文信息.

Fig.2 Some samples which include a frame extracted from video, spatial layout, corresponding caption and long-term and short-term memorability ground-truth

图 2 数据样例:从视频中截取的一帧,对应标题、空间布局、视频长时记忆度和短时记忆度

2.2 空间布局表示模块

直观来看,图像或视频中的空间布局有两条线索,即物体之间的排列规则和空间分布.对于物体之间的排列规则,我们可以想象一下:当人们拍照时,他们通常站得井然有序,照片看起来干净整齐,或是阅兵式的仪仗队整齐划一.在空间分布上,摄影师通常遵循一些摄影技巧和规则,如黄金分割法.因此,在记忆度预测中,我们认为并假设物体的空间布局是非常重要的.为了探索视频的空间布局,我们设计了一个简单描述视频空间布局的模板.首先用 Faster-RCNN 检测视频帧中的对象,并得到相应的特征和对应边框.然后用值 1 填充对象的像素,用值 0 填充其余的像素.最后,我们可以得到一个由 0 和 1 组成的掩模,其中物体整体被黑色像素代替,而非物体背景被白色像素代替.图 3 给出了计算给定帧的掩码的过程.

Algorithm 1. 空间布局掩模.

```
Input: X(视频帧像素矩阵), H(视频帧高度),
       C(视频帧宽度)
Output: M(掩模矩阵)
所有物体边框bboxes=ObjectDetection(X)
for i from 1 to H do
    for j from 1 to W do
        M[i][j]=0;
        for bbox in bboxes do
            if X[i][j] in bboxes do
                M[i][j]=1;
                break;
            end
        end
    end
end
```

Fig.3 Spatial layout mask algorithm

图 3 空间布局掩模算法

最后,我们把得到的视频内所有掩模进行平均,并输入到一个简单的 CNN 中,其结构为:卷积层 Conv1 包含 30 个 5×5 的卷积核,步长为 1;Maxpooling 层 maxpool 1,pool size 为 2×2,步长为 2;卷积层 Conv2 包含 15 个 3×3 的卷积核,步长为 1;Maxpooling 层 maxpool 2,pool size 为 2×2,步长为 2;全连接层 hidden_layer 1 维度为 512;dropout 层,dropout rate 为 0.5;全连接层输出维度为 1,表示最终记忆度值.

2.3 局部物体注意力模块

在观看一段视频时,人们的注意力是在不断转换的,也许我们把目光一直锁定在某个物体上,抑或是在不同的物体间来回切换.很多情况下,我们第一眼会看到最大的物体,所以一种简单的方法是使用最大对象来表示帧.但是,它缺少来自其他对象的补充信息.例如,在两帧中检测到的最大对象都是人,一帧位于海滩上,周围是度假者,另一帧则位于办公室,周围是桌子和打印机.环境和周围的物体对语义有很大的影响.

Soft Attention 机制首次应用于机器翻译中.Soft Attention 为每个元素生成概率权重,我们可以利用它为视频帧中的每个对象赋予不同的注意力权值.为了充分利用隐藏在所有对象中的信息,我们提出利用 Soft Attention 将框架中的所有对象嵌入到单个表示中.将第 2.1 节中提到的标题嵌入特征作为一个 query,并得到每个帧中所有对象功能的加权平均值.最后利用 3 层 GRU 网络对时间信息进行捕获,得到一个记忆度得分.

如果标题由 n 个单词组成,可将标题中的第 i 个单词定义为 W_i,而 $f_{ew}(x)$表示单词 x 的词嵌入表示,则全局上下文表示如下:

$$\frac{1}{n}\sum_{i=1}^{n} f_{we}(w_i).$$

g_t 表示注意力权重 α_t 以及在第 t 帧的所有物体特征的加权和.

$$g_t = \sum_{i=1}^{M} a_{i,t} x_i \alpha \in \mathbb{R}^M x_i \in \mathbb{R}^D.$$

注意力机制的权重有网络训练确定,最终通过 softmax 函数表示为一个权重向量:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{M}\exp(e_{t,k})}.$$

未经过 softmax 前的注意力权重是由物体特征和标题的嵌入特征乘积而得来的:

$$e_{i,j} = f_{score(x_i,c)}.$$

$f_{score(x_i,c)}$ 为计算每一个物体的注意力得分的得分函数:

$$f_{score(x_i,c)} = v^{\mathrm{T}} \tanh(Wc + Ux_i + b),$$

其中, $v \in \mathbb{R}^D, W \in \mathbb{R}^{D\times C}, U \in \mathbb{R}^{D\times D}, b \in \mathbb{R}^D$ 分别为网络的权重和偏置.

对于单个视频来说,我们首先抽取视频帧,并用 Faster-RCNN 检测帧中的对象并得到对应的特征和边界框,并按照面积由大到小排列.最后,我们建立了一个 3 层 GRU 网络来捕获整个视频的时序信息以预测出一个记忆度分值.

3 实验与分析

3.1 数据与任务描述

数据集由 8 000 个短的无声视频组成,视频根据许可证共享,许可证允许在 MediaEval 2018 环境中使用和重新发布.我们首先根据视频的记忆度得分进行排名,升序降序皆可(保证记忆度值的分布),并以固定步长值 4 对视频样本进行采样,每 5 个视频采样出一个,作为测试数据.最后,将开发集分为两部分,分别是训练集 6 000 个视频和测试集 2 000 个视频.

这些视频是从专业人士在制作视频时使用的原始视频中提取出来的.每个场景的持续时间为 7s,内容丰富,包含不同的场景类型.每个视频还附带其原始标题.这些标题通常被视为文本元数据,可能有助于预测视频的记忆度.

数据集包含两种标签,即长期记忆标签和短期记忆标签,分别对应于两个子任务.

- 短期记忆度子任务:该任务包括预测给定视频剪辑的“短期”记忆度得分,这反映了观看视频几分钟后记住的可能性;
- 长期记忆度子任务:这项任务包括预测给定视频剪辑的“长期”记忆度得分,这反映了观看后 1~3 天记住的可能性.

对于这两个子任务,官方的评估指标是所有视频的真实记忆度和预测记忆度之间的 Spearman's rank correlation.

3.2 基线系统

一般来说,我们使用回归方法来预测每个视频的记忆度分数,并考虑后期融合来结合不同的特征.采用了两种融合策略,即分数平均和二层回归.

首先,我们使用不同的单一特征进行回归,得到视频的记忆度分数.为了融合多个特征,考虑了两种策略.第 1 种是对同一视频中不同类型特征的分数进行平均,得到的分数是该视频的最终记忆度分数.第 2 种是对于第 2 层回归,我们将来自同一视频的不同特征的分数作为特征连接起来,并输入第 2 层回归模型,从而预测最终的记忆度分数.

因为视频是无声的,因此我们探索了视频和文本的特征,特别是一些高级和语义的表达.视频的标题很短,只有几个字.我们认为人们可能会对某些特定的物体或它们的组合印象深刻.每个词的意义都应该嵌入句子的表示中,以便于记忆预测.

预训练嵌入包含大量语义信息,有助于对句子的语义进行编码.我们尝试用 GloVe[21]词向量作为文本特征.结合每个单词的嵌入,以不同的方式生成句子的表示.首先,简单地把它们加起来,取每个维度的平均值.之后,将平滑的 IDF(inverse document frequency,逆文本频率指数)作为每个单词的权重[22].然后,尝试预先训练的 skip-thought[23]模型.最后,我们还尝试 ConceptNet[10].通过这 4 种方法,可以获得不同类型的视频级表示.

对于视觉特征,我们考虑一些神经网络学的特征和美学特征,包括 C3D[24]、HMP[25]比较、I3D[26]、美学[27].在基线系统,采用 C3D、HMP 和美学特征.此外,我们还提取了 I3D 中 RGB 分支倒数第 2 层的特征.

我们用两种回归器作为基线模型,即支持向量回归(SVR)和随机森林回归(RFR).参数由网格搜索确定.SVR 中的惩罚参数 C 从 0.125~32.内部评估器数量的搜索范围是[100,1000],步长为 100;最大深度范围是[2,10],其中,步长为 2.I3D 模型在 ImageNet 和 Kinetics 进行了预训练.

3.3 模型表现

长时和短时记忆度预测的不同特征结果分别如图 4 和图 5 所示.从图 4 图 5 中可知,文本特征普遍比视觉表示表现得更好.我们认为,标题中包含了关于视频元素的更清晰的描述.如果一个特定的对象是用一个词来描述的,那么嵌入这个词就可以描述这个对象和整个环境中其他对象的关系.视觉特征可能包含一些区域的细节,但不太直观,也许尚未捕捉到与记忆度相关的部分.

对于空间布局,我们考虑 3 种不同的设置,即简单遮罩、重叠遮罩和面积大小遮罩.简单遮罩即为我们用值 1 填充对象的像素,其他像素用值 0 填充.带重叠的遮罩表示重叠区域由包含此区域的对象数填充.而面积大小遮罩是用区域大小而不是值 1 填充像素.对于每种设置,我们还考虑沿时间维度融合所有帧的两种策略:平均和 LSTM.表 1 和表 2 给出不同空间布局策略的结果.可以看到,重叠在空间布局中并不重要,简单遮罩就可以很好地表示空间布局.此外,时序信息对性能没有帮助.我们认为,首先,大多数视频的场景没有剧烈的变化,捕捉到的时序信息并不能很好地和平均策略有所区分.其次,对于场景变化不大的视频来说,人们可能更关注视频的整体空间布局,而不是随时间变化的模式.如果我们将实验数据换成电影广告等视频,也许动作场面的设计、场景的变幻更能抓住人的眼球.

表 3 为我们所提方法中不同策略的结果.可以发现,将所有帧的特征取平均结果即可比基线系统中的最佳结果表现得更好.用 GRU 进行时序信息的捕捉,可以进一步提高效果.在此基础上,我们又加入了物体的注意力

机制,捕捉物体之间的权重分配信息,从而进一步提高模型整体的表现.由此可知,记忆度预测可以从全局语义信息、空间分布、时序信息以及局部的物体信息中受益.

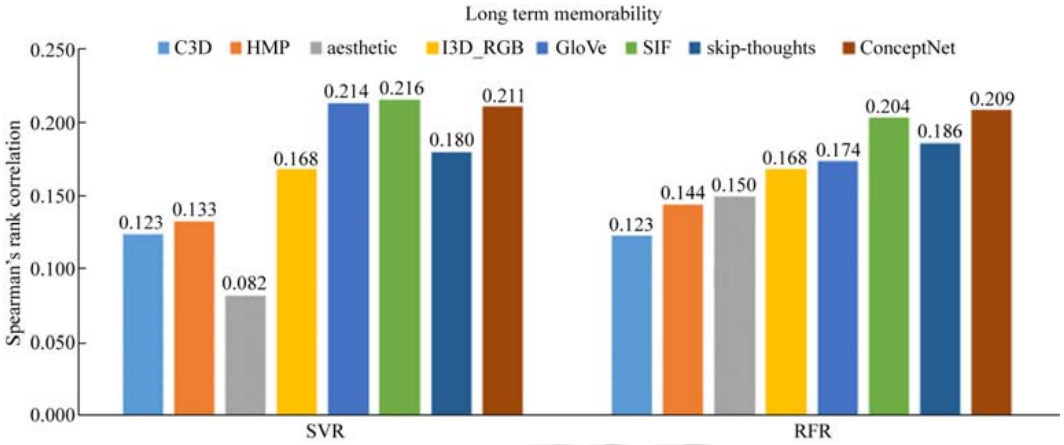

Fig.4 Results of different features for long-term memorability on the local test set

图 4 不同特征在本地测试集上长时记忆度的结果

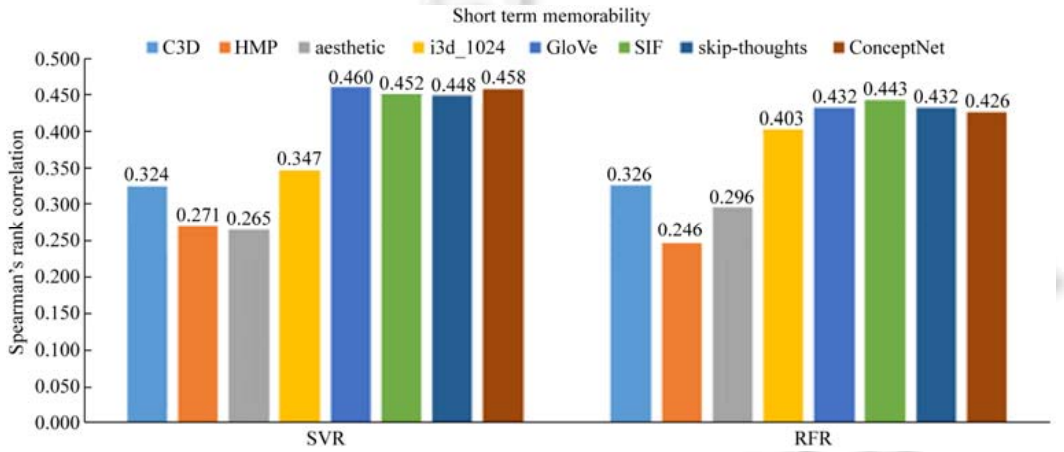

Fig.5 Results of different features for short-term memorability on the local test set

图 5 不同特征在本地测试集上短时记忆度的结果

Table 1 Results of spatial layout for long-term memorability

表 1 空间布局模块的长时记忆度结果

策略	简单遮罩	重叠遮罩	面积大小遮罩
视频帧平均	0.144 0	0.129 0	0.141 1
LSTM	0.138 8	0.121 5	0.138 4

Table 2 Results of spatial layout for short-term memorability

表 2 空间布局模块的短时记忆度结果

策略	简单遮罩	重叠遮罩	面积大小遮罩
视频帧平均	0.280 3	0.270 5	0.277 3
LSTM	0.265 4	0.262 3	0.260 5

Table 3 Results of two strategies in object branch on the test set
表 3 两种处理物体特征策略在测试集上的结果

策略	物体特征平均	物体特征时序	物体注意力
Long-term(无空间布局模块)	0.224 8	0.228 1	0.234 0
Long-term(空间布局)	0.231 2	0.238 5	0.241 6
Short-term(无空间布局模块)	0.447 1	0.458 0	0.465 2
Short-term(空间布局)	0.469 5	0.470 2	0.471 1

从表 1 和表 2 对比以及表 3 的内部对比来看,短时记忆度比长时记忆度的可预测性要高很多.原因在于短时记忆度所表示的是测试者在几分钟内能记住该视频的程度,而长时记忆度的测试时间是 1~3 天后.因此这会造成两个现象,一是从数据的标注来看,短时记忆度普遍高于长期记忆度,因为人短期内能够较清晰地记住几分钟前所看的视频;二是从实验结果来看,短期记忆度更好预测.从图 4 和图 5 来看,记忆度长短时对不同特征的记忆度预测能力影响不大,也就是说,在长时预测较好的特征普遍在短时记忆度上也有良好表现.当然会有一些例外,如 C3D 特征在短时记忆度上的相对预测能力要高于长时记忆度,我们未来工作中也会分析这些原因,比如是否是因为时序上的信息更能影响人的短期记忆而长期记忆更受全局的视频表征影响.

我们挑选了一些视频样例进行分析,发现其中一些视频描绘了物体的具体部位或是特写镜头,而其中一些视频显示了整体的场景,如自然景观、一些人物的故事.

图 6 展示了一些数据样例.图 6(a)和图 6(b)所示长短时记忆度都很高;图 6(c)和图 6(d)所示长时记忆度很高而短时记忆度很低;图 6(e)~图 6(h)所示都有着较低的长时记忆度和较高的短时记忆度.

通过观察很多视频样例及其标签后,我们得出如下一些猜想.

- 短期标签较低的视频通常具有长期标签较低的特征;
- 具有较高短期标签和较低长期标签的视频通常描述一些特写镜头或一些静态的常见对象和场景;
- 有少量的视频具有较低短期标签和较高长期标签.这些视频通常有开放和广阔的场景;
- 一些有趣的物体或场景可以使长期和短期记忆度得分很高,例如一个戴潜水镜的男人坐在海滩上用笔记本电脑工作,样品如图 6(a)所示.因此,视觉和文本内容背后的语义信息是一个值得探讨的重要问题.

从这些样例可以看出,如果一个视频在长期内是值得纪念的,那么在短期内通常也是值得纪念的.相反,具有较高的短期记忆度视频无法决定长期记忆度.所以,可以将对这些样本的分析结论作为一个猜想,推动后面的研究和分析.

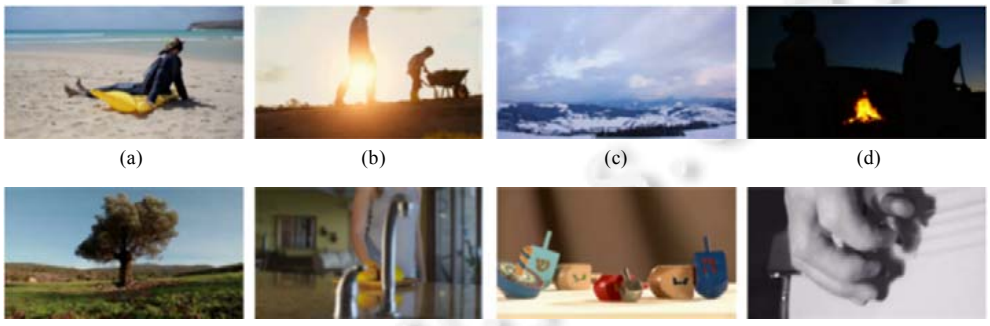

Fig.6 Some samples of the dataset
图 6 数据集中的一些样例

4 总结与展望

互联网、移动设备以及软件服务等不同因素的共同发展,使得互联网上的视频也发生爆炸式的增长.精确预测视频的记忆度可能会给人们的生活带来更大的便利,也能够给企业带来大量商机与发展,例如多媒体检索

与推荐、教育系统、广告设计等等.在本文中,我们从全局和局部两个角度探索了视频的视觉和文本的特征表示,并提出一个视频记忆度预测模型.实验结果表明,在全局的表示上,文本表示的性能优于视觉特征.空间布局特征的表现也十分良好,甚至比一些深度神经网络的特征表示更加有力.同时,局部的物体注意力机制的学习也能很好地捕捉到一些记忆度的信息.对不同对象使用注意力机制,可以有效地将所有对象嵌入到一个单一的表示中,并显示出显著的性能提升.在未来的工作中,我们将重点关注视觉语义的表达和视频是对象与记忆度关系的相关工作,如探索不同对象之间的关系,设计更稳定的模型来预测视频的记忆度.

References:

[1] Romain C, Claire-Hélène D, Duong NQK, Sjöberg M, Ionescu B, Do TT, Rennes F. In: Proc. of the MediaEval 2018: Predicting Media Memorability Task. Sophia Antipolis, 2018. 29–31.

[2] Fajtl J, Argyriou V, Monekosso D, Remagnino P. AMNet: Memorability estimation with attention. In: Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition. 2018. 6363–6372. [doi: 10.1109/CVPR.2018.00666]

[3] Gygli M, Grabner H, Riemenschneider H, Nater F, Van Gool L. The interestingness of images. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2013. 1633–1640. [doi: 10.1109/ICCV.2013.205]

[4] Zhong ZM, Guan Y, Hu Y, Li CH. Mining user interests on microblog based on profile and content. Ruan Jian Xue Bao/Journal of Software, 2017,28(2):278–291 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/5030.htm [doi: 10.13328/j.cnki.jos.005030]

[5] Bhattacharya S, Sukthankar R, Shah M. A frame-work for photo-quality assessment and enhancement based on visual aesthetics. In: Proc. of the ACM Int'l Conf. on Multimedia. 2010. 271–280. [doi: 10.1145/1873951.1873990]

[6] Wang CH, Pu YY, Xu D, Zhu J, Tao ZE. Evaluating aesthetics quality in portrait photos. Ruan Jian Xue Bao/Journal of Software, 2015,26(Suppl.(2)):20−28 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15012.htm

[7] Khosla A, Das Sarma A, Hamid R. What makes an image popular. In: Proc. of the Int'l Conf. on World Wide Web. 2014. 867–876. [doi: 10.1145/2566486.2567996]

[8] Kong QC, Mao WJ. Predicting popularity of forum threads based on dynamic evolution. Ruan Jian Xue Bao/Journal of Software, 2014,25(12):2767–2776 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4730.html. [doi: 10.13328/j.cnki.jos.004730]

[9] Isola P, Xiao JX, Parikh D, Torralba A, Oliva A. What makes a photograph memorable. IEEE Trans. on Pattern Analysis & Machine Intelligence, 2014,36(7):1469–1482. [doi: 10.1109/TPAMI.2013.200]

[10] Speer R, Chin J, Havasi C. ConceptNet 5.5: An openmultilingual graph of general knowledge. In: Proc. of the 31st AAAI Conf. on Artificial Intelligence. San Francisco, 2017. 4444–4451. http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972

[11] Phillips WA. On the distinction between sensory storage and short-termvisual memory. Perception & Psychophysics, 1974,16(2):283–290. [doi: 10.3758/BF03203943]

[12] Wang S, Chen SZ, Zhao JM, Jin Q. Video interestingness prediction based on ranking model. In: Proc. of the Joint Workshop of the 4th Workshop on Affective Social Multimedia Computing and first Multi-ModalAffective Computing of Large-Scale Multimedia Data. ACM, 2018. 55–61. [doi: 10.1145/3267935.3267952]

[13] Squalli-Houssaini H, Duong NQK, Gwenaëlle M, Demarty CH. Deep learning for predicting image memorability. In: Proc. of the 2018 IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018. 2371–2375. [doi: 10.1109/ICASSP.2018.8462292]

[14] Baveye Y, Cohendet R, Da Silva MP, LeCallet P. Deep learning for image memorability prediction: The EmotionalBias. In: Proc. of the ACM on Multimedia Conf. 2016. 491–495. [doi: 10.1145/2964284.2967269]

[15] Zarezadeh S, Rezaeian M, Sadeghi MT. Image memorability prediction using deep features. In: Proc. of the Electrical Engineering. 2017. 2176–2181. [doi: 10.1109/IranianCEE.2017.7985423]

[16] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. Computer Science, 2014.

[17] Shekhar S, Singal D, Singh H, Kedia M, Shetty A. Show and recall: Learning what makes videos memorable. In: Proc. of the 2017 IEEE Int'l Conf. on Computer Vision Workshops (ICCVW). Venice: IEEE, 2017. 2730–2739. [doi: 10.1109/ICCVW.2017.321]

[18] Demarty CH, Sjöberg M, Ionescu B, Do TT, Gygli M, Duong NQK. In: Proc. of the Mediaeval 2017 Predicting Media Interesting Nesstask. 2017.

[19] Dubey R, Peterson J, Khosla A, Yang M, Ghanem B. What makes an object memorable. In: Proc. of the 2015 IEEE Int'l Conf. on Computer Vision (ICCV). 2015. 1089–1097. https://doi.org/10.1109/ICCV.2015.130

[20] Isola P, Xiao J, Torralba A, Oliva A. What makes an image memorable. In: Proc. of the CVPR. 2011. 145–152. https://doi.org/10.1109/CVPR.2011.5995721

[21] Pennington J, Socher R, Manning CD. GloVe: Global vectors for word representation. In: Proc. of the Empirical Methods in Natural Language Processing (EMNLP). 2014. 1532–1543. http://www.aclweb.org/anthology/D14-1162

[22] Arora S, Liang YY, Ma TY. A simple but tough-to-beat baseline for sentence embeddings. In: Proc. of the ICLR. 2017.

[23] Kiros R, Zhu YK, Salakhutdinov RR, Zemel R, Urtasun R, Torralba A, Fidler S. Skip-thought vectors. In: Advances in Neural Information Processing Systems 28. Curran Associates, Inc., 2015. 3294–3302.

[24] Du T, Bourdev L, Fergus R, Torresani L. Learning spatio temporal features with 3D convolutional networks. In: Proc. of the IEEE Int'l Conf. on Computer Vision. 2015. 4489–4497. [doi: 10.1109/ICCV.2015.510]

[25] Almeida J, Leite NJ, Torres RDS. Comparison of video sequences with histograms of motion patterns. In: Proc. of the IEEE Int'l Conf. on Image Processing. 2011. 3673–3676. [doi: 10.1109/ICIP.2011.6116516]

[26] Carreira J, Zisserman A. Quo Vadis, action recognition? A new modeland the kinetics dataset. In: Proc. of the 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017. 4724–4733. https://doi.org/10.1109/CVPR.2017.502

[27] Haas AF, Guibert M, Foerschner A, Co T, Calhoun S, George E, Hatay M, Dinsdale E, Sandin SA, Smith JE, Vermeij MJA, Felts B, Dustan P, Salamon P, Rohwer F. Can we measure beauty? Computational evaluation of coral reef aesthetics. 2015. https://doi.org/10.7717/peerj.1390

附中文参考文献:

[4] 仲兆满,管燕,胡云,李存华.基于背景和内容的微博用户兴趣挖掘.软件学报,2017,28(2):278–291. http://www.jos.org.cn/1000-9825/5030.htm [doi: 10.13328/j.cnki.jos.005030]

[6] 王朝晖,普园媛,徐丹,祝娟,陶则恩.人像照片的美感质量评价.软件学报,2015,26(Suppl.(2)):20−28. http://www.jos.org.cn/1000-9825/15012.htm

[8] 孔庆超,毛文吉.基于动态演化的讨论帖流行度预测.软件学报,2014,25(12):2767–2776. http://www.jos.org.cn/1000-9825/4730.html [doi: 10.13328/j.cnki.jos.004730]

王帅(1993－),男,学士,CCF 学生会员,主要研究领域为情感计算.

陈师哲(1994－),女,学士,CCF 学生会员,主要研究领域为多模态内容理解.

王维莹(1996－),女,学士,主要研究领域为多媒体计算.

金琴(1972－),女,博士,副教授,博士生导师,CCF 专业会员,主要研究领域为多媒体语义理解,情感计算.