ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 6,Nov. 2016,pp. 1171－1177
DOI:10. 16337/j. 1004-9037. 2016. 06. 012

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax: ＋86-025-84892742

教学视频的文本语义镜头分割和标注

王 敏 王 斌 沈钧戈 高新波

(西安电子科技大学电子工程学院,西安,710071)

摘 要: 为了对教学视频这一专门类别视频进行自动标注,本文首先提取视频中的字幕信息,通过文本预处理后,使用视频中的字幕文本信息内容结合潜在狄利克雷分布(Latent Dirichlet allocation,LDA)主题模型方法获得视频镜头在主题上的概率分布,通过计算主题概率分布差异,进行语义层面镜头分割。然后以镜头为样本,使用安全的半监督支持向量机(Safe semi-supervised support vector machine,S4VM)方法,通过少量的标注镜头样本,完成对未标注镜头的自动标注。实验结果表明,本文方法利用字幕文本信息和LDA模型,有效完成了视频的语义镜头分割,不仅可以对镜头完成标注,而且可以对整个视频进行关键词标注。

关键词: 教学视频;字幕文本;半监督学习;潜在狄利克雷分布

中图分类号: TP391　　**文献标志码**:A

Lecture Video Text Semantic Shot Segmentation and Annotation

Wang Min, Wang Bin, Shen Junge, Gao Xinbo

(School of Electronic Engineering, Xidian University, Xi′an, 710071, China)

Abstract: To automatically annotate a special kind of video, i. e. , lecture videos, a method is proposed to extract caption information from video, Then subtitle information is utilized with latent Dirichlet allocation(LDA). The document distribution probability on the topics is obtained. The distance between these probability distributions is calculated. Finally the semantic shot segmentation is realized. A shot is set as a sample based on safe semi-supervised support vector machine(S4VM) method. A small amount of labeled semantic shots are taken as samples. The unlabeled shots are automatically annotated. Experimental results show that the proposed method can not only effectively complete the shot semantic segmentation, but also annotate key words for the video.

Key words: lecture video;caption text;sem-supervised learning; latent Dirichlet allocation(LDA)

引　　言

教学视频是将要传授给学生的知识、技能等内容制成视频进行传播,目前已成为现代化多媒体教学的重要手段。教学视频的内容包括教师的言谈、手势动作、黑板(白板)板书、投影幕布上的幻灯片、学生

基金项目:国家杰出青年科学基金(61125204)资助项目;陕西省科技攻关(2013K06-30)资助项目;西安电子科技大学宁波信息技术研究院种子基金资助项目。
收稿日期:2014-10-29;**修订日期**:2014-11-09

听课情况以及课堂讨论等,含有丰富的教学信息,已成为在线学习的重要素材。计算机技术、网络技术的日新月异使得在线学习平台发挥着日益重要的作用。如何在海量的教学视频中快速有效地定位某个知识点所对应的视频片段或感兴趣的视频片段,对于高效学习显得尤为重要。但是由于教学视频的场景比较均匀、视觉差异不明显,从而无法有效使用视觉信息进行分析。为了对这些海量教学视频数据进行存储、管理和索引,需要研究高效的检索和搜索方法,而视频标注是视频检索和视频搜索的基础。在地域广泛、教育程度严重区域不平衡的中国,在线学习或者远程教育缩短了教育水平的差距,对于教育而言,教学视频的标注是一项很有意义的工作。教学视频存在以下特点:(1)基本上是在室内录制,反应教学现场情况,背景可能是幻灯片、教师和黑板(白板)板书等,背景比较单一。(2)视频的背景转换较少,大都是幻灯片、白板(黑板或者板书等),教师这些镜头之间的转换,视觉信息无法作为有效信息。借助于一般的场景视频标注方法,对教学视频标注进行分析,一般的视频标注方法主要分为 3 类:手工标注、基于规则的标注方法[1,2]以及基于机器学习的方法[3]。手工标注的方法费时费力,受标注者主观影响较大,因此不适用于海量视频处理。基于规则的标注方法利用相关领域的专家知识建立规则进行标注,但规则对视频语义的刻画能力有限,通用性有限且实用性不高;基于机器学习的方法利用视频中的视觉信息进行标注,该类方法可分为无监督学习、有监督学习和半监督学习。表 1 给出了基于机器学习的标注方法分类。主要思想是通过从视频中提取出的视觉特征集,进行训练、建立模型,最后对能表达视频的特征预测进而对视频进行标注。手工标注和基于规则的标注方法因费时费力,不具备通用性和实用性,不适用于教学视频。流行的机器学习方法先对视频提取视觉特征(颜色、纹理和形状等),通过训练特征集建立模型,最后对视频进行预测、标注。由于教学视频的视觉信息特征比较单一,仅依靠视觉信息,很难对视频作出比较准确的标注。如图 1 所示,通过观察可知,不同镜头不同关键帧的颜色特征可能相同或类似,对于分类标注来说,不具备差异的特征作为特征输入,分类效果往往不理想。

表 1 基于机器学习的视频标注方法

Tab. 1 Video annotation method based on machine learning

类别	具体方法
无监督学习	K-means 聚类[4]
	Parse Coding[5]
	独立子空间分析(Independent subspace analysis, ISA)[6]
有监督学习	核密度估计(Kernel density estimation, KDE)[7]
	高斯混合模型(Gaussians mixture model, GMM)[8,9]
	支持向量机(Support vector machine, SVM)[10]
半监督学习	KDE[11]
	基于图的方法[12]
	SVM[13,14]

为了解决教学视频这一专门视频无法有效使用传统视频标注方法进行标注的问题,本文借助基于内容的文本、视频检索技术[15-17],从语义层次角度完成教学视频标注。提出了基于文本信息并结合潜在狄利克雷分布(Latent Dirichlet allocation, LDA)模型的半监督教学视频标注方法。实验之前,要先采集教学视频数据,为此文章对知名的教育网站进行分析。通过对比分析,了解到多数的优秀教学视频大都是国外知名大学或机构制作的,为了更加有效、有意义地进行教学视频标注,本文采集的实验视频都是英文教学视频,中英文字幕,但只提取英文字幕。语义镜头分割方法过程见图 2,先对视频做结构化处理,依照等间隔进行视频关键帧的提取。通过对各个关键帧进行光标字符识别(Optical character recognition, OCR)提取形成关键帧对应的文档,对重复的文档合并,并进行文本预处理。然后利用无监督的 LDA 方法,得到文档在主题下的概率分布,通过比较相邻关键帧的字幕文档之间的特征距离与预先设

定的阈值关系,对文档进行主题分配,获得语义文本镜头分割结果。最后利用少量的标注镜头样本和大量未标注样本采用半监督学习方法对大量的未知镜头进行语义标注,从而完成整个标注过程。

图 1　教学视频的关键帧视图

Fig. 1　Visual graph of key frames for lecture video

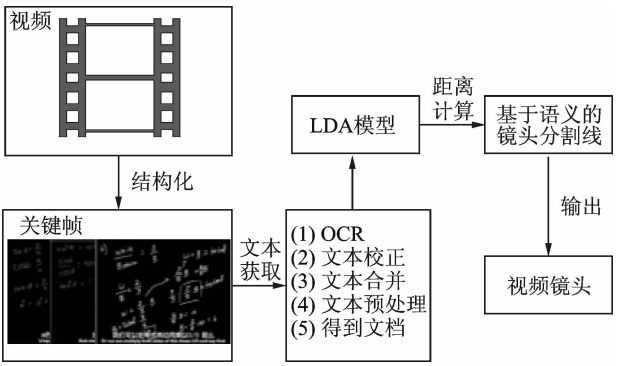

图 2　基于语义的教学视频镜头分割过程

Fig. 2　Shot segmentation procedure of lecture video based on semantic

1　基于语义的镜头分割和半监督学习标注

输入视频集 $V=\{V_1,V_2,\cdots,V_{|V|}\}$,结构化得到关键帧 $K_f=\{k_{f_1},k_{f_2},\cdots,k_{f_{|K|}}\}$,提取相应的字幕文档 $D=\{k_{d_1},k_{d_2},\cdots,k_{d_{|D|}}\}$,使用 LDA 主题模型进行语义镜头分割,得到镜头集合 $S=\{s_1,s_2,\cdots,s_{|S|}\}$,其中 $|V|,|S|,|D|,|K|$ 分别为各个集合的长度。

1.1　字幕文本文档获取

为了能够利用视频的字幕文本信息,本文需要获取字幕文本,获取的过程包括关键帧提取,OCR 处理,文本预处理 3 个过程。

(1) 关键帧提取。先对视频 v_i 结构化处理[18],得到关键帧 k_{f_i},为了保证字幕信息的完整性,使得获取的关键帧包括所有的字幕帧,本文先读取视频,然后按照等间隔差($T=20$)的方法对视频进行关键帧提取,以保证抓取到所有字幕信息。

(2) OCR 处理。OCR 处理包括文本区域定位和文本识别。由于关键帧中的字幕信息一般都是在帧的底部,并且字幕也可能会有双语字幕,为简化处理,本文只提取英文字幕,文本区域定位的范围可以限定在帧的底部,通过多次区域阈值的选择,确定最适合阈值,得到英文字幕的文本区域;然后利用基于灰度的图像匹配方法,将相同含字幕的图片进行处理,若是相同的字幕图片,则只保留一个字幕图片;最后得到所有含字幕的区域图像。确定文本区域之后进行 OCR 识别,本文采用开源 OCR 引擎 Tesseract-OCR[19],结果得到字幕的文本文档 $D=\{k_{d_1},k_{d_2},\cdots,k_{d_{|D|}}\}$。

(3) 文本预处理。OCR 处理后的文档需要做进一步的预处理,得到有效的文本输入数据。对所有的文本进行文本校正,确保所有的单词正确可信;然后过滤文档中的停用词(Stopwords)[20],对一些出现频率很高的但从语义上不影响文本的单词进行过滤(I, was, here, the 等),过滤这些无用词,防止在文本中其大量出现将有用词"淹没"。这可解释为对特征空间降维,获得更加紧凑的文本文档。不仅可以减少计算量以提高后续阶段的运行效率,而且减少输入噪声提高了算法的准确率;最后,对所有文档再进行词干抽取(Stemming)[21],将文本中同一单词的多种形式合并成一个单词,若同一单词的多种形式作为多种特征,不仅增加了特征空间的维数,而且还会分散特征的权重计算,影响文本的正确表示,词干抽取的目的是将由词干派生的词还原为词干,得到最能表示视频的最简洁文档集。

1.2 LDA 主题模型合并语义镜头

文档集合 D 为所有字幕的文档集合,Topic 集合为主题集合,D 中的文档 d 理解为单词序列 $\{w_1, w_2, \cdots, w_n\}$,$t$ 为隐含的主题。w_i 为第 i 个单词,d 有 n 个单词。以文档集合 D 作为 LDA[19,22] 的输入,可以得到两个概率表示。$p(t/k_{d_i}) = p(t_1, t_2, t_3, \cdots, t_T \mid k_{d_i})$ 为给定文档 Document 下,对应主题的概率。$p(w/t_i) = p(w_1, w_2, w_3, \cdots, w_m \mid t_i)$ 表示给定的主题下,生成不同单词的概率。通过比较 $p(t/k_{d_i})$ 与 $p(t/k_{d_j})$ 的距离,得到文档间的相似度,预先设定相似度阈值 T_1,若相似度超过或等于阈值 T_1,则这两个文档属于同一主题,否则属于不同主题,其实际含义表示关键帧之间的相似度比较,若相似度很高则为同一个镜头,否则划分为不同镜头。基于此原理,即可利用字幕文本信息来合成视频的语义镜头。文档 k_{d_i}, k_{d_j} 之间的距离定义为

$$\text{distance} = \sum_{i,j=1}^{D} \| k_{d_i} - k_{d_j} \| \tag{1}$$

1.3 半监督学习自动标注镜头

无监督学习单使用未标记的样本,有监督学习单使用标记样本进行训练,现实问题中,标注代价比较高,有少量的标记样本,但存在大量未标记的数据,因此,本文选择可同时使用两类样本的半监督学习方法。安全的半监督支持向量机(Safe semi-supervised support vector machine, S4VM)[23] 在使用未标记数据时,能保证不产生性能下降,在给定的众多不同间隔较大的分界线时,通过优化未标记样本的类别划分,使得在最坏的情况下,最大化提升相对于只使用标记样本的 SVM 性能。基于半监督学习的视频标注过程如图 3 所示。将镜头 s_i 作为一个样本,当属于某一个主题时,Y 标记为 1,不属于标记为 -1,用少量标记的镜头样本和大量未标记的镜头样本,采用 S4VM 的方法做半监督训练,得到训练模型后,对未标记样本进行自动标注。

标记样本集 $\{\boldsymbol{s}_i, y_i\}_{i=1}^{l}$,未标记样本集 $\{\hat{\boldsymbol{s}}_j\}_{j=1}^{u}$,通过半监督学习函数 $f: S \to Y$ 可准确地对样本预测标记。其中 $\boldsymbol{s}_i, \boldsymbol{s}_j \in S$ 均为 $|S|$ 维向量,$y \in \{\pm 1\}$ 是样本 $\boldsymbol{s}_i$ 的标记,l, u 分别为标记样本集、未标记样本集的长度。$h(f, \hat{y})$ 为半监督支持向量机需要优化的目标函数,具体可定义为

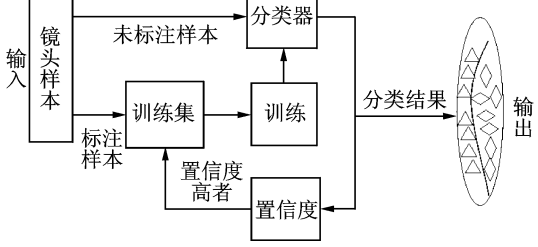

图 3 基于半监督的标注过程

Fig. 3 Semi-supervised annotation procedure

$$h(f,\hat{y})=\frac{\|f\|_H}{2}+C_1\sum_{i=1}^{l}l(y_i,f(\boldsymbol{s}_i))+C_2\sum_{j=1}^{u}l(\hat{y}_j,f(\hat{s}_j)) \tag{2}$$

目标是要找到多个间隔较大的低密度分界线$\{f_t\}_{t=1}^{T}$以及相应的类别划分$\{\hat{y}_t\}_{t=1}^{T}$,使得下面的函数最小化

$$\min_{\{f_t,y_t\in\beta\}_{t=1}^{T}}\sum_{t=1}^{T}h(f_t,y_t)+M\Omega(\{y_t\}_{t=1}^{T}) \tag{3}$$

式中:T为分界线数量,Ω为惩罚函数,M为常数,用来保证差异性。最小化式(3)可以保证分界线的差异性和较大的间隔,不失一般性,假设f为线性函数,$f(x)=w'\phi(s)+b$,则需要求解的最优化问题表示为

$$\begin{aligned}&\min_{\{w_t,b_t,y_t\in\beta\}_{t=1}^{T}}\sum_{t=1}^{T}\left(\frac{1}{2}\|w_t\|^2+C_1\sum_{i=1}^{l}\xi_i+C_2\sum_{j=1}^{u}\xi_j\right)+M\Omega(\{y_t\}_{t=1}^{T})\\&\text{s.t.}\quad y_i(w'_t\phi(\boldsymbol{s}_i+b_t))\geqslant 1-\xi_i\quad \xi_i\geqslant 0,\hat{j}_{t,j}(w'_t\phi(\hat{\boldsymbol{s}}_j+b_t)\geqslant 1-\hat{\xi}_i\qquad \hat{\xi}_i\geqslant 0\\&\forall i=1,\cdots,l;\forall j=1,\cdots,u;\forall t=1,\cdots,T\end{aligned} \tag{4}$$

通过找到多个低密度分界线,使得目标函数最优化,完成半监督学习过程,从而完成对未标注样本的预测,实现对视频镜头的关键词标注。

2　实验结果及分析

2.1　实验结果

实验中使用的数据为从网易公开课等教育网站下载的教学视频,先对少量视频进行镜头的标记作为标记样本,其余的样本作为未标记样本。根据上述方法,本文选定了数学科目类的34个教学视频,对34个教学视频进行结构化,得到关键帧,然后分别提取关键帧的字幕信息,同时合并相同关键帧的字幕文本,得到不同的字幕文档文本共4 933。对所有文档文本预处理后,得到关键有效的文本文档集。使用LDA模型,对视频的字幕文本信息进行主题模型分析,根据下载到数据库中的视频内容,进行大致分析后,确定将主题数量设为11,得到每个文档属于不同主题下的概率$p(t_1,t_2,\cdots,t_n|k_{d_i})$,通过比较主题概率分布差异,对文档进行分类,对应得到视频语义的镜头分割。算法的平均标注分类正确率表示为

$$\text{正确率}=\frac{\text{正确分类个数}}{\text{总的分类个数}}=\frac{\text{正确检出数}}{\text{正确检出数}+\text{错误检出数}} \tag{5}$$

设定不同的阈值T_2,镜头分割的效果如表2所示。

表2　不同阈值T_2对应的分割结果

Tab.2　Results of different threshold T_2

视频	阈值T_2	关键帧数	镜头转换数	正确率/%
math01～math34	0.455	4 933	141	90.3
	0.460	4 933	141	91.7
	0.465	4 933	141	92.1
	0.470	4 933	141	90.4
	0.475	4 933	141	89.8

利用S4VM[23]半监督学习方法,对分割出的镜头进行标记。实验对比的算法为:直推式支持向量机(Transductive SVM,TSVM)[24],meanSVM[25]的两种实现(meanSVM-iter和meanSVM-mkl)结合和S4VM。实验结果如表3所示。

2.2 结果分析

表 2 表明，利用字幕信息结合 LDA 的方法，通过选择适合的阈值 T_2，能够有效地对教学视频进行语义层面上的镜头分割，解决了传统方法不能有效标注教学视频的问题。阈值过大，文档间的差异不能够精确描述，导致将不同类别划分为同一类别；阈值过小，文档间的差异会被放大，导致应属于同一类别的而错分成两类，表 4 中，给出了不同阈值 T_2 下的视频 v_{29} math29 镜头分割的结果及镜头所属主题的关键词；同时，从表 3 中可看出 S4VM 半监督学习方法能够很好地对未标记样本进行预测标注。

表 3 各类算法的实验结果对比表

Tab. 3 Comparison of algorithms result

标记样本数量	算法	正确率/%
10	TSVM	75.35
10	meanSVM-iter	76.20
10	meanSVM-mkl	79.40
10	S4VM	73.10
80	TSVM	91.70
80	meanSVM-iter	91.80
80	meanSVM-mkl	89.13
80	S4VM	92.45

表 4 不同 T_2 下 v_{29} 镜头分割结果及关键词

Tab. 4 v_{29} shot segmentations and keywords under different threshold T_2

视频	阈值 T_2	镜头转换数	分割结果	关键词
math29	0.450	5	8	minus, value, equation, line, angle, number, time, plus, function, power, line, angle
math29	0.465	5	5	equation, value, line, angle, number, line function, power, angel
math29	0.470	5	4	equation, line, number, power, line, angle

3 结束语

不同于一般场景的视频，教学视频场景转换较少，视觉信息对关键帧或镜头的刻画能力有限，传统的基于视觉信息的视频标注方法效果不理想。本文针对有字幕的教学视频，利用视频中的字幕文本信息采用 LDA 模型对视频进行语义镜头分割，然后结合半监督学习方法，对分割后的镜头进行自动标注。实验表明，该方法能够有效地在语义层面上分割教学视频的镜头，半监督学习的方法也可以有效地对镜头完成自动标注。本文方法还有待于进一步提升，比如教学视频中会有一些幻灯片内容，对于其中的文本信息，有待于进一步探索和研究，挖掘更多的文本信息对于教学视频的标注有着重要意义。

参考文献：

[1] Xie Lexing. Structure analysis of soccer video with domain knowledge and hidden Markov models [J]. Pattern Recognition Letters, 2004, 25(7): 767-775.

[2] Ekin A, Tekalp A M, Mehrotra R. Automatic soccer video analysis and summarization [J]. Image Processing IEEE Transactions on, 2003, 12(7): 796-807.

[3] Mei Tao, Wang Meng, Hua Xiansheng, et al. Coherent image annotation by learning semantic distance[C]// Computer Vision and Pattern Recognition. [S. l.]: IEEE, 2008: 1-8.

[4] Hartigan J A, Manchek A W. Algorithm AS 136: A k-means clustering algorithm [J]. Applied Statistics, 1979, 28(1): 100-108.

[5] Olshausen B A, David J F. Sparse coding with an over complete basis set: A strategy employed by V1[J]. Vision Research, 1997, 37(23): 3311-3325.

[6] Le Quoc V. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis [C]// Computer Vision and Pattern Recognition (CVPR). [S. l.]: IEEE, 2011: 3361-3368.

[7] Sheather S J, Michael C J. A reliable data-based bandwidth selection method for kernel density estimation [J]. Journal of the Royal Statistical Society: Series B(Methodological), 1991, 53(3): 683-690.

[8] Fan Jianping, Luo Hangzai, Lin Xiaodong. Semantic video classification by integrating flexible mixture model with adaptive EM algorithm [C]// Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval. [S. l.]: ACM, 2003: 9-16.

[9] Wu Jun. An online-optimized incremental learning framework for video semantic classification [C]// Proceedings of the 12th Annual ACM International Conference on Multimedia. [S. l.]: ACM, 2004: 320-323.
[10] Yanagawa A. Columbia University's baseline detectors for 374 lscom semantic visual concepts [J]. Columbia University Advent Technical Report, 2007,18(1):222-2006.
[11] Vondrick C, Donald P, Deva R. Efficiently scaling up crowdsourced video annotation[J]. International Journal of Computer Vision, 2013,101(1): 184-204.
[12] Wang Meng, Hua Xiansheng, Mei Tao, et al. Semi-supervised kernel density estimation for video annotation [J]. Computer Vision and Image Understanding,2009,113(3):384-396.
[13] Xu Xinshun. Semi-supervised multi-instance multi-label learning for video annotation task [C]// Proceedings of the 20th ACM International Conference on Multimedia. [S. l.]:ACM, 2012: 737-740.
[14] 万建平,彭天强,李弼程. 基于证据理论的视频语义概念检测[J]. 数据采集与处理,2011,26(5):536-541.
Wan Jianping,Peng Tianqiang,Li Bicheng. Video semantic concept detection based on evidence theory[J]. Journal of Data Acquisition and Processing,2011,26(5):536-541.
[15] 李广东,高新波,赵力. 一种基于静帧特征分析的视频检索方法[J]. 数据采集与处理,2011,26(3):334-338.
Li Guangdong,Gao Xinbo,Zhao Li. Video retrieval based on static frame feature analysis[J]. Journal of Data Acquisition and Processing,2011,26(3):334-338.
[16] 程娟,平西建,周冠玮. 基于多特征和 SVM 的文本图像版面分类方法[J]. 数据采集与处理,2008,23(5):569-574.
Cheng Juan,Ping Xijian,Zhou Guanwei. Layout analysis based on multi-feature and SVM[J]. Journal of Data Acquisition and Processing,2008,23(5):569-574.
[17] 周之昊,王士同. 在线聚类算法用于基于内容的镜头检索[J]. 数据采集与处理,2008,23(1):84-88.
Zhou Zhihao,Wang Shitong. Content-based shot retrieval by on-line clustering algorithm[J]. Journal of Data Acquisition and Processing,2008,23(1):84-88.
[18] Rui Y, Huang T S. Mehrotra S. Constructing table-of-content for videos [J]. Multimedia Systems, 1999,7(5):359-368.
[19] Smith R. An overview of the tesseract OCR engine [C]// Proceedings of the 9th International Conference on Document Analysis and Recognition. [S. l.]:IEEE, 2007: 629-633.
[20] Wilbur W J, Karl S. The automatic identification of stop words[J]. Journal of Information Science, 1992,18(1): 45-55.
[21] Hull D A. Stemming algorithms: A case study for detailed evaluation [J]. JASIS,1996,47(1): 70-84.
[22] Rosen-Zvi M. The author-topic model for authors and documents [C]// Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. [S. l.]:AUAI Press, 2004: 487-494.
[23] Li Yufeng, Zhou Zhihua. Towards making ublabeled data never hurt [C]// Proceedings of the 28th International Conference on Machine Learning (ICML'11). Bellevue: WA,2011: 1-1.
[24] Joachims T. Transductive inference for text classification using support vector machines [C]// Proceedings of the 16st International Confernce on Machine Learning. [S. l.]:ACM, 1999: 200-209.
[25] Bennett K, Ayhan D. Semi-supervised support vector machines [J]. Advances in Neural Information Processing Systems, 1999,9(2): 368-374.

作者简介:

王敏(1989-),女,硕士研究生,研究方向:图像视频内容分析处理,E-mail:wangminxidian@126.com。

王斌(1977-),男,副教授,研究方向:图像分割与分析。

沈钧戈(1987-),女,博士研究生,研究方向:模式识别与智能系统。

高新波(1972-),男,教授,长江学者,研究方向:机器学习与计算智能、视觉感知与人脑认知和医学影像处理与科学可视化。