

文章编号:1001-9081(2008)01-0199-03

一种多模态信息融合的视频检索模型

张 静¹, 俞 辉²

(1. 华东理工大学 计算机科学与工程系, 上海 200237; 2. 复旦大学 计算机科学与工程系, 上海 200433)

(jingzhang@ecust.edu.cn)

摘要:针对包含复杂语义信息的视频检索的需要,提出了一种基于关系代数的多模态信息融合视频检索模型,该模型充分利用视频包含的文本、图像、高层语义概念等多模态特征,构造了对应于多个视频特征的查询模块,并创新地使用关系代数表达式对查询得到的多模态信息进行融合。实验表明,该模型能够充分发挥多模型视频检索及基于关系代数表达式的融合策略在复杂语义视频检索中的优势,得到较好的查询结果。

关键词:TRECVID; 视频检索; 多模态信息融合; 关系代数表达式

中图分类号: TP391 文献标志码:A

Video retrieval model based on multimodal information fusion

ZHANG Jing¹, YU Hui²

(1. Department of Computer Science and Engineering, East China University of Science and Technology, Shanghai 200237, China;

2. Department of Computer Science and Engineering, Fudan University, Shanghai 200433, China)

Abstract: In allusion to the complex requirement of query, a new video retrieval model based on multimodal information fusion was brought forward in this paper. It included multi-models like text retrieval, image query, semantic features extraction, and used relational algebra expression to fuse these multimodal information. Experimental results demonstrate that our method could fully utilize the advantages of multimodal information fusion based on relational expression in video retrieval, and achieve good performance on complex semantic video retrieval.

Key words: TRECVID; video retrieval; multimodal information fusion; relation algebra expression

随着计算机软硬件和网络技术的快速发展,视频信息已经逐渐成为信息传输和存储的主体。如何在大规模视频信息上进行有效的基于内容的检索也已成为当前多媒体领域一个重要研究课题。由于视频由多种媒体形式共同构成,其表达的内容也包含在各种媒体格式之中,因而有效的视频检索需要对视频数据进行分析,综合利用从图像、文本、语音等媒体格式中提取的多模态信息进行查询。本文针对 TRECVID^[1]中的搜索任务提出了一种多模态信息融合的视频检索模型,综合利用了文本、图像、高层语义概念等多模态信息,并采用关系代数表达式对多模态信息进行有效融合。实验证明该视频检索模型对于复杂语义的视频查询十分有效。

1 相关研究

正是因为视频具有复杂的表现形式以及包含多种媒体格式等特点,多模态信息融合的视频检索方法便成为视频检索研究重要的组成部分。文献[2]提出一种基于伪相关反馈的多模态信息融合视频检索方法。该方法采用文本、图像、特定镜头检测,以及人脸识别等多模型,并利用线性融合的方法对多模型的检索结果进行融合。在融合权重的设计方面,提出了两种训练模式:基于查询分类的模式和基于查询题目的模式^{[3][207]}。文献[4]借鉴了文献[2]中的方法,在视频检索中融合了语音识别(ASR)、字符识别(VOCR)、说话人身份识别和人脸识别等多模态信息,并采用基于查询分类的方法训练线性融合的权重对多模态信息进行融合。

上述视频检索方法均采用 LR(Logistic Regression)训练

得到多模态信息融合的权重^{[2][302][4][143]},这种权重设置方法对训练样本非常依赖,如果训练样本选择得不好,或者分布与测试样本相差很大,都会极大地影响权重设置的准确性,从而影响融合效果。针对这一点的不足,本文提出了一种基于关系代数的多模态信息融合视频检索模型,该模型利用关系代数表达式对查询得到的多模态信息进行融合,避免了线性融合中权重设置不当对查询结果的影响。

2 基于关系代数的多模态信息融合视频检索模型

视频是一种没有结构的流数据,是集图像、声音和文本为一体的综合性媒体信息。正是由于视频结构的复杂性和难描述性,单一的特征或模型很难得到较好的查询效果。多模态信息融合的方法针对视频的图像、声音、文本等多模态信息分别进行查询,并通过有效融合得到优于任何单一检索模块的查询结果。本文提出一种基于关系代数的多模态信息融合的视频检索模型,如图 1 所示。该检索模型把视频检索分成文本、图像、高层语义概念、相机运动等若干方面,分别针对视频的不同特征进行检索,然后利用关系代数表达式完成多模态信息融合。下面重点介绍该检索模型中的特征查询模块以及融合排序模块。

2.1 特征检索模块

文中提出的多模态信息融合的视频检索模型将视频信息按照不同的模态进行分解,根据视频信息的不同特征,如文本、图像、高层语义概念等构造出多个表达视频特征的检索模块,这些检索模块是非同构的,它们按照不同的方法建立以实

收稿日期:2007-07-30;修回日期:2007-09-29。 基金项目:国家自然科学基金资助项目(60533100)。

作者简介:张静(1978-),女,河南三门峡人,讲师,博士,主要研究方向:高维索引结构、多媒体信息检索; 俞辉(1983-),男,江苏南通人,硕士研究生,主要研究方向:视频信息检索。

现最佳的检索结果，并且可以根据特征数据库的情况随时添加或者删减，使检索系统实现动态更新。

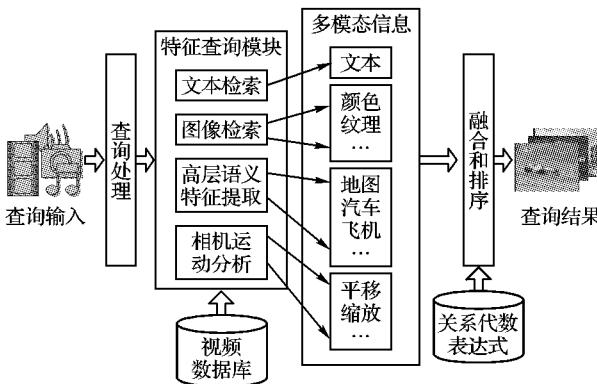


图 1 多模态信息融合视频检索模型

文本检索模块主要针对视频中包含的文本信息进行查询，它将自动语音识别得到的脚本信息，字符识别得到的画面文字信息和从视频解码中得到的字幕信息(Close Caption)进行综合整理，并对应到相应的镜头，然后利用布尔查询方法针对查询处理得到的关键词在已建立的视频文本特征索引结构上进行查询。该模块利用 $W = TF \times IDF$ 计算查询词在查询结果中的权重，并利用其计算每一个查询得到的镜头的置信度，按照置信度的值对结果进行排序，其中 TF 和 IDF 分别代表单词频率(Term Frequency)以及逆文档频率(Inverse Document Frequency)。考虑到通过语音识别得到的文本信息与镜头内容不能准确对应，算法中设定了一定大小的查询窗口，对查询的结果进行相应的扩展，扩展镜头的置信度由一个指数递减函数计算，即 $w(S_i) = w(S_0) \times e^{-i}$ ，其中 S_0 表示文本检索得到镜头， S_i 表示 S_0 的相邻镜头， $w(S_0)$ 表示镜头 S_0 的置信度， i 表示 S_i 距离 S_0 的镜头数。该计算公式假定镜头位置距离文本关键词越近，则其包含该关键词的可能性越大。

图像检索模块提取了镜头中关键帧的颜色和纹理方面的若干特征形成特征库，提供全局特征的近似查询。该模块根据查询的需要可以提取多种全局特征，如：HSV 颜色直方图、LAB 颜色直方图、YUV 颜色直方图、GABOR 纹理特征以及 MPEG 7 中的若干视觉描述子，包括 SCD(Scalable Color Descriptor)、CLD(Color Layout Descriptor)、EHD(Edge Histogram Descriptor)等。这些全局特征检索模块可以根据查询的需要随时进行添加或者删减。查询时计算从样例视频中提取的特征与视频特征库中特征的距离，从而得到相似视频。

高层语义概念分类器是针对给定的高层语义概念建立分类器。分类器采用多模型融合以及机器学习的方法进行构造。为某一个高层语义概念构造分类器时，首先将训练集中的每一幅图像分割成 5×5 的小块，并针对每一个小块提取 CLD、SCD、LAB 颜色直方图、EHD 和 GABOR 纹理特征这五种低层特征形成高维特征矢量，由于提取的特征矢量维数过高，因此在训练分类器之前先需要对其进行降维操作，这里采用 PCA(Principal Components Analysis) 对得到的高维特征数据进行降维，然后利用支持向量机^[6,7]针对这五种特征分别进行训练，得出针对不同特征的语义概念分类器，预测时将这些单个分类器的分类结果通过线性融合得到最终的分类结

果。分类器构造以及预测过程如图 2 所示。

相机运动分类器是根据文献[8]中提出的基于特征的相机运动检测算法完成的。该方法从压缩视频流中提取 P 帧的运动向量，并根据镜头中连续 P 帧的运动统计特性，构造分类器完成不同相机运动的分类。算法首先对视频数据进行中值滤波以去掉视频中的部分噪音，然后根据运动矢量的相互关系提取 14 维的运动特征向量刻画视频中的运动，然后利用基于规则的决策树对运动特征向量进行分类，分析得出镜头中的若干种相机运动。利用相机运动分类器进行查询时，只需从视频库中提取每个镜头 14 维的运动特征向量，然后在训练好的分类器进行预测即可得到每个镜头所包含的相机运动情况。通过分析镜头的相机运动能够得到一些深层的视频内容信息，为复杂语义的视频检索提供帮助。

2.2 融合与排序模块

该模型利用关系代数表达式对多模态搜索结果进行融合，由于关系代数本身就是针对关系数据库中查询操作的运算集合，其中定义的各种关系运算符具有很强的表达能力，能够有效地表示各类复杂的查询要求，而多个检索模块查询得到的查询结果又是相对独立分布的多模态信息，类似于关系数据库的表结构，因此可采用关系代数表达式对搜索结果进行融合。下面通过一个实例介绍利用关系代数表达式对多模态信息进行融合的过程。

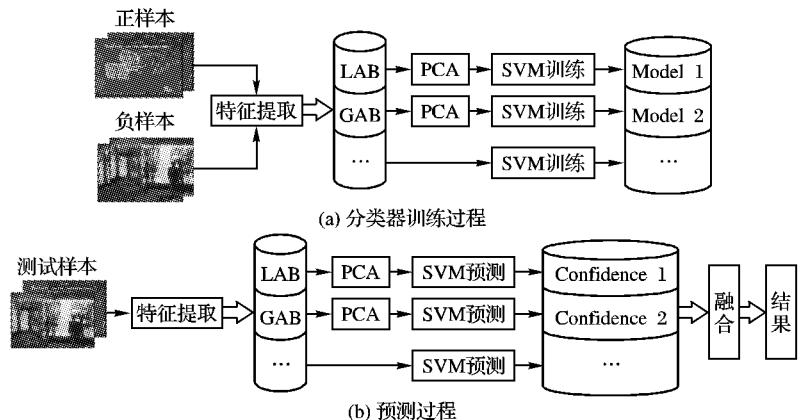


图 2 高层语义概念分类器

例如查询题目为：检索办公室场景。利用本文提出的视频检索模型，通过已定义的检索模块进行查询得到如下多模态查询结果：文本(T)、HSV 颜色特征(C)、LAB 颜色特征(L)、GABOR 纹理特征(G)、EHD 纹理特征(E)、高层语义特征 Office(O)、高层语义特征 Anchor Shot(A)，这些搜索结果都是独立的二元表结构，包含两个属性：镜头编号(ShotId)和置信度(Confidence)。可以表示成如下关系：

$$\begin{array}{ll} T_{\text{office}}(\text{ShotId}, \text{Confidence}) & C_{\text{office}}(\text{ShotId}, \text{Confidence}) \\ L_{\text{office}}(\text{ShotId}, \text{Confidence}) & G_{\text{office}}(\text{ShotId}, \text{Confidence}) \\ E_{\text{office}}(\text{ShotId}, \text{Confidence}) & O(\text{ShotId}, \text{Confidence}) \\ A(\text{ShotId}, \text{Confidence}) & \end{array}$$

根据现有的多模态搜索结果以及查询要求可以设计相应的关系代数表达式如下：

$$\begin{aligned} E_{\text{office}} = & (\pi_{\text{ShotId}}(\delta_{\text{Confidence} > \text{Th}_1}(T_{\text{office}})) \cup \\ & \pi_{\text{ShotId}}(\delta_{\text{Confidence} > \text{Th}_2}(C_{\text{office}})) \cup \\ & \pi_{\text{ShotId}}(\delta_{\text{Confidence} > \text{Th}_3}(L_{\text{office}})) \cup \\ & \pi_{\text{ShotId}}(\delta_{\text{Confidence} > \text{Th}_4}(G_{\text{office}})) \cup \\ & \pi_{\text{ShotId}}(\delta_{\text{Confidence} > \text{Th}_5}(E_{\text{office}})) \cup \\ & \pi_{\text{ShotId}}(\delta_{\text{Confidence} > \text{Th}_6}(O)) - \end{aligned}$$

$$\pi_{Should}(\delta_{Confidence} > Thg) \quad (1)$$

其中: $Th_t, Th_c, Th_l, Th_g, Th_e, Th_o, Th_a$ 分别为 $T_{office}, C_{office}, L_{office}, G_{office}, E_{office}, O$ 和 A 的置信度阈值。多模块检索的结果,通过关系代数表达式的融合便可以得到相关镜头的集合。排序方法是将关系代数表达式中所有采用“并”的方式融合的那些多模态信息的置信度值进行归一化并求和作为排序的依据,排序的结果就是最后的查询结果。

3 实验

3.1 实验数据与评估标准

为了验证算法的有效性并使实验结果具有可参考性,实验按照 TRECVID 中搜索任务的要求进行,实验数据为 TRECVID 2005 提供的约 80 小时的新闻视频,查询题目为包括特定人物、物体、场景搜索在内的 24 个包含复杂语义内容的查询题目^[1]。

评估除了沿用查全率和准确率这两个传统的标准外,又引入了针对每一次搜索结果的 AP (Average Precision) 和针对每一组搜索结果的 MAP (Mean Average Precision)。 AP 是与查询结果的排序有关的度量方法,它的定义如下:对于一个给定的查询,如果仅仅考虑结果排序中前 N_r 个查询结果的成绩,则可以定义平均准确率(AP)如下:

$$AP = \frac{\sum_{i=1}^N \rho_i P_i}{N_{GT}} \quad (2)$$

其中:

$$\rho_i = \begin{cases} 1, & \text{第 } i \text{ 个结果是正确的} \\ 0, & \text{第 } i \text{ 个结果是错误的} \end{cases}$$

$$P_i = \text{前 } i \text{ 个结果中正确的个数} / i$$

N 为计算 AP 的深度。

N_{GT} 是 GT 中正样本的个数。

由上面的定义可知, AP 倾向于排序比较好的查询结果,并且可以用来比较不同规模的结果集。 MAP 则是对一组搜索结果中所有题目的 AP 求平均值。假定在一组搜索任务中共有 n 个查询题目,则 MAP 定义如下:

$$MAP = \left(\sum_{i=1}^n AP_i \right) / n \quad (3)$$

3.2 实验结果

文中提出的视频检索模型包括文本信息,全局特征,高层语义特征和相机运动特征多个检索模块。文本信息主要包括通过语音识别和机器翻译得到的对应视频的脚本信息和通过对视频进行光学字符识别得到的画面上的字符信息。全局特征是指从视频关键帧上提取的颜色、纹理分布方面的特征,主要包括颜色特征 HSVCH 和 LABCH,纹理特征 GABOR 和边缘特征 EHD。高层语义特征包含通过机器学习方法得到的 14 个高层语义特征分类器,它们分别是 airplane, boat, building, car, fire, map, marching, meeting, military, office, road, sports, waterfront, anchor shot。相机运动特征包括 7 种相机运动特征 pan left, pan right, tilt up, tilt down, zoom in, zoom out 和 still。

首先利用本文提出的基于关系代数的多模态信息融合的视频检索模型对 TRECVID2005 提供的 24 个查询题目进行测试,得到的平均准确率(AP)如图 3 所示。从实验结果可以看出,该视频检索模型对大部分的查询题目都能够得到较好的平均准确率,特别是包含复杂语义的事件类题目,如:156

(tennis count) 的 AP 值为 88.86%;172 (office setting) 为 87.87%;164 (ship or boat) 为 63.03%;161 (people with banners or signs) 的 AP 值为 61.65%。图 3 中其他编号所对应的查询题目请参阅 TRECVID 官方网站^[1]。

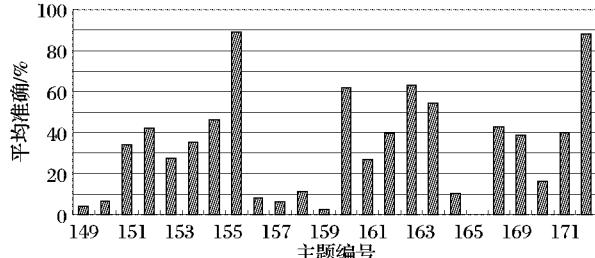


图 3 针对所有查询题目查询所得的平均准确率

由于复杂的查询题目包含众多语义特征,采用多模态信息融合的方法可以有效地综合视频的多个特征,以得到满足多方面查询要求的最优结果。文中介绍的视频检索模型充分考虑了多模态信息融合在语义视频检索中的优势,并将在关系数据库中用于处理复杂查询的关系代数表达式引入视频检索对多模态信息进行有效融合,该方法能够将包含多种语义信息的多模态查询结果有效地结合起来以满足多概念的查询要求,从而得到较好的查询结果。

为了进一步说明该算法的有效性,本文做了一组对比实验,实验中比较了采用相同的多模态信息但不同融合方式的检索模型的查询结果。Run 1 采用的是利用关系代数表达式进行融合的方法;Run 2 和 Run 3 均采用基于 LR^[9] 的线性融合方法,只是采用了不同的权重训练策略,Run 2 采用的是伪相关反馈(Co-retrieval)和基于查询题目的权重训练方式^{[3]208},而 Run 3 采用的是手工标注样本和基于查询题目的权重训练方式^{[2]304}。

表 1 中给出了三组实验在深度为 50 和 100 时的 MAP 值,即在查询结果排序中第 50 和 100 排名处计算得到的 MAP 值。从实验结果看到,在深度为 50 时,Run 1 的 MAP 值比 Run 2 高出 9.66%,比 Run 3 高出 19.93%;在深度为 100 时,仍比 Run 2 高出 7%,比 Run 3 高出 15.25%。由于这三组实验均采用上文介绍的视频检索模块进行多模态信息搜索,只是采用了不同的融合方式,因而充分证明了基于关系代数的多模态信息融合方法在复杂语义的视频检索上优于基于逻辑回归的线性融合方法。另外,Run 2 的搜索结果明显优于 Run 3,也说明了在测试集上通过伪相关反馈训练得到的合并权重优于在训练集上通过标注信息训练得到的权重。

表 1 三种多模态信息融合视频检索方法的查询结果

MAP 值	深度为 50	深度为 100
Run 1	0.3897	0.3302
Run 2	0.2931	0.2602
Run 3	0.1904	0.1777

分析实验结果可以发现,基于线性融合的方法其融合效果明显不如基于关系代数表达式的融合方法。主要是因为基于线性融合的方法其性能在很大程度上依赖于权重的设置,而权重的训练与样本质量密切相关,如果训练样本的质量不好或者分布与测试样本相差很大,则会严重地干扰权重设置的合理性,从而影响融合效果。另外,Run 2 的查询结果明显优于 Run 3,说明了利用伪相关反馈在测试集上在线训练权

(下转第 213 页)

取加权均值便得到运动模糊角度的估计。

3 实验结果

本文算法针对运动模糊角度识别的改进,实验主要基于人工生成的运动模糊图像。在 Matlab 6.5 的环境下分别对标准图像库中的图像 BABOO. BMP 和 PEPPERS. BMP,采用在文献[1]中的数据对其进行人工的运动模糊效果,然后分别应用相应算法进行运动模糊角度识别,结果比较见表 1。

表 1 FBD 和 FBDII 对两幅标准图的运算结果比较

图像名	模糊带宽 (像素)	模糊 角度(°)	算法 FBD 的 估计结果(°)	算法 FBDII 的 估计结果(°)
BABOO. BMP	15	0	0	0.0
	15	30	30	30.0
	23	18	19	18.7
	38	30	30	30.0
	60	40	40	40.0
PEPPERS. BMP	15	45	45	45.0
	20	50	45	47.3
	30	60	56	57.3
	35	15	21	16.7
	50	0	0	0.0

表 1 结果表明,针对两幅标准图,当算法 FBD 得到的结果误差比较大的时候,算法 FBDII 得到的解误差相对小,较小误差的结果将有助于运动模糊图像的精确复原。

表 2 分别应用 FBD 和 FBDII 进行 100 次实验的结果误差统计

各类情况	算法 FBD 的 估计结果(°)	算法 FBDII 的 估计结果(°)
最佳估计值误差	0.00	0.00
最坏情况估计误差	9.00	5.90
平均估计误差	2.41	1.84
估计误差的标准方差	10.89	6.84

(上接第 201 页)

重的方法明显优于在训练数据集上离线训练权重的方法。也再次证明了训练数据集和测试数据集在特征分布上的差异,将严重影响线性融合中权重训练的准确性,进而影响查询结果。

4 结语

本章介绍了一种新的多模态信息融合视频检索模型,该模型充分考虑了多模态信息查询和融合在复杂语义视频检索中的重要性,构造了对应于文本、图像、高层语义等多个视频特征的查询模块,并有效地利用在关系数据库中用于处理复杂查询的关系代数表达式对多模态信息进行融合,将多模态查询结果有效地结合起来以满足复杂语义内容的查询要求。实验结果表明,本文提出的视频检索模型对于复杂语义内容的视频检索能够得到较好的查询结果。并且在相同的多模态信息条件下,本文提出的基于关系代数的多模态信息融合方法明显优于基于逻辑回归的线性融合方法,能够得到更准确的查询结果。

下一步的工作将从引入相关反馈和提取有效特征两个方面进一步改进检索模型的性能。

参考文献:

- [1] TRECVID[EB/OL].[2007-08-12].<http://www-nplir.nist.gov/projects/trecvid/>.

针对更多图像,更多情况下的运动模糊步长和角度进行人工运动模糊效果后,再分别调用 FBD 和 FBDII 对模糊图像进行模糊角度识别(100 次实验)的结果误差统计见表 2。

表 2 结果表明,算法 FBDII 比 FBD 在最差情况下误差要小,平均误差和标准方差也都相对要小。因此本文提出的算法 FBDII 比 FBD 更加稳定。

4 结语

本文在文献[1]提出的基于霍夫变换的图像运动模糊角度识别法的基础上,对该方法进行了改进,改进主要分为两方面:1)在预处理操作上,增加了更多的处理,使得得到的边界二值图杂点更少,边界更清晰。2)在霍夫变换操作中,由原来的只寻找一条直线变为寻找最合适的三条直线,然后再将三条直线对应的角度做均值得到结果。实验结果表明,改进算法能得到更为精确的结果,并且比原算法有更小的平均误差和标准方差,因而更加稳定。

参考文献:

- [1] LOKHANDE R. Identification of parameters and restoration of motion blurred images [M]. Dijon: ACM Press, 2006: 301 - 305.
- [2] FABIAN R, MALAH D. Robust identification of motion and out-of-focus blur parameters from blurred and noisy images [J]. CVGIP: Graphical, Models and Image Processing, 1991(53): 403 - 412.
- [3] LI Q, YOSHIDA Y. Parameter estimation and restoration for motion blurred images [J]. IEICE Transaction Fundamentals, 1997, E80-A (8): 1430 - 1437.
- [4] MOGHADDAM M E, JAMZAD M. Motion blur identification in noisy images using fuzzy sets [J]. IEEE international Symposium on Signal Processing and Information Technology, 2005(5): 862 - 866.
- [5] GONZALEZ R C, WOODS R E. 数字图像处理[M]. 2 版. 阮秋琦, 阮宇智, 译. 北京: 电子工业出版社, 2003: 175 - 176.

- [2] HAUPTMANNN A, CHEN M Y, CHRISTEL M, et al. Confounded expectations: Informed media at TRECVID 2004[C]// NIST TRECVID 2004 Workshop. Gaithersburg: TRECVID press, 2004.
- [3] YAN R, HAUPTMANNN A. Co-retrieval: A boosted reranking approach for multimedia retrieval[C]// International Conference on Image and Video Retrieval. Dublin: ACM Press, 2004: 21 - 23.
- [4] CHUA T S, NEO S Y, LI K Y. TRECVID 2004 search and feature extraction task by NUS PRIS[C]// NIST TRECVID workshop. Gaithersburg: TRECVID press, 2004.
- [5] YANG H, CHAISOM L, ZHAO Y L, et al. VideoQA: Question answering on news video[C]// Proceeding of the ACM conference on Multimedia. [S. l.]: ACM Press, 2003.
- [6] BURGES C J C. A tutorial on support vector machines for pattern recognition [J]. Data Mining and Knowledge Discovery, 1998, 2 (2): 121 - 167.
- [7] SVM[EB/OL].[2005-08-12].<http://www.csie.ntu.edu.tw/~cjlin/libsvm/>.
- [8] ZHU X, ELMAGARMID A K, XUE X, et al. InsightVideo: Toward hierarchical video content organization for efficient browsing, summarization and retrieval [J]. IEEE Transaction on Multimedia, 2005, (7): 648 - 666.
- [9] HO T K, HULL J J, SRIHARI S N, et al. Combination of decisions by multiple classifiers [M]. Structured Document Image Analysis. [S. l.]: Springer-Verlag, 1992: 188 - 202.