]. Journal of Computer-Aided Design & Computer Graphics,2018,30(5):893-900(in Chinese).

[16] 张佳婧,彭韧,王健,等. 水墨画计算审美评估[J]. 软件学报,2016,27(增刊 2):220-233.
ZHANG J J,PENG R,WANG J,et al. Computational aesthetic evaluation of Chinese wash paintings[J]. Journal of Software, 2016,27(Suppl. 2):220-233(in Chinese).

[17] 陈丽君. 美感与积极情绪的关系及对变化觉察的影响[D]. 重庆:西南大学,2010.
CHEN L J. The relationship between aesthetic experience and positive emotion and the impact on change detection[D]. Chongqing:Southwest University,2010(in Chinese).

[18] ISRAELI N. Affective reactions to painting reproductions: A study in the psychology of esthetics[J]. Journal of Applied Psychology,1928,12(1):125-139.

[19] HAGTVEDT H,HAGTVEDT R,PATRICK V M. The perception and evaluation of visual art[J]. Empirical Studies of the Arts,2008,26(2):197-218.

[20] STAMATOPOULOU D. Integrating the philosophy and psychology of aesthetic experience:Development of the aesthetic experience scale[J]. Psychological Reports,2004,95(2):673-695.

[21] SILVIA P J,FAYN K,NUSBAUM E C,et al. Openness to experience and awe in response to nature and music:Personality and profound aesthetic experiences[J]. Psychology of Aesthetics Creativity & the Arts,2015,9(4):376-384.

[22] MARKOVIĆ S. Aesthetic experience and the emotional content of paintings[J]. Psihologija,2010,43(1):47-64.

[23] ROWOLD J. Instrument development for esthetic perception assessment[J]. Journal of Media Psychology Theories Methods & Applications,2008,20(1):35-40.

[24] HAGER M,HAGEMANN D,DANNER D,et al. Assessing aesthetic appreciation of visual artworks—The construction of the art reception survey (ARS)[J]. Psychology of Aesthetics Creativity & the Arts,2012,9(4):320-333.

[25] KARINA V. Die emotionale Wirkung moderner Kunst[D]. Deutschland:Universität Wien,2010.

[26] 丁月华. 概念隐喻理解中的美感体验对科学概念理解的作用研究[D]. 重庆:西南大学,2008.
DING Y H. A research on the role of aesthetic experience of concept metaphor understanding to the scientific concept understanding[D]. Chongqing:Southwest University,2008(in Chinese).

[27] HEVNER K. Experimental studies of the elements of expression in music[J]. American Journal of Psychology,1936,48(2): 246-268.

[28] 孟子厚. 音质主观评价的实验心理学方法[M]. 北京:国防工业出版社,2008:84-89.
MENG Z H. Experimental psychological method of subjective evaluation of sound quality[M]. Beijing:National Defense Industry Press,2008:84-89(in Chinese).

[29] CATTELL R B. The scientific use of factor analysis in behavioral and life sciences[M]. Berlin:Springer,1978.

[30] 湛颖,高妍,谢凌云. 中国国画情感-美感数据库[J/OL]. 中国图像图形学报(2019-06-19)[2019-07-03]. http://www.cjig.cn/jig/ch/reader/view_abstract.aspx? flag=2&file_no=201903210000001&journal_id=jig.
ZHAN Y,GAO Y,XIE L Y. A database for emotion and aesthetic analysis on Chinese traditional paintings[J/OL]. Journal of Image and Graphics(2019-06-19)[2019-07-03]. http://www.cjig.cn/jig/ch/reader/view_abstract.aspx? flag=2&file_

no = 201903210000001&journal_id = jig(in Chinese).

[31] HE K, SUN J, TANG X, et al. Single image haze removal using dark channel prior[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 1956-1963.

[32] CANNY J F. A computational approach to edge detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8(6): 679-698.

[33] DONG Z, TIAN X. Multi-level photo quality assessment with multi-view features[J]. Neurocomputing, 2015, 168(30): 308-319.

[34] KOCH C, ULLMAN S. Shifts in selective visual attention: Towards the underlying neural circuitry[J]. Human Neurobiology, 1987, 4(2): 115-141.

[35] GUYON I, WESTON J, BARNHILL S, et al. Gene selection for cancer classification using support vector machines[J]. Machine Learning, 2002, 46(1): 389-422.

[36] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 26-30.
ZHOU Z H. Machine learning[M]. Beijing: Tsinghua University Press, 2016: 26-30(in Chinese).

[37] GEURTS P, ERNST D, WEHENKEL L, et al. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

作者简介：

湛颖 女，硕士研究生。主要研究方向：图像处理、视听交互。

高妍 女，博士研究生。主要研究方向：心理声学、视听交互。

谢凌云 男，博士，副研究员，硕士生导师。主要研究方向：感知计算、视听交互。

Aesthetic feature analysis and classification of Chinese traditional painting

ZHAN Ying, GAO Yan, XIE Lingyun*

(Key Laboratory of Media Audio & Video of Ministry of Education, Communication University of China, Beijing 100024, China)

Abstract: Automatic classification of aesthetics in images has been a popular research field in these years. Chinese traditional painting is a pivotal embodiment of Chinese traditional arts, so its aesthetics shows a great potential for researching. In this paper, the automatic classification study and relevant feature analysis of aesthetics were conducted in a Chinese painting database annotated with 5 classes of aesthetics. First, based on subjective annotation, by employing feature extraction and selection, 33 optimal image features were filtered out for aesthetic classification. Then, a mapping analysis was conducted on the relationship among objective features, subjective aesthetics and image artistic elements. Finally, an automatic recognition using a variety of mainstream classifiers was implemented on the optimal feature set, and an acceptable performance was obtained, which proves the feasibility and effectiveness of automatic classification of Chinese painting aesthetics. The results show that the main artistic elements (in order) of aesthetic classification for Chinese traditional painting are: color, brushwork, brightness and lines.

Keywords: aesthetic classification; aesthetic feature; Chinese traditional painting; feature selection; image classification

Received: 2019-07-09; **Accepted**: 2019-08-19; **Published online**: 2019-08-27 10:34

URL: kns.cnki.net/kcms/detail/11.2625.V.20190827.0924.002.html

Foundation item: the Fundamental Research Funds for the Central Universities (18CUCTJ086)

* **Corresponding author**. E-mail: xiely@cuc.edu.cn

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0376

结合轮廓及骨架序列编码的二维形状识别

卢勇强，栗志扬*，陈祎楠，刘朝斌，黄一鸣
（大连海事大学 信息科学技术学院，大连 116026）

摘 要：二维形状识别是物体识别中的一个基本问题，被广泛地应用于商标检索、指纹识别、物体定位、图像检索等多个领域。其中，基于生物信息学的二维形状识别是近期一个新的研究方向，基本思想是把二维形状的轮廓转化为生物信息序列，借助标准的生物信息序列分析工具来进行二维形状的匹配和识别。不过，利用轮廓进行信息序列编码存在编码冗余和编码准确性不高的问题，本文提出了一种新型的结合形状轮廓和骨架的序列编码方法。该方法利用骨架表示形状的细长分支，减少编码的冗余；并分别对轮廓和骨架进行不同类型的编码，具备编码简洁、后续匹配准确性高等优点。最后，本文在三个公开数据集上进行大量的形状识别实验，并与多种通用形状识别方法进行了比较。实验表明，本文方法在多个实验中均取得了较高的识别准确率，相比基本的形状特征描述方法，准确率提高了近5%。

关 键 词：形状匹配；轮廓特征；骨架特征；特征描述子；生物信息序列
中图分类号：TP391
文献标识码：A **文章编号**：1001-5965(2019)12-2523-10

形状是人类认知物体的一个重要线索。某些情况下，物体可能并不具备亮度、纹理或颜色等信息，只能通过其形状来表达。仅通过形状信息，人类仍然可以轻松识别不同的物体及其类别。而且，形状对于物体的颜色、光照和纹理的变化是具有稳定性的[1]，也可以作为其他物体表示的一种辅助，以提高物体表示及识别的准确性。由于形状信息具备上述优点，基于形状的物体高效表示及识别一直是研究的热点。

计算机视觉中研究的形状一般分为二维（平面）形状和三维形状两种数据类型，两类形状有多种应用领域。其中，二维形状被广泛地应用于诸如商标检索、指纹表示及识别、物体定位、图像检索等领域中[2]。这些应用的一个基本问题是如何形成一种丰富、简洁和高效的形状特征表示[3]。典型的二维形状表示方法可以大体上分为三类：基于轮廓的形状表示、基于区域的形状表示和基于骨架的形状表示。本文主要关注的是基于轮廓和基于骨架的二维形状表示及识别方法。

形状的轮廓提供了形状大部分的整体及细节信息，是形状表示中最为频繁采用的对象。基于轮廓的形状表示[1-2]多数把轮廓进行离散化采样，通过统计轮廓序列上采样点的空间位置分布关系，来提取形状轮廓上一些稳定的特征量，一般称为形状特征描述符。这些特征量可以是全局的[3]或者是局部的[1]，其中最为典型的特征子应该是形状上下文（Shape Context，SC）[1]。形状上下文在每个采样点上，利用二维直方图统计其余采样点相对于该点的距离和角度变化信息。两个形状之间的相似性通过各自采样点集的最优匹配计算，采样点之间的匹配误差由其形状上下文描述向量衡量。实验表明，基于轮廓的形状表示方

收稿日期：2017-07-09；**录用日期**：2019-08-12；**网络出版时间**：2019-08-20 14:04
网络出版地址：kns.cnki.net/kcms/detail/11.2625.V.20190820.1035.001.html
基金项目：国家自然科学基金（61300187，61672379）；辽宁省自然科学基金（2019-MS-028）
* **通信作者**. E-mail：lizy0205@gmail.com

引用格式：卢勇强，栗志扬，陈祎楠，等. 结合轮廓及骨架序列编码的二维形状识别[J]. 北京航空航天大学学报，2019，45(12)：2523-2532. LU Y Q，LI Z Y，CHEN Y N，et al. Two-dimensional shape recognition based on contour and skeleton sequence coding[J]. Journal of Beijing University of Aeronautics and Astronautics，2019，45(12)：2523-2532 (in Chinese).

法,在形状简单变化时具备较好的稳定性,可以有效应对形状中的噪声和轻度的变形。

不过在形状发生一些等距变换(如关节变换)时,基于轮廓的表示方法往往稳定性不好,而此时形状骨架的拓扑结构会保持不变。可以借助骨架提出形状更简洁且稳定的描述方法。一般来说,骨架是形状内所有最大内切圆圆心的集合,提供了形状的厚度沿骨架的变化信息,对非脊变形和关节具有不变性,且对于轮廓上的噪声比较鲁棒。这些优点使得基于骨架的形状表示独具优势。

然而现实中的形状往往存在较多变形,如放缩、遮挡、大尺度噪声、同一形状的多样性等,给形状的准确表示和识别带来很大挑战。近期,二维形状方面的研究进展包括基于深度神经网络的方法[4]和基于生物信息学的方法[5]等。其中,深度神经网络一般需要大规模的标记样本,而二维形状公开数据集的样本规模往往较小,制约了深度神经网络在二维形状研究领域中的推广。与传统研究方法不同,Bicego 和 Lovato[5]提出了利用生物信息学模型解决二维形状的识别问题,其基本思想是把形状轮廓按照其空间分布编码为生物信息序列,即 DNA 分子序列,然后借助标准的生物信息序列分析工具来进行序列之间的相似性分析,以及在此基础上的形状识别或分类。

而目前两个形状之间的相似性分析,通用的方法还是借助于其特征描述点集之间的最优匹配。匹配过程由传统的动态规划算法实现,存在算法复杂度较高且匹配精确度不够等问题。利用生物信息序列分析工具替代传统基于动态规划的匹配方法,无疑将有助于丰富轮廓之间相似性的计算方法,同时,生物序列相似性匹配具有几十年的研究历史,涌现出了多种针对不同场景的匹配算法。所以,利用生物信息序列分析方法,进行二维形状分析显然颇具新意,是一个有不错前景且值得进一步深入研究的方向。

基于上述思路,引入生物信息序列分析方法到二维形状分析及识别中。研究中发现,轮廓中的细长分支会使得生物信息编码序列中出现大量方向相反且长度相近的编码,这些编码对于形状相似性匹配贡献较少,且有时会制约匹配的精度。区别于 Bicego 和 Lovato[5]的工作,提出了结合形状轮廓和骨架的协同编码方案。

本文的主要贡献:①提出了一种新型的二维形状表示及识别方法,该方法基于生物信息学,把二维形状转化为生物信息序列,可以借助生物信息学领域的序列匹配方法提高二维形状的识别精度。②提出了一种二维形状的轮廓及骨架协同编码的方案。该方案利用骨架表示形状中的细长分支,并分别对轮廓和骨架进行不同类型的编码,具备编码简洁、后续匹配准确性高等优点。同时在基于草图图像检索(SBIR)也有一定的应用前景[6]。③在三个公开形状数据集上做了大量的形状识别实验,并采用不同的准确性评价指标,与多种通用形状识别方法进行了详细的比较,证明了本文方法的优势和有效性。

1 形状描述及匹配

首先简单介绍一些典型的基于轮廓和基于骨架的二维形状表示及识别方法,然后给出在形状匹配中所采用的生物信息序列分析工具的相关概述。

1.1 基于轮廓的形状描述

基于轮廓的形状表示方法是二维形状描述中最为通用的一类方法,在二维形状识别中有着广泛的应用。其描述方法的共同特点是通过统计轮廓序列上点的空间位置分布关系来提取形状轮廓某种不变特征量。

Belongie 等[1]根据轮廓点空间位置关系提出了一种形状上下文的描述方法。该描述方法对轮廓序列上的每个点采用一个向量来表示,通过构造一组特征向量对形状目标进行描述。该方法着重考察轮廓序列上的一个点与该轮廓序列上的其他所有采样点的空间位置分布关系,具有包含信息量大、描述能力强和鲁棒性良好等特点。

Ling 和 Jacobs[2]基于形状上下文的思想,提出了利用轮廓点间的内部距离(Inner-Distance Shape Context,IDSC)来代替 Belongie 等[1]计算形状上下文中采用的欧氏距离,把轮廓序列上两个采样点经形状内部的最短路径长度定义为内部距离。该方法适用于产生柔性变化目标的识别,实验中取得了较好效果。

Biswas 等[7]基于内部距离的形状上下文描述方法,提出了一个新的形状索引和检索框架。通过索引,高效计算待检索形状与数据库中形状的相似度,改进了传统基于动态规划的形状匹配算法的效率。总体上说,基于轮廓形状描述方法的优点主要表现在以下三个方面:

1)可以把二维形状整体信息与局部信息有机地结合在一起,较准确地描述形状结构特征。

2)可以与多种形状匹配算法组合,灵活地采用基于动态规划的形状匹配或基于词典的形状索

引及匹配。

3）可以通过图像分割及物体边界提取,方便地应用于自然图像的形状识别,具有较好的实用性。

这类方法主要存在以下两个方面的不足:

1）仅考虑了目标形状的轮廓边界信息,而轮廓容易受到形状变形的影响,导致提取的特征量有时无法准确反映物体信息。

2）对于复杂场景中的多个物体,基于轮廓的表示方法往往存在轮廓线参数化困难等问题,大多只能处理简单场景下的物体图像。

1.2　基于骨架的形状描述

骨架是形状数据一种重要的几何特征描述符,由 Blum 在 1967 年首次提出[8],并将其应用到形状识别中。使用一种基于最大内切圆模型的骨架表示方法,具体指圆心在形状内部、且至少内切形状轮廓上两个点的圆形。形状的骨架点就是这些内切圆圆心的集合。

陈展展等[9]提出利用树形结构表示形状骨架,首先根据欧氏距离变换定义一个骨架点的中心点,然后构造以骨架中心点为根节点的骨架树,使用骨架中心点到骨架端点路径上经过的骨架点构造特征向量,并利用改进的最优子序列双射算法来进行骨架点之间的匹配,以用于二维形状的分类和识别。

Bai 和 Latecki[10]提出了一种基于骨架路径相似性的形状识别,属于整体对局部的形状描述方法。通过提取当前骨架端点到其他所有骨架端点间骨架路径上的特征量,对骨架端点进行描述,结合最优匹配算法对形状目标进行识别。

基于骨架形状描述方法的优点主要表现在以下三个方面:

1）骨架表示了目标形状的整体结构,也能反映目标形状的边界。

2）骨架简化了目标形状的描述难度,也能方便目标形状的匹配。

3）骨架保持了形状的重要拓扑结构和几何特征,也能降低因噪声而引起的轮廓失真。

这类方法存在以下两个方面的不足:

1）骨架不能有效抵抗大尺度噪声的影响,会产生较多的冗余骨架枝,易出现主次不分、结构混乱的现象,从而影响对物体真实形状和连接关系的判断。

2）轮廓边缘的突起或遮挡会引起形状骨架的变化,有时会较大改变骨架的拓扑连接,造成目标形状骨架的信息多余,从而影响骨架的度量和匹配。

1.3　生物信息序列分析工具

生物信息学中,基因序列在区分生物种类上有决定性的作用。经过过去 40 年生物信息学的研究,在基因序列分析方面已经有了大量的工具和解决方案。在基因序列分析工具中,主要分为双序列比较工具和多序列比较工具。

双序列比较工具分为全局比较和局部比较:全局比较尝试将两个完整的基因序列进行对准,而局部比较则是将两个基因序列的高相似性短区域进行对准。最有名的双序列比对算法是 Needleman-Wunsch（全局比较）[11]和 Smith-Waterman（局部比较）[12]。近些年,比较通用的基因序列比较方法是 BLAST 工具(基本局部对比较工具)[13],在比对的准确性和效率上都有着较好的表现。

多序列比较工具是对三个以上基因序列比较,首先对所有序列对准,然后进行多序列的比较。其最广泛使用的方法是渐进技术,该技术通过从最相似的一对基因序列进行对齐到最不相似的基因序列,近些年主要使用多序列比较工具包括 Clustal[14]、MUSCLE 等。

基于生物信息序列进行形状分析的优点主要表现在以下两个方面:

1）把二维形状编码为生物信息序列,是一种新型的描述二维形状几何分布的方法,拓展了传统的二维形状表示技术。

2）可以利用多种生物信息序列比对工具,进行形状编码的相似性匹配,从而改进了传统基于动态规划的匹配算法,提高了算法效率。

这类方法存在以下两个方面的不足:

1）本质上,生物信息序列表示属于一种形状签名技术,相比于形状上下文,其表示准确性需要进一步提高。

2）目前二维形状的生物信息编码技术,尚存在编码冗余问题,同时也未充分考虑如何使形状编码序列具有更多基因层面的信息。

2　基于轮廓及骨架序列的编码方法

形状编码及匹配流程如图 1 所示。本节具体给出形状轮廓和骨架的协同表示及编码方案。首先,假定二维形状数据 F 的轮廓已经得到,$C(F)=\{c_1,c_2,\cdots,c_n\}$ 为轮廓上的 n 个等距采样点,并按顺时针排列,其中 $c_i=(x_i,y_i)$。其次,利

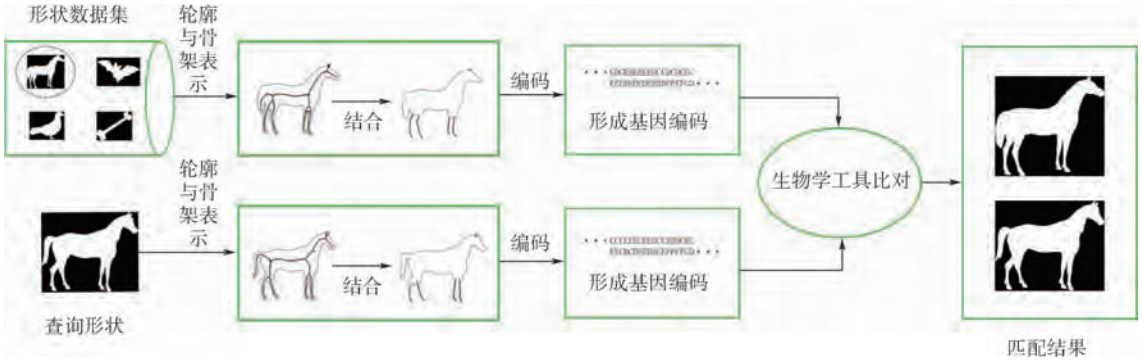


图 1 二维形状匹配流程图

Fig. 1 Flowchart of 2D shape matching

用二维形状骨架提取方法[15]得到形状 F 的骨架 $S(F)$。

2.1 形状归一化

在形状编码过程中,形状的方向对编码有着很大的影响。两个相同的形状,如果编码方向不同,得到的形状编码也会不同,从而影响编码之间的距离。为了解决这个问题,需要对形状进行归一化。首先,旋转形状使其长轴与 x 轴平行,消除不同旋转角度对形状分类结果的影响。

形状的长轴可以由形状数据的主成分分析得到,也可以通过形状中距离最长的两个点简单确定。实验中,发现两种选择的形状识别结果相近。简单起见,本文选择了后者。旋转对齐后,还可能会存在形状与目标匹配形状 180° 翻转的情况。本文采用的解决方案是:以形状长轴作为分隔线,把形状分为两个部分,使得面积较大的那一部分处于长轴的上方。

此时,形状与目标匹配形状可能还存在大小比例不一致的问题,本文编码方法本质上可以有效应对形状之间的比例尺度不一致的问题,不过,为了与其他形状匹配方法进行比较,实验中,还是通过标准的尺度归一公式(式(1))对形状进行尺度归一化。其中,D 为目标长轴大小,D_0 为当前形状长轴大小。

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \begin{bmatrix} \frac{D}{D_0} & 0 & 0 \\ 0 & \frac{D}{D_0} & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{1}$$

2.2 形状的轮廓与骨架联合表示

给定形状 F,利用上述标准化方法得到其轮廓 $C(F)$,同时提取骨架 $S(F)$,如图 2 所示。图中左边是获取了轮廓点与骨架点之后的形状,右边蓝色代表骨架点,绿色代表轮廓点。假定在每个轮廓点或骨架点周围至多有两个轮廓点或骨架点位于其 8 个连通方向上。这里的骨架点不包括骨架的关节点。

Bicego 和 Lovato[5]提出可以把形状轮廓转化为生物信息序列即 DNA 分子序列,本质上是对每个轮廓点的 8 个方向上的轮廓点,利用字母 A ~ H 依次编码。不过发现,这种编码方案在形状细支处两边的编码会产生过多的冗余,使大量重复的编码被考虑进形状分类,如图 3 所示。图 3 中细支上,a 到 b 点的相对位置关系可以近似另一边 c 到 b 点相对位置关系。也即轮廓点从 a 到 b 的编码和从 b 到 c 的编码虽然不一样,但是存在一个变换,在该变换下,两个编码是相似的。所以,a 到 b 的编码和 b 到 c 的编码存在冗余。而发现,如果利用蓝色的骨架表示这个细支,也即通过 e 到 d 的编码代替 a 到 c 的编码可以有效减少

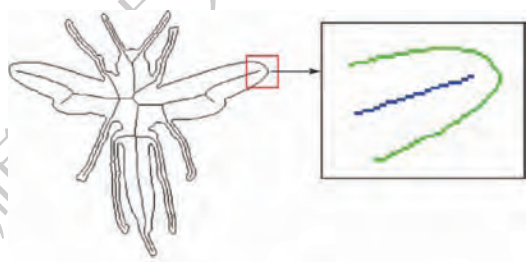


图 2 形状的轮廓与骨架

Fig. 2 Contour and skeleton of a shape

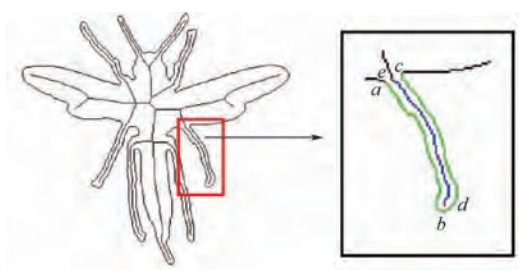


图 3 形状细支处的轮廓与骨架

Fig. 3 Contour and skeleton of branch of a shape

编码的冗余。同时，这种编码方案也可以继承骨架表示的优点。

另外，如果使用整个轮廓点进行编码，如编码图 3 所示的昆虫。昆虫在细支处发生变形，细支处两边轮廓上一般会产生大体相似的变形，从而形状的一处变形会产生编码两处较大的改变。而在非细支处，轮廓的一侧变形，很难使这段轮廓在骨架上对应的另一端轮廓发生变形。所以，仅依靠轮廓信息进行编码无法表示形状的这种特性，从而降低了编码的准确性。显然，如果在细支处采用骨架点编码，在主体轮廓处采用轮廓点编码，可以兼顾编码的简洁性和准确性，有效解决上述问题。

同时，形状中往往会存在形变的情况，如动物形状和日常物体形状等。常见的形变包括轻度变形、遮挡和关节变换等。而骨架处理这些变形问题独具优势[10]，由于本文引入了针对骨架的编码，所以也具备了对于形变物体一定的鲁棒性。

当然，在形状中准确定位其细支位置，本身也是一个比较有挑战性的问题。不过，由于本文编码方案可以单独在轮廓上编码，所以并不要求对于形状中分支的准确定位。实验中，认为骨架特征尺寸（内切圆半径）低于一个给定阈值的骨架点对应的轮廓点处于一个细支中，这些轮廓点不参与编码过程，由其骨架点的编码代替。图 3 的形状经上述方法可以确定形状中所有的细支和主体，从而得到一个联合部分骨架和部分轮廓所形成的形状表示，如图 4 黑色点所示。

图 4　轮廓与骨架的联合表示及默认的编码方向

Fig 4　A combined representation of contour and skeleton and default encoding direction

2.3　形状信息编码方法

在得到形状的轮廓和骨架的联合表示之后，可以对形状进行编码。区别于 Bicego 和 Lovato[5] 单一编码形式，对形状轮廓和骨架分别采用不同的编码形式，即把二维形状转化为字母和数字联合表示的字符编码。

同时，Bicego 和 Lovato[5] 形成编码序列的方法是对每个轮廓点的 8 个方向上的轮廓点依次使用字母 A ~ H 进行编码。如果轮廓存在一些局部变形，这种编码方式的结果是不太稳定的。为了改善这一方案，提出以下编码方法。

在轮廓上，每隔两个轮廓点取一个轮廓点，这样可以在一定程度上解决轮廓发生局部变形时稳定性不足的问题，也可以精简轮廓部分编码的长度，最大限度地保留轮廓上的大部分细节信息。在实验部分，给出了轮廓点的选取方式对准确率影响的比较，在实验结果看来，目前的方案是最优的。由于骨架部分往往仅占全部形状编码的一小部分，所以取所有骨架点进行编码，这样可以使骨架部分的编码长度最长，所占全部编码的比重也最大，使骨架编码的贡献度最高。为了能在字符编码中分别区分轮廓信息给予的贡献和骨架信息给予的贡献，分别用字母和数字来编码轮廓和骨架。

按照上述编码方式，在每个轮廓点周围会存在 24 个方向可能出现下一个轮廓点，对每个可能出现的下一个轮廓点的位置赋予从 A ~ X 的字母表示。起始点设为图像最左上方的轮廓点，即通过下一个轮廓点与此轮廓点的相对位置关系，来表示此轮廓点的编码。每个骨架点周围会有8 个方向出现下一个骨架点，对每个可能出现的下一个骨架点的位置赋予从 1 ~ 8 的数字表示。起始点为离轮廓点最近的骨架点，结尾的骨架点用 0 表示，即通过下一个骨架点与此骨架点的相对位置关系，来表示此骨架点的编码。通过把骨架和轮廓进行编码，会得到图像的编码，如图 5 所示。

关于编码的方向如图 4 红色箭头所示。编码开始的起点为轮廓部分最靠近左上方的点，沿逆时针方向对形状上所有点按照上述规则进行编码，最后回到起点。

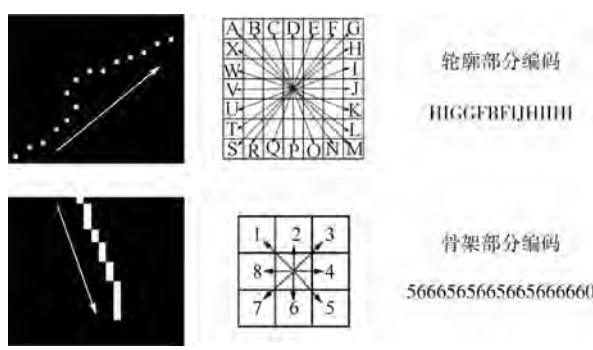


图 5　编码构造规则

Fig. 5　Encoding construction rule

2.4 使用生物学工具形状分类

在通过 2.3 节转换后,可以得到不同形状唯一的字符编码表示。通过生物信息学的映射规则,把字符编码映射为基因序列,在分类时,就可以使用大量生物信息学领域的对齐工具。区别于 Bicego 和 Lovato[5] 使用的 NW(全局比较)和 SW(局部比较)对齐工具,使用更为高效和准确的 MUSCLE 对齐工具。因为 NW 和 SW 对齐工具只能比较 2 个基因序列,效率不高且对齐效果太过依赖生物基因信息的氨基酸占比。在此使用的是支持多序列同时比较、对生物多样性更为适合的 MUSCLE 对齐工具。之后实验会给出使用不同对齐工具对结果的影响。

通过对齐工具的比较,可以得到形状之间的相似度的评分,由于该评分是通过生物学的基因对齐工具获得,为了使该评分能表现形状上的基本信息,引入惩罚项,使评分更为合理。

在惩罚项的选择中,使用简单形状描述符。文献[16]已证实简单形状描述符能达到这一目的。这里选用最小边界矩形面积及离心率作为惩罚项。惩罚项如式(2)。最小边界矩形是通过在原形状上利用周长最小原则做外界矩形,计算此矩形的面积为最小边界矩形面积,离心率是最小外界矩形的宽度与长度之比。

加入惩罚项之后,通过计算式(2),最终得到 2 个形状之间的相似度评分:

$$D'_{a,b} = D_{a,b} - (1 + |\ln E_F - \ln E_{F'}|)(1 + |S_F - S_{F'}|) \tag{2}$$

式中:S_F、$S_{F'}$ 表示 2 个形状最小边界矩形面积;E_F、$E_{F'}$ 表示 2 个形状所形成的最小边界离心率。

3 实验及分析

使用不同的数据集对本文方法进行性能评价。思路来源于文献[5]。首先把二维形状通过轮廓信息转换为生物信息编码,然后使用现有的编码分析工具:全局对齐工具(NW)和局部对齐工具(SW)计算形状之间的相似性。提出了一种结合轮廓和骨架的混合编码方案,并采用 MUSCLE 对齐工具,以进一步提高形状相似性度量的准确性。在形状的识别实验中,论文采用文献[5]中方法,利用 K 近邻分类器进行形状的分类和识别。

3.1 数据集和分类方法

3.1.1 MPEG-7 数据集

MPEG-7 数据集是形状分析研究领域中最流行的数据集之一。有 70 个不同的类别组成,每个类别中有 20 个形状,因此数据集中总计有 1 400 个形状。比较方法使用牛眼比较法(bulls-eye),即对每个查询形状,计算按相似度排序的前 40 个结果中,属于同一类的形状的百分比。同时,本文也采用留一法(leave one out)和留半法(half training)进行识别准确性的评价。留一法,即每次在同一类形状中抽取一个作为测试样本,其余作为训练集,检测样本集被准确分类的百分比。留半法,即每次在一类中选取一半作为测试集,其余全部作为训练集,检测样本集被成功分类的百分比。

3.1.2 ETH-80 数据集

ETH-80 数据集包含 8 类形状,每个类别中有 10 个对象,每个对象具有 41 幅不同角度拍摄的彩色图片。由于本文主要是对形状进行分类,不考虑图片中的颜色问题。在识别准确率评价中,采用基于 K 最近邻的留一法。即对每个形状找与其相似度最高的 K 个形状(不包括本身),记该形状的类别为 K 个形状中类别出现频次最高的那一类。最后统计所有形状识别(分类)成功的个数,从而得到总体的识别准确率。

3.1.3 Swedish leaf 数据集

Swedish leaf 数据集是全部由树叶图片组成的数据集,由 15 个类别组成,每个类别包含 75 个样本。Swedish leaf 数据集由高分辨率的彩色图像组成,虽然提供了丰富的信息,但是与形状分类无关。所以,在原分辨率不变的情况下,本文对图像取二值处理,即只保留图像中叶子的形状信息,这样既可以保留叶子的细节,也可以去除图像中的多余信息。识别准确率的评价采用基于 K 最近邻的留一法进行。

3.2 MPEG-7 数据集的识别结果

MPEG-7 数据集每一类形状的代表如图 6 所示。

MPEG-7 数据集利用牛眼比较法,和现有的方法的对比结果如表 1 所示。可以看出,在牛眼比较的准确率上,本文方法并不是最好的。离性能表现最好的方法尚有一定的差距,但是对于文献[5]得到的 77.24% 的准确率,本文提出的编码

图 6 MPEG-7 数据集

Fig. 6 MPEG-7 dataset

方案还是有较大提升。主要得益于形状中细支处使用骨架和主体上使用轮廓的协同编码方式。

性能表现最好的 2 种方法是 IDSC + LP 和 IDSC + SSP,准确率均在 90% 以上。不过,这 2 种方法性能的提升均建立在对于形状相似度矩阵的距离学习上,属于形状识别领域中的后处理方法。该类方法必须提前得到一个质量较好的形状相似度矩阵,且并不能提供形状的表示。这与本文方法及形状上下文(SC 和 IDSC)等方法有本质上的区别。

进一步,在 MPEG-7 数据集上,本文分别使用了留一法和留半法的评价原则和现有的方法进行了比较,结果如表 2 所示。

由表 2 可见,本文编码方法在使用留一法和留半法时都取得了一个较高的准确率,和文献[5]的方法结果相当,不过本文的编码较为简短,冗余较小。由于一些形状存在比较明显的细支,另一些形状不存在。下面分别给出含有细支处形状和没有含有细支处形状对于形状编码的

表 1　MPEG-7 数据集上牛眼比较法分类准确率对比

Table 1　Comparison of classification accuracy rate of bullseye method on MPEG-7 dataset

方法	准确率/%
IDSC + LP[17]	91.61
IDSC + SSP[18]	93.35
Layered graph[19]	88.75
Aspect shape context[3]	88.3
Shape tree[1]	87.7
MDS + SC + DP[2]	84.35
HPM[20]	86.35
Symbolic representation[21]	85.92
IDSC + DP[2]	85.4
Bioinformatics classification[5]	77.24
本文方法	88.64

表 2　MPEG-7 数据集上的分类准确率对比

Table 2　Comparison of classification accuracy rate on MPEG-7 dataset

方法	准确率/%	
	留半法	留一法
Skeleton paths[22]	86.7	
Class segment set[23]	90.9	97.93
Contour segments[22]	91.1	
ICS[22]	96.5	
Robust symbolic[21]		98.57
Kernel-edit distance[21]		98.93
BCF + SVM[24]	97.16	98.93
Bioinformatics classification[5]	95.85	98.1
本文方法	96.07	98.07

可视化查询,如图 7 所示。

图 7 中,形状上点的颜色表示点与点匹配相似度的高低。通过可视化查询的片段中可以看出,在未含有细支的形状中,如上半部分所示,主体轮廓所形成的编码可以准确地描述形状。在含有细支的形状中,如下半部分所示,本文的方法由于在细支处使用骨架来进行编码,可以更有效地表示出细支处的形状信息,对该类形状的贡献度比较大。所以,在细支处使用骨架、在主体处使用轮廓的编码方案可以有效地表示形状信息。

图 7　MPGE-7 数据集可视化查询

Fig. 7　Visual query of MPEG-7 dataset

3.3　ETH-80 数据集的识别结果

在 ETH-80 数据集,每组中随机选取一个作为测试形状,然后对所有形状进行编码,计算测试形状和训练形状之间的相似度,然后通过相似度排序选取前 79 个形状,统计它们的类别,得到测试形状的分类,并最终计算识别准确率。ETH-80 数据集中 8 类物体的代表性形状如图 8 所示。

本文方法和现有方法的识别准确率如表 3 所示。可以看出,本文方法取得了较高的准确率,相对于传统的描述子如基于内部距离的形状上下文(IDSC + DP),准确率提高了将近 5%。

图 8　ETH-80 数据集

Fig. 8　ETH-80 dataset

3.4　Swedish leaf 数据集的识别结果

由于不同叶子类之间的高度相似性以及叶子类中往往存在较大变形,Swedish leaf 数据集具有较大挑战性。识别准确率的计算方式和 ETH-80 数据集方式相同,在 Swedish leaf 数据集每类中选取一个作为测试形状,然后对所有形状进行编码,通过相似度降序排序选取前 14 个形状的相似度,

判定该形状的类别为 14 个形状中比重最大的类别,最终得到识别准确率。Swedish leaf 数据集 15 类叶子的代表性形状如图 9 所示。

本文方法和现有方法的识别准确率如表 4 所示。可以看出本文在 Swedish leaf 数据集上取得了较高的准确率。

当然,本文方法的识别准确率,与 BCF 和 CNN 方法的准确率相比还有一些差距。BCF 方法利用多个轮廓片段及其池化表示,借助支持向量机(SVM)进行分类及识别。而 CNN 方法把形状表示和识别融合于卷积神经网络的训练和学习,其准确性依靠于样本的规模和有效标记。和这 2 种方法不同,本文的方法基于生物信息序列分析的理论。把形状数据表示为基因序列,可以直接用于形状数据的表示与匹配,也是一个颇具前景的方向。

表 3 ETH-80 数据集上的分类准确率对比

Table 3 Comparison of classification accuracy rate on ETH-80 dataset

方法	准确率/%
Color histogram[25]	64.86
PCA gray[25]	82.99
PCA mask[25]	83.41
SC + DP[4]	86.40
IDSC + DP[4]	88.11
IDSC + Morphological Strategy[26]	88.04
Robust symbolic[21]	90.28
Kernel-edit[27]	91.33
BCF[24]	91.49
Bioinformatics classification[5]	91.33
本文方法	91.37

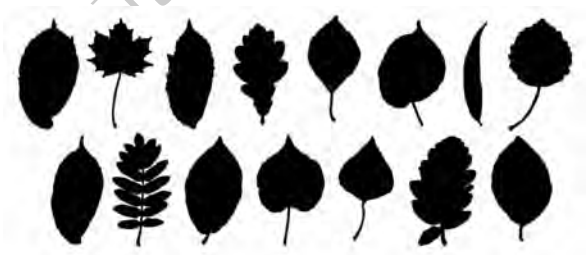


图 9 Swedish leaf 数据集

Fig. 9 Swedish leaf dataset

表 4 Swedish leaf 数据集上的分类准确率对比

Table 4 Comparison of classification accuracy rate on Swedish leaf dataset

方法	准确率/%
Fourier[2]	89.60
SC + DP[2]	88.12
IDSC + DP[2]	94.13
IDSC + Morphological Strategy[26]	94.80
Robust symbolic[21]	95.47
Shape-tree[20]	96.28
BCF[24]	96.56
CNN[28]	99.11
本文方法	94.67

3.5 深入分析

在结合轮廓与骨架的编码方案中,除了本文提出的策略外,还有其他多种选择,不同的编码策略会产生不同的编码形式和编码长度。本节通过实验说明本文选择策略的有效性。

在主体轮廓上中,选取不同间隔的轮廓点:相邻点、间隔一个点、间隔两个点和间隔三个点,产生了不同的生物信息序列;并结合不同的序列对齐工具:局部对齐(SW)、全局对齐(NW)、MUSCLE 对齐以及 MUSCLE 对齐加惩罚项,进行不同组合之间的比较。实验中,选取 MPEG-7 数据集和 ETH-80 数据集,使用上述的留一法进行比较,实验结果如图 10 和图 11 所示。

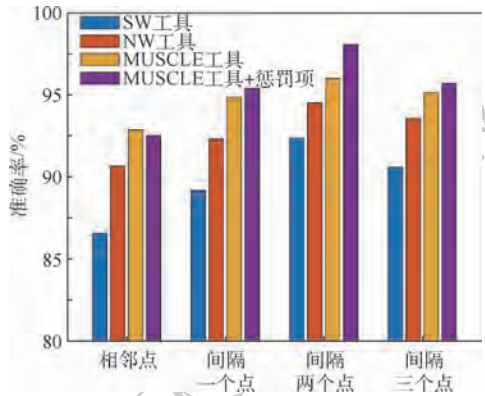


图 10 MPEG-7 数据集上不同编码策略分类准确率的比较

Fig. 10 Comparison of classification accuracy rate on MPEG-7 dataset by different encoding strategies

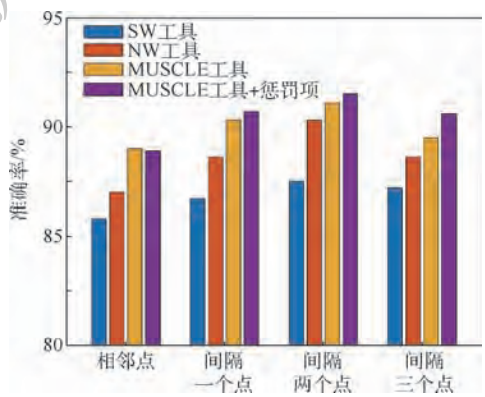


图 11 ETH-80 数据集上不同编码策略分类准确率的比较

Fig. 11 Comparison of classification accuracy rate on ETH-80 dataset by different encoding strategies

可以看出,使用本文的编码选择策略,即在主体轮廓上间隔两个点选取一个轮廓点进行编码,同时采用 MUSCLE 对齐工具和惩罚项,得到了最高的识别准确率。

4　结　论

本文提出了一种新型的二维形状识别方法。该方法首先得到形状主体轮廓和细支骨架的组合表示,其次通过一个轮廓及骨架的协同编码方案,把形状编码为生物信息序列,最后由成熟的生物序列对齐工具进行形状的分类及识别。在 3 个公开形状数据集的实验表明,本文方法相较于现有方法,有一定准确度的提升。

本文方法的优势在于可以有效降低原始编码方案的信息冗余。同时,骨架和轮廓的组合表示及编码可以增强形状编码的鲁棒性,提高后续编码匹配的效率和准确性。最后,把形状识别转化为生物信息序列匹配,可以采用广泛的生物序列对齐工具。

在对形状转化为基因编码的过程中,如何使形状编码序列具有更多基因层面的信息、几何信息与基因信息有更深层次的映射关系[29]仍是一个颇具前景的研究方向。

参考文献(References)

[1] BELONGIE S, MALIK J, PUZICHA J. Shape matching and object recognition using shape contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(4): 509-522.

[2] LING H, JACOBS W D. Shape classification using the inner-distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(2): 286-299.

[3] LING H, YANG X, LATECKI J L. Balancing deformability and discriminability for shape matching[C]// 11th European Conference on Computer Vision. Berlin: Springer-Verlag, 2010, 6313(3): 411-424.

[4] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. ImageNet classification with deep convolutional neural networks[C]// NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems, 2012: 1097-1105.

[5] BICEGO M, LOVATO P. A bioinformatics approach to 2D shape classification[J]. Computer Vision and Image Understanding, 2016, 145: 59-69.

[6] XU D, ALAMEDA-PINEDA X, SONG J, et al. Cross-paced representation learning with partial curricula for sketch-based image retrieval[J]. IEEE Transactions on Image Processing, 2018, 27(9): 4410-4421.

[7] BISWAS S, AGGARWAL G, CHELLAPPA R. Efficient indexing for articulation invariant shape matching and retrieval[C]// 2007 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2007: 1-8.

[8] BLUM H. A transformation for extracting new descriptors of shape[M]// WATHEN DUNN W. Models for the perception of speech and visual form model. Cavnbridge: MIT Press, 1967: 362-380.

[9] 陈展展,汤进,罗斌,等. 基于最优子序列双射的骨架树匹配[J]. 计算机工程与应用,2011,41(7):162-165.
CHEN Z Z, TANG J, LUO B, et al. Skeleton tree matching based on optimal subsequence bijection[J]. Computer Engineering and Applications, 2011, 41(7): 162-165(in Chinese).

[10] BAI X, LATECKI J L. Path similarity skeleton graph matching[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(7): 1282-1292.

[11] NEEDLEMAN S, WUNSCH C. A general method applicable to the search for similarities in the amino acid sequence of two proteins[J]. Journal of Molecular Biology, 1970, 48(31): 443-453.

[12] SMITH T, WATERMAN S M M. Identification of common molecular subsequences[J]. Journal of Molecular Biology, 1981, 147(1): 195-197.

[13] ALTSCHUL S, GISH W, MILLER W, et al. Basic local alignment search tool[J]. Journal of Molecular Biology, 1990, 2145(3): 403-410.

[14] LARKIN M, BLACKSHIELDS G, BROWN N, et al. Clustal w and clustal x version 2.0[J]. Bioinformatics, 2007, 23(21): 2947-2948.

[15] MARIE R, LABBANI-IGBIDA O, MOUADDIB M E. The delta medial axis: A fast and robust algorithm for filtered skeleton extraction[J]. Pattern Recognition, 2016, 56: 26-39.

[16] GOŚCIEWSKA K, FREJLICHOWSKI D. Silhouette-based action recognition using simple shape descriptors[C]// Lecture Notes in Computer Science. Berlin: Springer-Verlag, 2018, 11114: 413-424.

[17] BAI X, YANG X, LATECKI J L, et al. Learning context-sensitive shape similarity by graph transduction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 32(5): 861-874.

[18] WANG J, LI Y, BAI X, et al. Learning context-sensitive similarity by shortest path propagation[J]. Pattern Recognition, 2011, 44(10): 2367-2374.

[19] LIN L, ZENG K, LIU X, et al. Layered graph matching by composite cluster sampling with collaborative and competitive interactions[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2009: 1351-1358.

[20] MCNEILL G, VIJAYAKUMAR S. Hierarchical procrustes matching for shape retrieval[C]// 2006 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2006: 885-894.

[21] DALIRI R M, TORRE V. Robust symbolic representation for shape recognition and retrieval[J]. Pattern Recognition, 2008, 41(5): 1782-1798.

[22] BAI X, LIU W, TU Z. Integrating contour and skeleton for shape classification[C]// 2009 IEEE 12th International Conference

on Computer Vision Workshops. Piscataway, NJ: IEEE Press, 2009:360-367.

[23] SUN B K, SUPER J B. Classification of contour shapes using class segment sets[C]//2005 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2005:727-733.

[24] WANG X, FENG B, BAI X, et al. Bag of contour fragments for robust shape classification[J]. Pattern Recognition, 2014, 47(5):2116-2125.

[25] LEIBE B, SCHIELE B. Analyzing appearance and contour based methods for object categorization[C]//2003 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2003:409-415.

[26] HU X R, JIA W, ZHAO Y, et al. Perceptually motivated morphological strategies for shape retrieval[J]. Pattern Recognition, 2012, 45(9):3222-3230.

[27] DALIRI R M, TORRE V. Shape recognition based on Kernel-edit distance[J]. Computer Vision and Image Understanding, 2010, 114(10):1097-1103.

[28] ATABAY A H. A convolutional neural network with a new architecture applied on leaf classification[J]. IIOAB Journal, 2016, 7(5):326-331.

[29] GAO L, SONG J, NIE F, et al. Optimal graph learning with partial tags and multiple features for image and video annotation[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE Press, 2015:4371-4379.

作者简介:

卢勇强 男,硕士研究生。主要研究方向:计算机视觉、图像处理。

栗志扬 男,博士,副教授,硕士生导师。主要研究方向:计算机视觉、云计算与大数据。

陈祎楠 男,硕士研究生。主要研究方向:计算机视觉。

刘朝斌 男,博士,教授,博士生导师。主要研究方向:云计算与云安全。

黄一鸣 男,硕士研究生。主要研究方向:图像处理、组合优化。

Two-dimensional shape recognition based on contour and skeleton sequence coding

LU Yongqiang, LI Zhiyang*, CHEN Yinan, LIU Zhaobin, HUANG Yiming

(School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China)

Abstract: Two-dimensional shape recognition is a fundamental problem in object recognition, which is widely used in trademark retrieval, fingerprint recognition, object location, image retrieval and other fields. Recently, two-dimensional shape recognition based on bioinformatics has become a new research direction, whose basic idea is to transform the contour of a planar shape into a biological information sequence. The two-dimensional shape matching and recognition are then achieved by the standard alignment tools of such biological information sequence. However, the classic coding method still suffers from the problems of code redundancy and low accuracy. In this paper, we present a new coding method based on both the shape contour and skeleton sequence. Firstly, skeletons are used to represent slender branches of the shape to reduce coding redundancy. Secondly, the contour and skeleton are coded in different ways to compact the code and improve the matching accuracy. Finally, extensive shape recognition experiments are conducted on three public datasets and the proposed method is compared with a variety of shape recognition methods. The experimental results demonstrate that the proposed method has achieved higher performance in several experiments, and the recognition accuracy rate is improved by nearly 5% compared with basic shape feature description methods.

Keywords: shape matching; contour features; skeleton features; feature descriptors; biological information sequences

Received: 2019-07-09; **Accepted**: 2019-08-12; **Published online**: 2019-08-20 14:04

URL: kns.cnki.net/kcms/detail/11.2625.V.20190820.1035.001.html

Foundation items: National Natural Science Foundation of China (61300187, 61672379); Liaoning Provincial Natural Science Foundation of China (2019-MS-028)

* **Corresponding author**. E-mail: lizy0205@gmail.com

http://bhxb.buaa.edu.cn jbuaa@buaa.edu.cn
DOI: 10.13700/j.bh.1001-5965.2019.0388

基于粒子群优化 LSTM 的股票预测模型

宋刚[1,2], 张云峰[1,2,*], 包芳勋[3], 秦超[1,2]
(1. 山东财经大学 计算机科学与技术学院, 济南 250014; 2. 山东财经大学 山东省数字媒体技术重点实验室, 济南 250014;
3. 山东大学 数学学院, 济南 250100)

摘 要: 为了提高股票时间序列预测精度,增强预测模型结构参数可解释性,提出一种基于自适应粒子群优化(PSO)的长短期记忆(LSTM)股票价格预测模型(PSO-LSTM),该模型在 LSTM 模型的基础上进行改进和优化,因此擅长处理具有长期依赖关系的、复杂的非线性问题。通过自适应学习策略的 PSO 算法对 LSTM 模型的关键参数进行寻优,使股票数据特征与网络拓扑结构相匹配,提高股票价格预测精度。实验分别以沪市、深市、港股股票数据构建了 PSO-LSTM 模型,并对该模型的预测结果与其他预测模型进行比较分析。结果表明,基于自适应 PSO 的 LSTM 股票价格预测模型不但提高了预测准确度,而且具有普遍适用性。

关 键 词: 粒子群优化(PSO); LSTM 神经网络; 自适应; 股票价格预测; 预测精度
中图分类号: TP183
文献标识码: A **文章编号**: 1001-5965(2019)12-2533-10

股票市场行情预测一直是投资者迫切关注的话题,然而股票数据具有高噪声、动态、非线性和非参数等特点[1],准确地预测股票价格仍是一项具有挑战性的工作。随着人工智能技术的发展,深度学习以其在机器翻译[2]、语音情感识别[3]、图像识别[4]等方面的优异表现受到广泛关注。与传统统计模型相比,深度神经网络(DNN)可以通过分层特征表示来分析深层和复杂的非线性关系,适合处理股票数据分析这种多因素影响、不稳定、复杂的非线性问题。

循环神经网络(Recurrent Neural Network, RNN)将时序的概念引入到网络结构设计中,使其在时序数据分析中表现出更强的适应性。Hochreiter 和 Schmidhuber 通过对 RNN 网络单元结构进行改进提出了长短期记忆(Long Short-Term Memory, LSTM)模型[5],通过设计控制门结构弥补了 RNN 的梯度消失和梯度爆炸、长期记忆能力不足等问题,使得 RNN 能够真正有效地利用长距离的时序信息[6]。目前,LSTM 神经网络已经成功应用于语音识别[7-8]、文本处理[8]等方面,基于 LSTM 神经网络在这些方面的优异表现,应用其对金融时间序列进行研究受到广泛关注。2015 年,Chen 等[9]使用 LSTM 模型预测了中国股票市场的股票价格;2016 年, Jia[10]验证了 LSTM 模型预测股票价格趋势的有效性;2017 年,Nelson 等[11]基于股票历史数据和技术指标使用 LSTM 模型预测了未来股票市场趋势,并与其他机器学习方法进行比较,该实验表明 LSTM 预测模型具有更高的预测精度;2018 年,Fischer 和 Krauss[12]应用 LSTM 模型对标准普尔 500 指数波动情况进

收稿日期: 2019-07-10; **录用日期**: 2019-08-23; **网络出版时间**: 2019-09-02 09:20
网络出版地址: kns.cnki.net/kcms/detail/11.2625.V.20190830.1655.004.html
基金项目: 国家自然科学基金(61672018,61772309); 国家自然科学基金-浙江两化融合联合基金(U1609218); 山东省重点研发计划(2016GSF120013,2017GGX10109,2018GGX101013); 山东省高等学校优势学科人才队伍培育计划; 山东省自然科学杰出青年(ZR2018JL022); 山东省自然科学基金(ZR2019MF051)
* **通信作者**. E-mail: yfzhang@sdufe.edu.cn

引用格式: *宋刚, 张云峰, 包芳勋, 等. 基于粒子群优化 LSTM 的股票预测模型[J]. 北京航空航天大学学报, 2019, 45(12): 2533-2542. SONG G, ZHANG Y F, BAO F X, et al. Stock prediction model based on particle swarm optimization LSTM[J]. Journal of Beijing University of Aeronautics and Astronautics, 2019, 45(12): 2533-2542 (in Chinese).*

行预测,与随机森林(RAF)、DNN 和逻辑回归分类器(LOG)相比 LSTM 模型具有较好的预测效果。

与其他神经网络模型类似,LSTM 神经网络模型中部分参数需要人为设置,如时间窗口大小、批处理数量、隐藏层单元数目等。这些参数直接控制网络模型拓扑结构,不同参数训练出模型的预测性能差异巨大,因此选择合适的模型参数就显得尤为重要。目前,对于网络模型超参数的选择往往依赖研究者的经验和多次实验结果,耗费大量的人力和计算资源。因此,本文提出一种基于自适应粒子群优化(PSO)的 LSTM 股票价格预测模型(PSO-LSTM),利用自适应学习策略的 PSO 算法对股票数据特征与 LSTM 神经网络拓扑结构进行匹配,获得更高的预测性能。在此基础上,本文选取沪市浦发银行、深市五粮液、港股恒隆集团股票数据,构建 PSO-LSTM 模型对第 2 日股票收盘价进行预测。

1 相关工作

1.1 LSTM 神经网络

LSTM 是一种特殊的循环神经网络。它通过精心设计"门"结构,避免了传统循环神经网络产生的梯度消失与梯度爆炸问题,能有效地学习到长期依赖关系。因此,在处理时间序列的预测和分类问题中,具有记忆功能的 LSTM 模型表现出较强的优势。

如图 1 所示,LSTM 是由多个同构单元格组成,该结构能够通过更新内部状态来长时间存储信息,A 表示 3 个单元具有相同的单元结构。每个单元格由 4 个主要元素构成:输入门、遗忘门、输出门和单元状态。

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_{\mathrm{f}} \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{f}}) \tag{1}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_{\mathrm{i}} \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{i}}) \tag{2}$$

$$\widetilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_{\mathrm{C}} \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{C}}) \tag{3}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \widetilde{\boldsymbol{C}}_t \tag{4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_{\mathrm{o}} \cdot [\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_{\mathrm{o}}) \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \tanh(\boldsymbol{C}_t) \tag{6}$$

式中:$\boldsymbol{x}$ 为 LSTM 单元的输入向量;$\boldsymbol{h}$ 为单元格输出向量;$\boldsymbol{f}$、$\boldsymbol{i}$、$\boldsymbol{o}$ 分别表示遗忘门、输入门和输出门;$\boldsymbol{C}$ 表示单元状态;下标 t 表示时刻;σ、tanh 分别为 sigmoid、tanh 激活函数;$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别表示权重和偏差矩阵。

LSTM 的关键是单元状态 $\boldsymbol{C}$,它在 t 时刻保持单元状态的记忆,通过遗忘门 $\boldsymbol{f}_t$ 和输入门 $\boldsymbol{i}_t$ 进行调节。遗忘门的作用是让细胞记住或忘记它之前的状态 $\boldsymbol{C}_{t-1}$,输入门的作用是允许或阻止传入信号更新单元状态。输出门的作用是控制单元状态 $\boldsymbol{C}$ 输出和传输到下一个单元格。LSTM 单元的内部结构是由多个感知器构成的,反向传播算法通常是最常见的训练方法。

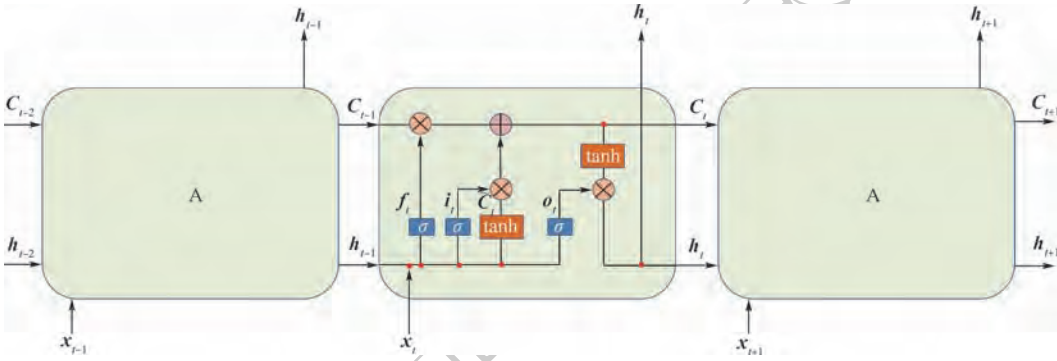


图 1 LSTM 单元结构

Fig. 1 LSTM unit structure

1.2 传统粒子群算法

粒子群算法的思想源于对鸟类社会行为的研究。鸟群捕食最简单有效的方法是搜索距离食物最近的鸟的所在区域,通过个体间的协助和信息共享实现群体进化。算法将群体中的个体看作多维搜索空间中的一个粒子,每个粒子代表问题的一个可能解,其特征信息用位置、速度和适应度值 3 种指标描述,适应度值由适应度函数计算得到,适应度值的大小代表粒子的优劣。粒子以一定的速度"飞行",根据自身及其他粒子的移动经验,即自身和群体最优适应度值,改变移动的方向和距离。不断迭代寻找较优区域,从而完成在全局搜索空间中的寻优过程。

传统粒子群算法描述为:假设在一个 D 维搜

索空间，m 个粒子构成的一个种群 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m\}$，其中 $\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{iD}]$，设 t 时刻 $\boldsymbol{x}_i$ 的特征信息为

位置：$\boldsymbol{X}_i^t=[x_{i1}^t,x_{i2}^t,\cdots,x_{iD}^t]^{\mathrm{T}}$

速度：$\boldsymbol{V}_i^t=[v_{i1}^t,v_{i2}^t,\cdots,v_{iD}^t]^{\mathrm{T}}$

个体最优位置：$\boldsymbol{p}_i^t=[p_{i1}^t,p_{i2}^t,\cdots,p_{iD}^t]^{\mathrm{T}}$

全局最优位置：$\boldsymbol{p}_{\mathrm{g}}^t=[p_{\mathrm{g}1}^t,p_{\mathrm{g}2}^t,\cdots,p_{\mathrm{g}D}^t]^{\mathrm{T}}$

则该粒子在 $t+1$ 时刻的速度和位置信息更新：

$$\begin{cases}v_{id}^{t+1}=wv_{id}^t+c_1r_1^t(p_{id}^t-x_{id}^t)+c_2r_2^t(p_{\mathrm{g}d}^t-x_{id}^t)\\x_{id}^{t+1}=x_{id}^t+v_{id}^{t+1}\end{cases}\tag{7}$$

式中：w 为惯性权重，控制粒子在全局探测和局部开采间的有效平衡；c_1 和 c_2 为学习因子，分别调整飞向自身和全局最好位置方向的步长；r_1 和 r_2 为均匀分布在[0,1]之间的随机数。为避免粒子搜索，一般将其速度和位置分别限制在 $[-V_{\max},V_{\max}]$ 和 $[-X_{\max},X_{\max}]$ 之内。

2　自适应 PSO 的 LSTM 神经网络预测模型

2.1　自适应学习策略的 PSO 算法

本文借助聚类思想，根据粒子自身的分布情况，将整个种群自适应地划分成若干子群[13]。对于每个子群，采用不同的学习策略分布对不同类型的粒子进行更新，提高种群多样性。

采用一种具有较高聚类性能的快速搜索聚类方法实现子群的划分[14]。该方法能自动发现数据集样本的类簇中心，同时能够实现任意形状数据集样本的高效聚类。其基本原理是：类簇中心具备 2 个基本特征，被局部密度较低的点包围，且与局部密度较高的点距离较大。

在一个 D 维搜索空间中，由 h 个粒子构成的一个种群 $\boldsymbol{S}=\{\boldsymbol{x}_i\}_{i=1}^h$，其中 $\boldsymbol{x}_i=[x_{i1},x_{i2},\cdots,x_{iD}]$，对于第 i 个粒子的第 d 维给出 2 个变量，粒子的局部密度 ρ_{id} 与到更高局部密度粒子间距离 δ_{id}，其定义为

$$\rho_{id}=\sum_{j\neq i}\exp\left(-\frac{d_{ij}^2}{d_{\mathrm{c}}}\right)\tag{8}$$

式中：d_{ij} 为 x_{id} 和 x_{jd} 之间欧氏距离；d_{c} 为截断距离。

$$\delta_{id}=\min_{j:\rho_{jd}>\rho_{id}}(d_{ij})\tag{9}$$

对于局部密度 ρ_{id} 最大的样本，其 $\delta_{id}=\max_j(d_{ij})$。

由式(9)可知，若 x_{id} 的密度是最大局部密度，则 δ_{id} 远大于其最邻近粒子的 δ 距离。因此，子群的中心往往是 δ 异常大的粒子，这些粒子的密度 ρ 也相对较高，即选择 ρ 和 δ 都较大的粒子作为聚类中心。对于其他粒子的 x_{jd}，将其归入密度比 x_{jd} 大且距离 x_{jd} 最近的样本所在的子群。

基于子群划分的结果，将每个子群中的粒子分为普通粒子和局部最优粒子两类。对于普通粒子，其主要在子群中最优粒子的引导下拓展局部搜索能力，更新公式为

$$x_{id}=\omega x_{id}+c_1\mathrm{rand}_{1d}(\mathrm{pbest}_{id}-x_{id})+c_2\mathrm{rand}_{2d}(\mathrm{cgbest}_{cd}-x_{id})\tag{10}$$

式中：ω 为惯性权重；c_1、c_2 为学习因子；rand_{1d}、rand_{2d} 为区间[0,1]上的均匀分布随机数；pbest_{id} 为第 i 个粒子第 d 维的最优位置信息；cgbest_{cd} 为第 c 个子群中的最优位置信息。

对于局部最优粒子，其主要通过综合各子群的信息进行更新，以加强子群间的信息交互，更新公式为

$$x_{id}=\omega x_{id}+c_1\mathrm{rand}_{1d}(\mathrm{pbest}_{id}-x_{id})+c_2\mathrm{rand}_{2d}\left(\frac{1}{C}\sum_{c=1}^{C}\mathrm{cgbest}_{cd}-x_{id}\right)\tag{11}$$

局部最优粒子一方面指导普通粒子的学习，另一方面作为子群间信息交互的媒介。在子群中，局部最优粒子引导着整个子群的搜索方向，若采用式(10)的学习策略，一旦局部最优粒子偏离最优解的搜索方向，会导致整个子群陷入局部最优。因此，局部最优粒子需要突破子群的控制，从其他子群中获得信息。考虑到局部最优粒子是子群中最有可能找到最优解的粒子，式(11)利用各个子群中局部最优粒子的平均信息来指导粒子的更新。通过这种子群间信息的共享，促进子群间寻优信息的传递，进一步提高种群的多样性，避免陷入局部最优。

2.2　PSO-LSTM 模型

股票数据作为一种金融时间序列，其受到多方面因素的影响，具有复杂的不稳定性、非线性与周期不确定性。为了准确地预测股票价格，本文以在时间序列分析中表现优异的 LSTM 模型为基础，构建针对股票数据的预测模型。LSTM 模型中某些超参数的取值控制着模型网络结构，为了使模型网络结构与股票数据特征相匹配，本文将自适应 PSO 算法与 LSTM 模型相融合，构建了 PSO-LSTM 预测模型。

PSO 算法相较于其他生物智能演化算法的最大优势在于算法设计简单、收敛速度快，但易陷入局部最优。自适应 PSO 算法能够根据种群自身的分布，通过自适应的子群划分和粒子更新来避

免局部最优，提高参数寻优的准确性。自适应 PSO 算法的自适应特性使得 LSTM 模型能够根据股票数据的特征，快速、准确地确定最优超参数，实现 LSTM 模型网络结构与股票数据特征的有效结合。

模型首先将时间窗口大小、批处理大小、隐藏层单元数目作为自适应 PSO 算法的优化对象，根据超参数取值范围随机初始化各粒子位置信息。通过式(8)、式(9)计算得到粒子的局部密度 ρ_{id} 及其到更高局部密度粒子的距离 δ_{id}，实现自适应种群划分。

其次，根据粒子位置对应的超参数取值建立 LSTM 模型，利用训练数据对模型进行训练。将验证数据代入训练好的模型进行预测，以模型在验证数据集上的平均绝对百分比误差作为粒子适应度值。

适应度函数 f 定义为

$$f = \frac{1}{K}\sum_{i=1}^{K}\frac{|\hat{y}_i - y_i|}{y_i} \tag{12}$$

式中：K 为验证数据集的数量；$\hat{y}_i$ 为第 i 个验证数据的预测值；y_i 为第 i 个验证数据的真实值。

根据各子群中粒子适应度值的取值情况，将粒子划分为普通粒子、子群局部最优粒子与全局最优粒子。通过式(10)、式(11)分别对不同类别粒子位置信息进行更新。判断是否达到终止条件，达到终止条件即得到优化目标的最优值；否则，重新根据粒子位置信息进行种群划分，计算各粒子适应度值，更新各粒子位置信息，直到满足终止条件。

最后，以超参数最优值构建 LSTM 模型，通过股票数据进行训练和预测。PSO-LSTM 模型架构如图 2 所示。

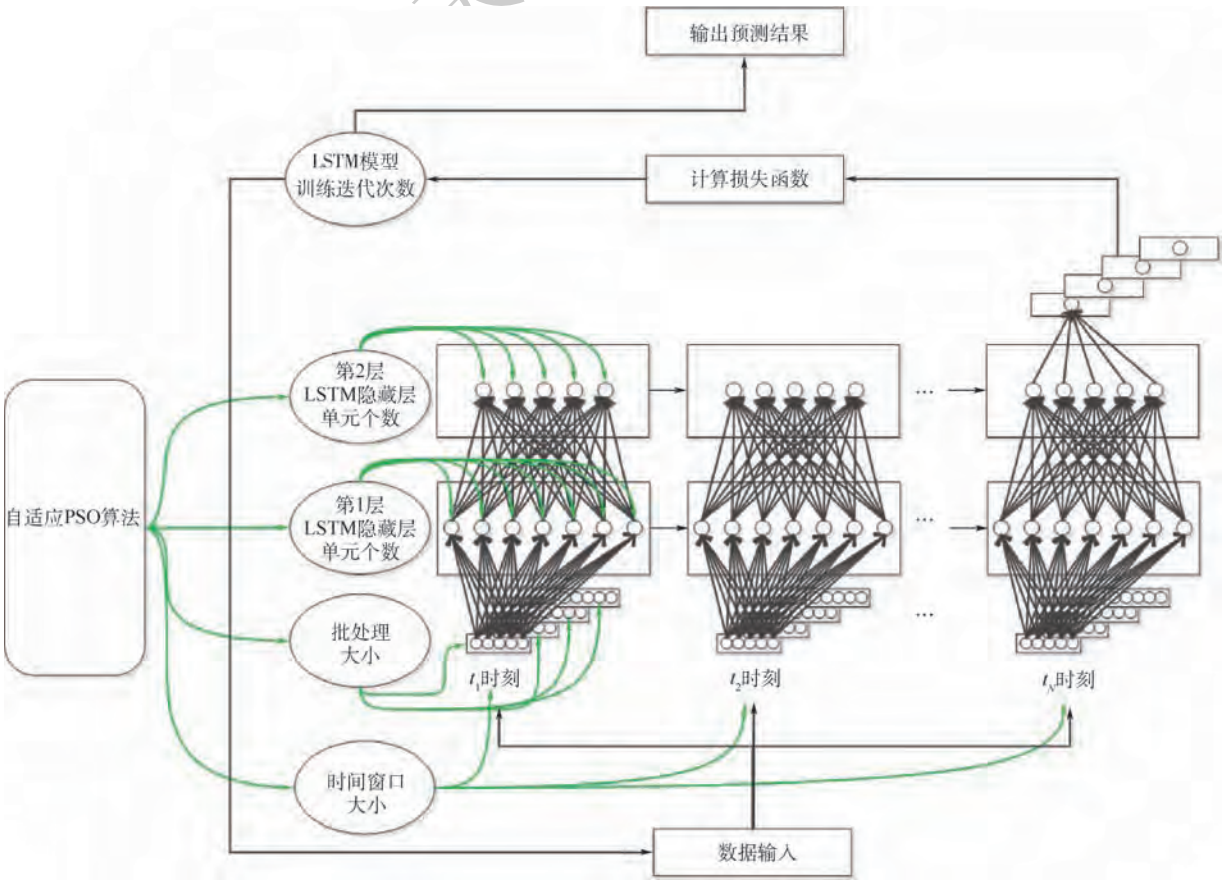


图 2 PSO-LSTM 模型架构

Fig. 2 PSO-LSTM model architecture

2.3 算法流程

PSO-LSTM 模型算法流程如下：

步骤 1 将实验数据分为训练数据、验证数据和测试数据。

步骤 2 将 LSTM 模型中时间窗口大小、批处理大小、神经网络隐藏层单元数目作为优化对象，初始化自适应 PSO 算法。

步骤 3 根据式(8)、式(9)划分子群。

步骤 4　根据式(12)计算每个粒子的适应度值。以各粒子对应参数构建 LSTM 模型,通过训练数据进行训练,验证数据进行预测,将预测结果的平均绝对百分比误差作为各粒子的适应度值。

步骤 5　根据粒子适应度值与种群划分结果,确定全局最优粒子位置 pbest 和局部最优粒子位置 gbest。

步骤 6　根据 PSO 算法的式(10)、式(11)分别对普通粒子和局部最优粒子位置进行更新。

步骤 7　判断终止条件。若满足终止条件,返回最优超参数取值;否则,返回步骤 3。

步骤 8　利用最优超参数构建 LSTM 模型。

步骤 9　模型通过训练数据和验证数据进行训练,测试集进行预测,得到预测结果。

3　实验与结果

实验选取沪市 A 股(600000 浦发银行)、深市 A 股(000858 五粮液)、港股(00010 恒隆集团)为例进行研究,预测股票第 2 日收盘价格。股票历史数据包含开盘价、收盘价、最高价、最低价、涨幅、振幅、成交量、成交额、换手率以及成交次数 10 个属性。其中可能包含停盘停等操作所造成的数据空缺,对于空缺数据进行删除操作,并按时间对数据进行排序。建立 PSO-LSTM 模型并与自回归移动平均模型(ARIMA)[15]、支持向量机(SVM)[16]、多层感知机(MLP)[17]、RNN[18]和 LSTM[19]模型进行对比实验。

3.1　模型评价标准

本文选取均方根误差(RMSE)、平均绝对百分比误差(MAPE)、平均绝对误差(MAE)、均方误差(MSE)和决定系数 R^2 作为评价指标对模型预测效果进行定量评价。其中 RMSE、MAPE、MAE、MSE 数值越小,模型预测结果与真实值偏差越小,结果越准确;决定系数 R^2 越接近 1,代表拟合优度越大,模型预测效果越好。具体公式定义为

$$\mathrm{RMSE}=\sqrt{\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n-y_n)^2} \tag{13}$$

$$\mathrm{MAPE}=\frac{1}{N}\sum_{n=1}^{N}\frac{|\hat{y}_n-y_n|}{y_n} \tag{14}$$

$$\mathrm{MAE}=\frac{1}{N}\sum_{n=1}^{N}|\hat{y}_n-y_n| \tag{15}$$

$$\mathrm{MSE}=\frac{1}{N}\sum_{n=1}^{N}(\hat{y}_n-y_n)^2 \tag{16}$$

$$R^2=\frac{\sum_{n=1}^{N}(\hat{y}_n-\bar{y})^2}{\sum_{n=1}^{N}(y_n-\bar{y})^2} \tag{17}$$

式中:N 为实验预测次数;$\hat{y}_n$ 为模型预测值;y_n 为真实值;$\bar{y}$ 为真实值的平均值。

3.2　模型参数设置

PSO-LSTM 模型结构由输入层、两层 LSTM 层、输出层组成,损失函数使用均方误差,模型训练过程采用 Adam 算法进行优化,网络模型的搭建在 Keras 框架下实现。该模型将时间窗口大小、批处理大小、训练次数和隐藏层单元个数设置为 LSTM 模型超参数。为了减少人为因素对模型的影响,实验根据股票数据具体情况,对超参数的取值范围设置如下:指定时间窗口大小取值范围[1,20],批处理大小取值范围[1,60],隐藏层单元个数取值范围[10,30]。同时设置粒子群粒子个数为 30,最大迭代次数为 500,速度惯性权重 ω 为 0.8,加速系数和为 2。

PSO-LSTM 训练次数由模型误差损失情况直接确定,模型迭代到 800 次时误差损失函数达到收敛状态。因此 PSO-LSTM 模型的训练次数为 800。

3.3　实验分析

3.3.1　沪市股票预测

浦发银行 2016-01-04—2018-12-21 日 K 数据共计 708 条,其中 70% 作为训练数据、20% 作为验证数、10% 作为测试数据。自适应 PSO 模型的最优超参数为:时间窗口大小为 10,批处理大小为 58,第 1 层隐藏层单元个数为 12,第 2 层隐藏层单元个数为 22。实验结果如图 3 所示:绿色实线表示测试数据的股票收盘价真实值,红色实线表示不同预测模型的股票收盘价预测值。预测结果显示 ARIMA 模型预测值曲线在趋势上非常符合股价真实值,然而在每个时间点处的预测值与真实值之间总保持一定误差,具有明显的滞后性。这是由于 ARIMA 模型采用了分步预测方法,每个时间点的预测值均为历史时刻的真实值预测产生。因此 ARIMA 模型预测曲线总是落后于股价真实值曲线说明 ARIMA 模型的预测效果并不好,其预测精度可由模型评价指标进行比较;从其他模型预测结果来看,本文提出的 PSO-LSTM 模型的预测曲线更接近股价真实值曲线,特别是在股价波动剧烈处的预测效果优于其他模型。

为进一步验证该模型的预测性能,表 1 给出各预测模型的评价指标计算结果。PSO-LSTM模

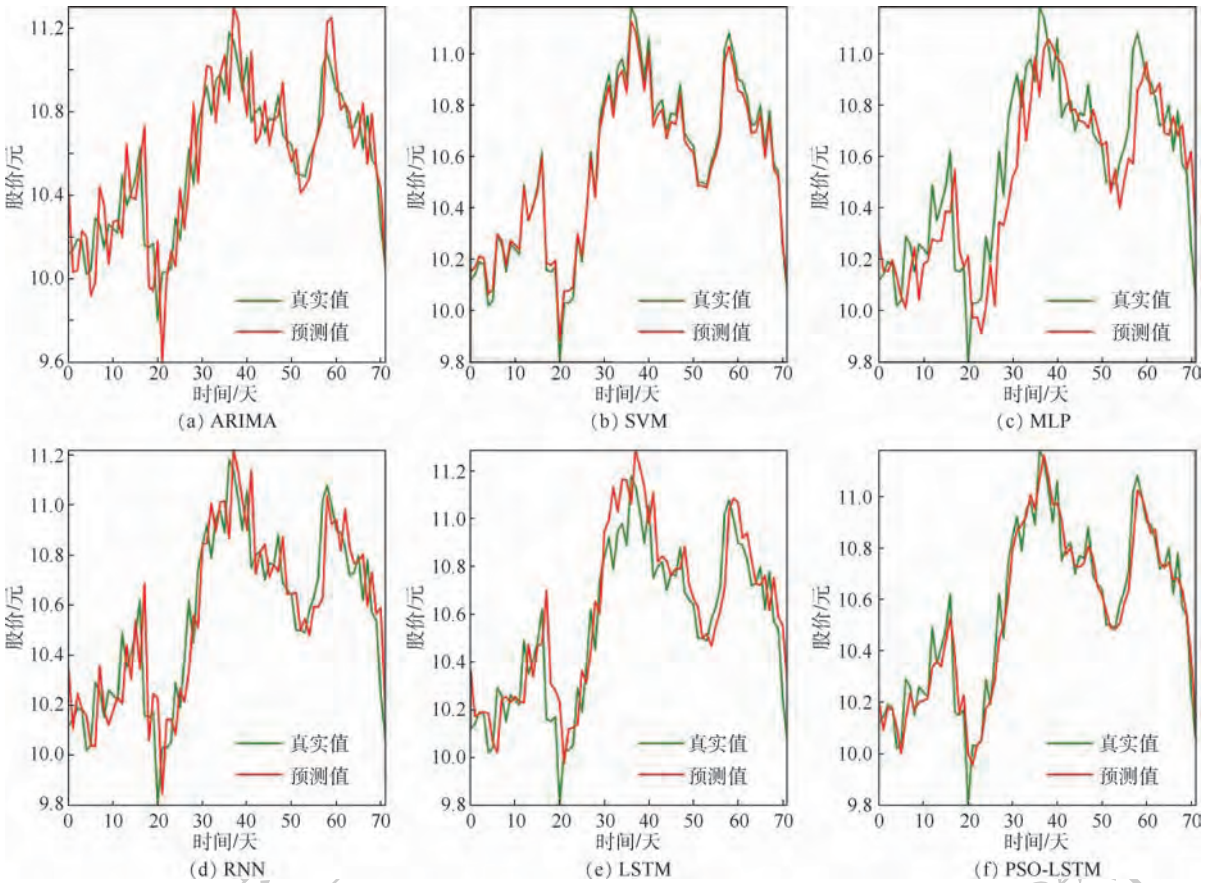


图 3 上海浦发银行各模型预测结果比较

Fig. 3 Comparison of prediction results of various models for Shanghai Pudong Development Bank

表 1 上海浦发银行各模型评价指标比较

Table 1 Comparison of various models' evaluation indicators for Shanghai Pudong Development Bank

模型	RMSE	MAPE/%	MAE	MSE	R^2
ARIMA	0.160 9	3.617 6	0.380 9	0.025 9	0.983 3
SVM	0.184 7	1.414 9	0.148 2	0.034 1	0.966 6
MLP	0.192 4	1.466 5	0.155 1	0.037 4	0.983 2
RNN	0.208 2	1.514 5	0.159 8	0.043 3	1.342 3
LSTM	0.179 9	1.379 2	0.145 1	0.032 3	1.151 8
PSO-LSTM	**0.142 0**	**1.036 9**	**0.106 8**	**0.020 2**	**1.012 0**

型预测误差在 RMSE、MAPE、MAE、MSE 评价标准下都低于其他预测模型;在决定系数 R^2 评价标准中,该模型计算结果比其他预测模型更接近 1,表明 PSO-LSTM 模型的预测性能优于其他模型。特别地,PSO-LSTM 模型 MAPE 指标比 ARIMA 模型低 71%,比 RNN 模型低 31%,模型预测精度显著提高;PSO-LSTM 模型的各项指标优于 LSTM 模型,但两者差距并不明显,主要原因是这两种预测模型具有相同的单元结构;而 PSO-LSTM 模型最突出的优势在于构建过程中不需要人工调参,而且预测结果比普通 LSTM 模型更优。

为了进一步研究该模型与传统 LSTM 模型的有效性与稳定性,实验将浦发银行 2018-12-24—2019-03-12 日的日 K 数据平均分成 5 组,分别利用 LSTM 模型和 PSO-LSTM 模型进行预测,预测结果如图 4 所示。图中:绿色曲线表示股价真实值,蓝色曲线表示 LSTM 模型预测值,红色曲线表示 PSO-LSTM 模型预测值。从图中可以看出,对于第 2 组、第 3 组和第 5 组数据,PSO-LSTM 模型的预测值曲线比 LSTM 模型更逼近股价真实值曲线。对于第 1 组和第 4 组数据,PSO-LSTM 模型与 LSTM 模型的预测结果非常接近。

表 2 给出两种预测模型的 RMSE 指标比较结果。除了在第 1 组数据中 LSTM 模型的 RMSE 指标具有微弱优势,PSO-LSTM 模型在其他 4 组数据预测的 RMSE 指标都优于 LSTM 模型。特别地,在第 5 组数据测试中 PSO-LSTM 模型的预测精度比 LSTM 模型高了 25%。实验结果表明,本文提出的 PSO-LSTM 模型对于不同时段的浦发银行股票价格均有较好的预测能力,在模型预测精度与模型预测稳定性方面比 LSTM 模型更具优势。

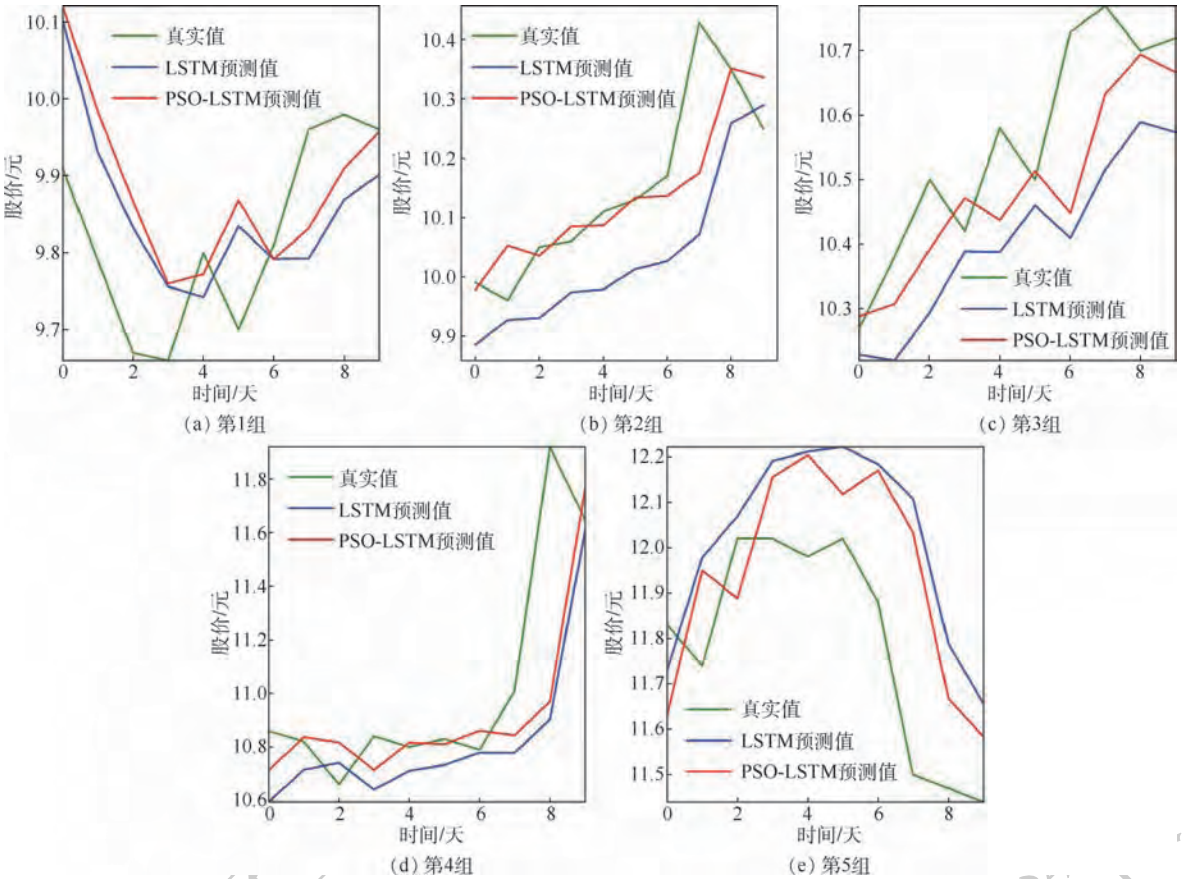


图 4　LSTM 与 PSO-LSTM 预测结果比较

Fig. 4　Comparison of prediction results between LSTM and PSO-LSTM

表 2　LSTM 与 PSO-LSTM 评价指标比较

Table 2　Comparison of evaluation indicators between LSTM and PSO-LSTM

模型	RMSE				
	第 1 组	第 2 组	第 3 组	第 4 组	第 5 组
LSTM	**0.1072**	0.1061	0.1442	0.1629	1.1534
PSO-LSTM	0.1174	**0.0913**	**0.1361**	**0.1240**	**0.1149**

3.3.2　深市股票预测

五粮液 2016-01-04—2018-12-21 的日 K 数据共计 726 条，其中 70% 作为训练数据、20% 作为验证数、10% 作为测试数据。自适应 PSO 模型的最优超参数为：时间窗口大小为 12，批处理大小为 60，第 1 层隐藏层单元个数为 16，第 2 层隐藏层单元个数为 32。实验结果如图 5所示。

如图 5 所示，ARIMA 模型在股票价格持续上涨或下跌时刻的预测效果较好，预测值接近股价真实值。但是在股价变化剧烈的时刻，ARIMA 模型预测结果具有一定的延后性；SVM 模型在股价波动剧烈时刻的预测效果较差，预测曲线与真实值之间具有明显的误差；LSTM 模型的预测效果优于 MLP 模型与 RNN 模型；对于五粮液股票价格的预测，PSO-LSTM 模型的预测效果最好，其预测值曲线能较好地逼近股价真实值。表 3 给出不同预测模型评价指标的比较结果，PSO-LSTM 模型除评价指标均优于其他预测模型。

3.3.3　港股股票预测

恒隆集团 2016-01-04—2018-12-21 的日 K 数据共计 733 条，其中 70% 作为训练数据、20% 作为验证数、10% 作为测试数据。自适应 PSO 模型的最优超参数为：时间窗口大小为 10，批处理大小为 48，第 1 层隐藏层单元个数为 18，第 2 层隐藏层单元个数为 24。实验结果如图 6 所示。ARIMA 模型在股票价格涨跌拐点处的预测结果较差，对于股票上涨或下跌趋势的预测有一定的滞后；SVM 模型对于恒隆集团股票数据的预测表现最差，在某些时间点的预测结果与真实值之间具有明显的差异。相较于其他模型，PSO-LSTM 模型的预测效果仍表现较好。

表 4 给出不同预测模型评价指标的比较结果，PSO-LSTM 模型的各项评价指标均优于其他预测模型，其中 R^2 指标非常接近 1，这说明 PSO-LSTM 模型对于恒隆集团股票数据的拟合最优度最佳。综合沪市、深市、港股股票的预测结果表明，PSO-LSTM 模型对于股票数据的预测具有

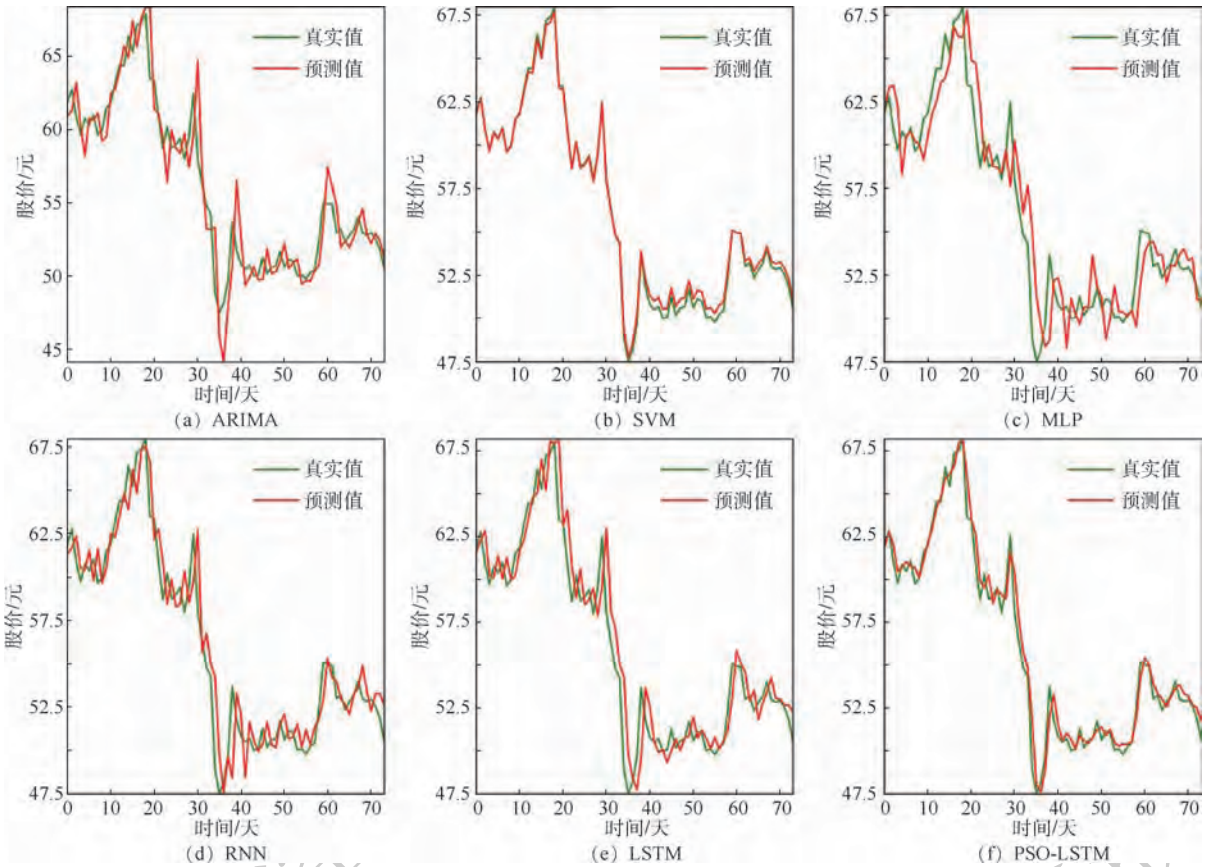


图 5 五粮液各模型预测结果比较

Fig. 5 Comparison of prediction results of various models for Wuliangye

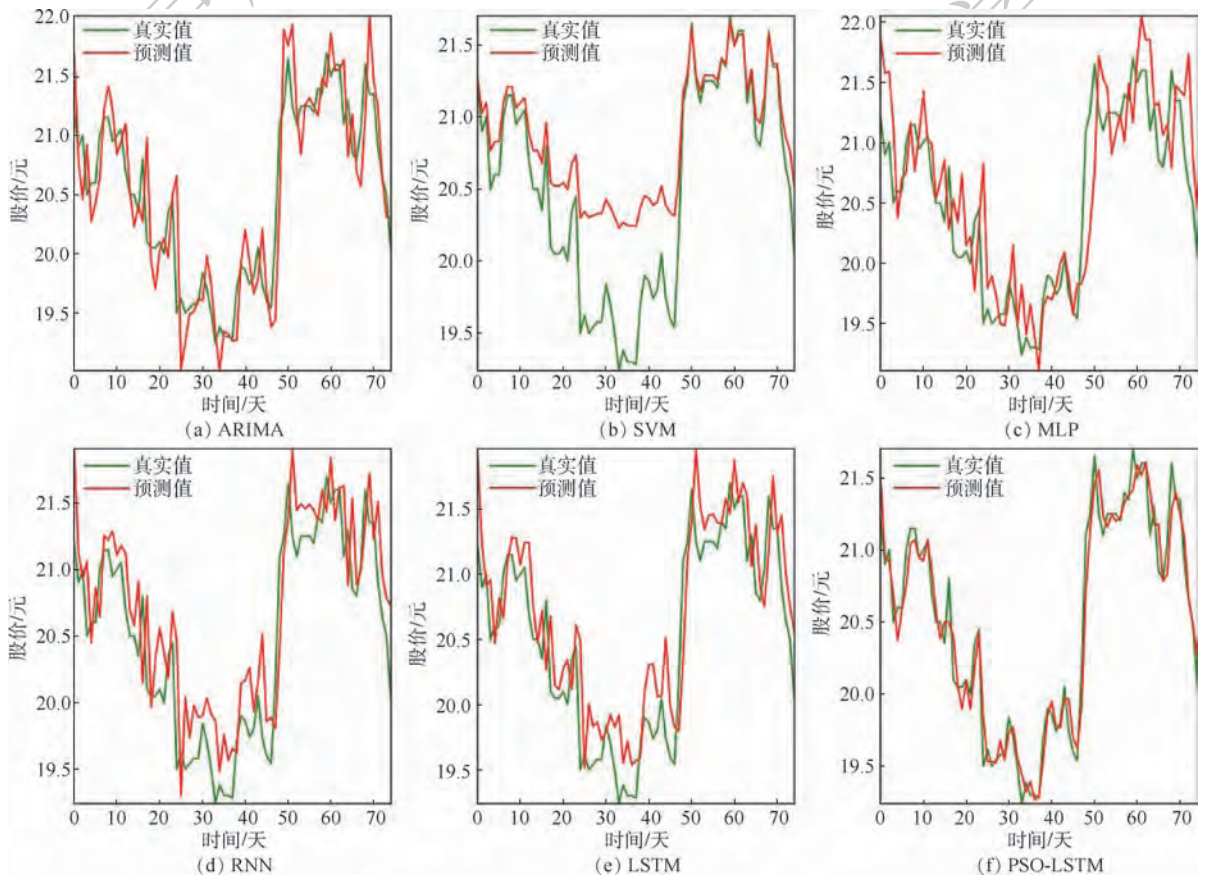


图 6 恒隆集团各模型预测结果比较

Fig. 6 Comparison of prediction results of various models for Hang Lung Group

表 3　五粮液各模型评价指标比较

Table 3　Comparison of various models' evaluation indicators for Wuliangye

模型	RMSE	MAPE/%	MAE	MSE	R^2
ARIMA	1.7814	11.1753	6.2734	3.2851	1.0178
SVM	1.6896	2.4872	1.3764	2.8548	0.7726
MLP	1.8892	2.6822	1.3879	3.4882	1.0605
RNN	1.6549	2.3478	1.3247	2.7203	1.0344
LSTM	1.6013	2.1247	1.1918	2.5734	1.0571
PSO-LSTM	**1.3427**	**1.7808**	**1.0342**	**1.8028**	**1.0124**

表 4　恒隆集团各模型评价指标比较

Table 4　Comparison of various models' evaluation indicators for Hang Lung Group

模型	RMSE	MAPE/%	MAE	MSE	R^2
ARIMA	0.3582	4.3628	0.8921	0.1282	1.2391
SVM	0.5216	2.1204	0.4247	0.2721	0.5797
MLP	0.3439	1.2296	0.2529	0.1183	0.9695
RNN	0.3844	1.5902	0.3242	0.1478	0.9568
LSTM	0.3649	1.4834	0.3026	0.1332	0.9567
PSO-LSTM	**0.2845**	**1.0472**	**0.2366**	**0.0812**	**1.0041**

较高的预测精度与稳定性。

4　结　论

针对复杂的股票价格预测问题，本文提出了 PSO-LSTM 模型，通过自适应学习策略的 PSO 算法对 LSTM 网络结构进行优化，减少人为因素影响，以提高模型捕获股票数据特征的能力。本文随机选取了沪市（浦发银行）、深市（五粮液）、港股（恒隆集团）股票数据进行实验。实验表明相较于统计模型与其他时间序列机器学习模型，PSO-LSTM 模型具有更高的预测精度，并且对于不用类型的股票数据具有一定的普遍适用性。该模型在金融时间序列研究中将具有广阔的应用前景。

参考文献（References）

[1] SI Y W, YIN J. OBST-based segmentation approach to financial time series[J]. Engineering Applications of Artificial Intelligence, 2013, 26(10): 2581-2596.

[2] COSTA-JUSSÀ, MARTA R, ALLAUZEN A, et al. Introduction to the special issue on deep learning approaches for machine translation[J]. Computer Speech & Language, 2017, 46(11): 367-373.

[3] FAYEK H M, LECH M, CAVEDON L. Evaluating deep learning architectures for speech emotion recognition[J]. Neural Network, 2017, 92(2): 60-68.

[4] XING J L, LI K, HU W M, et al. Diagnosing deep learning models for high accuracy age estimation from a single image[J]. Pattern Recognition 2017, 66(1): 106-116.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735-1780.

[6] DEIHIMI A, ORANG O, SHOWKATI H. Short-term electric load and temperature forecasting using wavelet echo state networks with neural reconstruction[J]. Energy, 2013, 57(3): 382-401.

[7] CAI M, LIU J. Maxout neurons for deep convolutional and LSTM neural networks in speech recognition[J]. Speech Communication, 2016, 77(3): 53-64.

[8] ZHOU C, SUN C, LIU Z, et al. A C-LSTM neural network for text classification[J]. Computer Science, 2015, 1(4): 39-44.

[9] CHEN K, ZHOU Y, DAI F. A LSTM-based method for stock returns prediction: A case study of China stock market[C]// IEEE International Conference on Big Data. Piscataway, NJ: IEEE Press, 2015: 2823-2824.

[10] JIA H J. Investigation into the effectiveness of long short term memory networks for stock price prediction[EB/OL]. (2016-03-25)[2019-07-10]. https://arxu-org/abs/1603.07893v1.

[11] NELSON D M Q, PEREIRA A C M, DEOLIVEIRA R A, et al. Stock market's price movement prediction with LSTM neural networks[C]// International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2017: 1419-1426.

[12] FISCHER T, KRAUSS C. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research, 2018, 270(2): 654-669.

[13] RODRIGUEZ A, LAIO A. Clustering by fast search and find of density peaks[J]. Science, 2014, 344(6191): 1492-1496.

[14] 胡旺，李志蜀. 一种更简化而高效的粒子群优化算法[J]. 软件学报，2007，18(4)：861-868.
HU W, LI Z S. A simpler and more effective particle swarm optimization algorithm[J]. Journal of Software, 2007, 18(4): 861-868(in Chinese).

[15] ADEBIYI A A, ADEWUMI A O, AYO C K. Stock price prediction using the ARIMA model[C]// UKSIM-AMSS International Conference on Computer Modelling & Simulation. Piscataway, NJ: IEEE Press, 2015: 106-112.

[16] LIN Y, GUO H, HU J. An SVM-based approach for stock market trend prediction[C]// International Joint Conference on Neural Networks. Piscataway, NJ: IEEE Press, 2013: 1-7.

[17] USMANI M, ADIL S H, RAZA K, et al. Stock market prediction using machine learning techniques[C]// International Conference on Computer & Information Sciences. Piscataway, NJ: IEEE Press, 2016: 322-327.

[18] RATHER A M, AGARWAL A, SASTRY V. Recurrent neural network and a hybrid model for prediction of stock returns[J]. Expert Systems with Applications, 2015, 42(6): 3234-3241.

[19] MCNALLY S, ROCHE J, CATON S. Predicting the price of Bitcoin using Machine Learning[C]// 26th Euromicro International Conference on Parallel Distributed and Network-based Processing. Piscataway, NJ: IEEE Press, 2018: 339-343.

作者简介：

宋刚 男，硕士研究生。主要研究方向：机器学习、数据分析。

张云峰 男，博士，教授，博士生导师。主要研究方向：计算机视觉、数据分析。

Stock prediction model based on particle swarm optimization LSTM

SONG Gang[1,2], ZHANG Yunfeng[1,2,*], BAO Fangxun[3], QIN Chao[1,2]

(1. School of Computer Science and Technology, Shandong University of Finance and Economics, Jinan 250014, China;
2. Shandong Key Laboratory of Digital Media Technology, Shandong University of Finance and Economics, Jinan 250014, China;
3. School of Mathematics, Shandong University, Jinan 250100, China)

Abstract: This paper proposes a stock price prediction model based on particle swarm optimization long short-term memory (PSO-LSTM). This model improves and optimizes the LSTM model, which makes it more appropriate for analyzing relationships such as long-term dependency and for solving complex nonlinear problems. Through finding the key parameters in LSTM model by the PSO algorithm with adaptive learning strategy, the stock data feature matches the network topology structure, and the model's prediction accuracy of stock price is improved. In the experiment, PSO-LSTM models are constructed respectively based on the stock datasets from Shanghai, Shenzhen and Hong Kong, and then they are compared to other prediction models. The comparison results show that the PSO-LSTM stock price prediction model achieves higher prediction accuracy and has general applicability.

Keywords: particle swarm optimization (PSO); LSTM neural network; adaptive; stock price forecast; prediction accuracy

Received: 2019-07-10; **Accepted**: 2019-08-23; **Published online**: 2019-09-02 09:20

URL: kns.cnki.net/kcms/detail/11.2625.V.20190830.1655.004.html

Foundation items: National Natural Science Foundation of China (61672018, 61772309); National Natural Science Foundation of China - Zhejiang Two Integration Joint Fund (U1609218); Key Research and Development Project of Shandong Province (2016GSF120013, 2017GGX10109, 2018GGX101013); Fostering Project of Dominant Discipline and Talent Team of Shandong Province Higher Education Institutions; Natural Science Foundation for Excellent Youth of Shandong Province (ZR2018JL022); Natural Science Foundation of Shandong Province (ZR2019MF051)

* **Corresponding author**. E-mail: yfzhang@sdufe.edu.cn

北 京 航 空 航 天 大 学 学 报

2019 年　第 45 卷　　总目次

（总第 311 期～总第 322 期）

北京航空航天大学学报 2019 年第 45 卷第 1 期（总第 311 期）

北京航空航天大学学报 2019 年第 45 卷第 2 期（总第 312 期）

北京航空航天大学学报 2019 年第 45 卷第 3 期（总第 313 期）

北京航空航天大学学报 2019 年第 45 卷第 4 期（总第 314 期）

北京航空航天大学学报 2019 年第 45 卷第 5 期（总第 315 期）

北京航空航天大学学报 2019 年第 45 卷第 6 期（总第 316 期）

北京航空航天大学学报 2019 年第 45 卷第 7 期 （总第 317 期）

北京航空航天大学学报 2019 年第 45 卷第 8 期（总第 318 期）

北京航空航天大学学报 2019 年第 45 卷第 9 期（总第 319 期）

北京航空航天大学学报 2019 年第 45 卷第 10 期（总第 320 期）

北京航空航天大学学报 2019 年第 45 卷第 11 期（总第 321 期）

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

2019 Vol. 45 Total Contents

(Sum 311 ~ Sum 322)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 1 (Sum 311)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 4 (Sum 314)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 5 (Sum 315)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 6 (Sum 316)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 7 (Sum 317)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 8 (Sum 318)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 9 (Sum 319)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 10 (Sum 320)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 11 (Sum 321)

Journal of Beijing University of Aeronautics and Astronautics 2019 Vol. 45 No. 12 (Sum 322)

《北京航空航天大学学报》征稿简则

《北京航空航天大学学报》是北京航空航天大学主办的以航空航天科学技术为特色的综合性自然科学学术期刊(月刊)。本刊以反映航空航天领域研究成果与动态、促进学术交流、培养科技人才和推动科技成果向社会生产力转化为办刊宗旨。本刊为中国自然科学技术核心期刊,并被 Ei Compendex 等国内外权威文献检索系统收录。本刊向国内外公开发行,为进一步提高办刊质量和所刊出文章的学术水平,特制定本简则。

1 论文作者及内容

1.1 本刊面向海内外所有学者。

1.2 主要刊载与航空航天科学技术有关的材料科学及工程、飞行器设计与制造、宇航科学与工程、信息与电子技术、控制技术和自动化工程、流体力学和动力工程、计算机科学及应用技术、可靠性工程与失效分析等领域的研究文章。航空航天科学技术民用方面以及具有航空航天工程背景的应用数学、应用物理、应用力学和工程管理等方面的文章也在本刊优先考虑之列。

2 来稿要求

2.1 论文应具有创新性、科学性、学术性和可读性。

2.2 论文为原创作品,尚未公开发表过,并且不涉及泄密问题。若发生侵权或泄密问题,一切责任由作者承担。

2.3 主题明确,数据可靠,图表清晰,逻辑严谨,文字精练,标点符号正确。

2.4 文稿撰写顺序:中文题名(一般不超过 20 个汉字),作者中文姓名、单位、所在城市、邮政编码,中文摘要(包括目的、方法、结果及结论),中文关键词(5 ~ 8 个),中图分类号,英文题名,作者英文姓名、单位、所在城市、邮政编码、国别,英文摘要,英文关键词,引言,正文,参考文献。首页下角注明基金项目名称及编号,作者信息。

2.5 作者请登录本刊网页进行在线投稿。

3 稿件的审核、录用与版权

3.1 来稿须经专家两审和主编、编委讨论后决定刊用与否。

3.2 若来稿经审查后认定不宜在本刊发表,将及时告知作者。如果在投稿满 3 个月后仍未收到本刊任何通知,作者有权改投它刊。在此之前,请勿一稿多投,否则一切后果自负。

3.3 来稿一经刊登,即赠送单行本。

3.4 来稿一经作者签字并在本刊刊出,即表明所有作者都已经认可其版权转至本刊编辑部。本刊在与国内外文献数据库或检索系统进行交流及合作时,不再征询作者意见。

邮寄地址:100191 北京市海淀区学院路 37 号 北京航空航天大学学报编辑部

办公地点:北京航空航天大学办公楼 405,407,409 房间

电 话:(010)82315594,82338922,82314839,82315426

E-mail:jbuaa@ buaa. edu. cn

http://bhxb. buaa. edu. cn

http://www. buaa. edu. cn

北京航空航天大学学报

Beijing Hangkong Hangtian Daxue Xuebao

（原《北京航空学院学报》）

（月刊　1956年创刊）

第45卷第12期　2019年12月

主管单位　中华人民共和国工业和信息化部
主办单位　北京航空航天大学
主　　编　赵沁平
编辑出版　《北京航空航天大学学报》编辑部
邮　　编　100083
地　　址　北京市海淀区学院路37号
印　　刷　北京科信印刷有限公司
发　　行　北航文化传媒集团
发行范围　国内外发行
联系电话　(010) 82315594　82338922　82314839
电子信箱　jbuaa@buaa.edu.cn

刊　　号　ISSN 1001-5965 / CN 11-2625/V
国内定价　50.00元／期

JOURNAL OF BEIJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS (JBUAA)

(Monthly, Started in 1956)

Vol.45 No.12　December 2019

Administrated by　Ministry of Industry and Information Technology of the People's Republic of China
Sponsored by　Beijing University of Aeronautics and Astronautics (BUAA)
(Beijing 100083, P. R. China)
Chief Editor　ZHAO Qinping
Edited and Published by　Editorial Board of JBUAA
Printed by　Beijing Kexin Printing Co., Ltd.
Distributed by　BUAA Culture Media Group Limited
Telephone　(010) 82315594　82338922　82314839
E-mail　jbuaa@buaa.edu.cn
http://bhxb.buaa.edu.cn

ISSN 1001-5965

9 771001 596199 12