第38卷 第21期
Vol.38 No.21

计 算 机 工 程
Computer Engineering

2012年11月
November 2012

·开发研究与设计技术· 文章编号：1000—3428(2012)21—0272—04 文献标识码：A 中图分类号：TP301.6

基于AdaBoost的多特征融合指纹检索方法

王富丽，欧阳建权

(湘潭大学信息工程学院，湖南 湘潭 411105)

摘 要：为提高视频内容检索方法的鲁棒性，提出一种基于AdaBoost的多特征融合指纹检索方法。通过对样本数据的训练，自适应地获得尺度不变特征变换特征、运动特征以及音频特征的权重，利用得到的权重融合音视频特征，以产生视频指纹。实验结果表明，该方法的准确性较高，在尺度变化、亮度变化、音频噪音攻击下具有较好的鲁棒性。

关键词：AdaBoost算法；视频指纹；运动特征；音频特征；动态贝叶斯网络

Fingerprint Retrieval Method of Multi-feature Fusion Based on AdaBoost

WANG Fu-li, OUYANG Jian-quan

(College of Information Engineering, Xiangtan University, Xiangtan 411105, China)

【Abstract】This paper proposes a fingerprint retrieval method of multi-feature fusion based on AdaBoost to improve the robust of video fingerprint. The proposed method can gain the weight of Scale Invariant Feature Transform(SIFT), temporal and audio feature adaptively by training the sample data, then fuse audio-video feature to produce video fingerprint according to the weights of the three features. Experimental results show that this method can gain higher accuracy, and have good robustness under various geometric, brightness modification and audio noise.

【Key words】AdaBoost algorithm; video fingerprint; motion feature; audio feature; Dynamic Bayesian Network(DBN)

DOI: 10.3969/j.issn.1000-3428.2012.21.072

1 概述

正文互联网的高速发展使得在线视频得到了前所未有的关注。据ComScore公司的统计，2011年10月份，将近12亿人观看了2 014亿个视频。YouTube是全球最大的视频服务商，约占视频观看量的40%。谷歌站点(包括YouTube)仅2011年10月就提供了883亿个视频服务，占视频总量的43.8%。同时，网络视频的盗版问题也变得越来越严峻，为此，催生了视频指纹技术。指纹是指从源视频中提取的少量的相关特征，其目标是提供鲁棒的方法来检索视频内容。实际上目前土豆、优酷、奇艺、酷6等视频网站已经开始利用视频指纹应对版权问题，但依然存在视频指纹的鲁棒性问题。

目前，可应用于视频检索的特征有图像的颜色、形状、纹理、局部特征、运动特征，以及音频特征等。单特征可以是视频的图像特征、运动特征或音频特征等，如文献[1]提出运用梯度矢量图心特征来检索拷贝视频。文献[2]采用音频特征来产生视频指纹，但单个特征由于其自身的特点而不能达到满意的效果。文献[3]证实一个单一的特征不足以检测有不同类型变化的拷贝视频片段，并提出融合面部镜头匹配、运动序列匹配和低水平特征的非面部镜头3种特征结合来检测拷贝视频，并获得了较好的效果和效率。但该方法是利用3种特征返回各自匹配最好的结果给检索视频，然后，联合这些候选结果与其他记录的匹配候选结果合并得到一个得分来实现。

文献[4]通过视频顺序特征和颜色特征相结合产生视频指纹，但该方法的查全率和查准确不高。文献[5]提出组合颜色布局描述、可伸缩颜色描述和边缘直方图描述3个特征进行视频检索，而该方法是将提出的3个特征整个作为一个指纹来实现。这些利用不同方法将多特征融合进行检索，都不能充分利用各个特征优势。

文献[6]提出基于动态贝叶斯模型(Dynamic Bayesian Network, DBN)融合音视频特征的说话人识别。文献[7]提出一种利用AdaBoost算法选择和融合多特征的方法。

针对单个特征自身的特点，本文提出一种基于AdaBoost的多特征融合指纹检索方法。采用SIFT特征、文献[8]提出的帧间差异法提取运动特征相结合来获得视频特征，利用文献[9]提出的方法来提取音频特征，并使用AdaBoost算法融合这3种特征来产生视频指纹。

基金项目：国家科技支撑计划基金资助项目“面向全网运营的数字卡通工程化技术研究与应用”(2007BAH14B05)

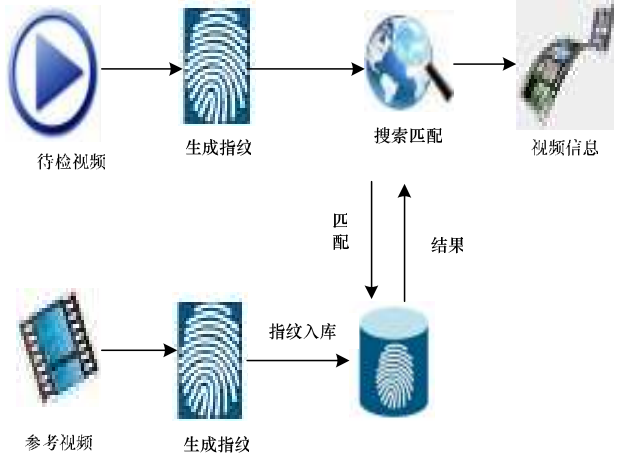
作者简介：王富丽(1986－)，男，硕士研究生，主研方向：多媒体处理；欧阳建权，教授

收稿日期：2012-01-31 **修回日期**：2012-03-10 **E-mail**：kissingman1@gmail.com

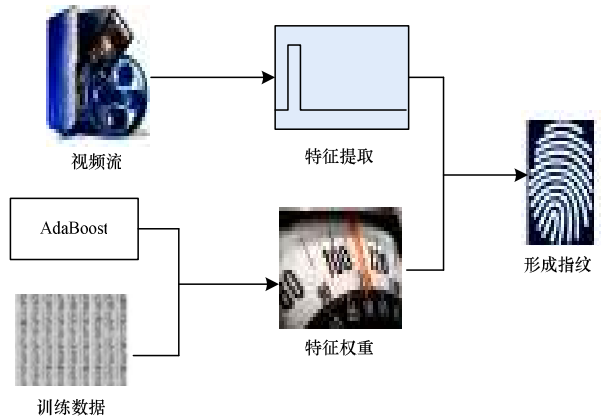
2　视频检索

2.1　基于内容的视频检索

为了利用各个特征的优势来实现鲁棒、准确的检索视频，本文从视频中提取出视频图像的尺度不变特征变换(Scale Invariant Feature Transform, SIFT)特征、基于帧间差异的运动特征及音频特征，并利用 AdaBoost 融合这些特征产生视频指纹进行视频检索。多模式视频指纹生成方法如图 1 所示。



(a)视频指纹生成方法



(b)产生指纹的过程

图 1　多模式视频指纹生成方法

2.2　特征提取

SIFT 特征是图像很重要的特征。本文首先将视频帧分成 $M\times N$ (本文取 M =3， N =4)块，分别求出每分块的 SIFT 特征点个数 $T(p,t)$，其中，$t=1,2,\cdots,12$，p 表示视频的第几帧，$p=1,2,\cdots,P$，P 为视频的总帧数，最后，每个视频的 SIFT 特征就是一个矩阵。

运动特征是视频区别于图像数据所特有的内容，能更好地标识视频。本文运用帧间差异法表示视频的运动特征，每 2 帧得到一个 12 位的二值行向量，从而得到的运动特征是一个维数和视频帧数量相关的矩阵 $\boldsymbol{T}$。

音频特征是从视频中解码出音频流，然后从音频流中提取特征，本文采用文献[9]提出的音频特征，每帧得到一个 32 位的二值行向量，最后音频特征是一个行数和音频帧数量相等的矩阵。

2.3　AdaBoost 算法

假定 X 表示样本空间，Y 表示样本类别标识集合，假设是二值分类问题，$Y=\{0,1\}$。令 $S=\{(x_i,y_i)|i=1,2,\cdots,m\}$ 为样本训练集，其中，$x_i\in X$，$y_i\in Y$，此处一个滑动窗口 $F_s(p_d,a)$、$F_m(p_m,a)$、$F_a(p_a,a)$ 和对应特征 $Q_s(P_q,a)$、$Q_m(P_m,a)$、$Q_a(P_a,a)$ 中的连续 p_d 行、p_m 行、p_a 行为一个样本，所以，一个待检视频 F 和一个被检视频 Q 总计有 P_a-p_a 个样本；在视频 Q 中，包含待检视频 F 处标记 Y 为 1，否则标记为 0。

基于 AdaBoost 的权值训练方法如下：

(1)初始化 3 个特征的权值为：$D_t(i,1)=1/3$，$D_t(i,2)=1/3$，$D_t(i,3)=1/3$，其中，$D_t(i)$ 表示在第 t 轮迭代中经过训练后更新的权值；1、2、3 分别对应于 SIFT 特征、运动特征及音频特征。

(2)令 *count* 表示迭代的次数，并且由特征个数决定，本文中 *count*=3。

(3)For t=1 to count do

用 3 种特征对原训练集 S 中的所有样本分类。

得到本轮的分类结果，并且有分类错误率：

$$\varepsilon_t=\Pr_{i-Di}\left[h_t(X_i)\neq y_i\right]$$

令：

$$\alpha_t=(1/2)\ln\left[(1-\varepsilon_t)\Big/\varepsilon_t\right]$$

分别更新每个特征的权值：

$$D_{i+1}(i)=\frac{D_t(i)}{Z_t}\times\begin{cases}\mathrm{e}^{-\alpha_i} & h_t(x_i)=y_i\\ \mathrm{e}^{\alpha_i} & h_t(x_i)\neq y_i\end{cases}$$

其中，Z_t 是一个正规因子，用来归一化确保 $\sum D_{t+1}(i)=1$。

end for

(4)得到 3 个特征的权重之后，利用时间轴对齐音视频特征，进而进行融合。

2.4　指纹匹配

在提取视频指纹后，3 种特征均采用滑动窗口匹配方法，SIFT 特征逐帧比较每个分块的特征点个数，对应块的个数之差小于 5，这样的块达到 8 块或者以上，则认为 2 帧是相似的；运动特征和音频特征逐帧计算城市距离，当小于阈值时，则认为 2 帧是相似的，再计算相似度，即相似帧数与总帧数之比。最后利用 AdaBoost 算法训练后得到 3 种特征的权重，求 3 种特征相似度的权值和作为总的视频相似度，如果总的相似度大于阈值 H (本文中 H = 0.6)则匹配成功。

设待检视频的 SIFT 特征为 $F_s(p_d,a)$，运动特征为 $F_m(p_m,a)$，音频特征为 $F_a(p_a,a)$；被检视频 SIFT 特征为 $Q_s(P_q,a)$，运动特征为 $Q_m(P_m,a)$，音频特征为 $Q_a(P_a,a)$；此处，$p_d<<P_q$，$p_m<<P_m$，$p_a<<P_a$。在匹配时把 $F_s(p_d,a)$、$F_m(p_m,a)$、$F_a(p_a,a)$ 均当作滑动窗口，与 $Q_s(P_q,a)$、$Q_m(P_m,a)$、$Q_a(P_a,a)$ 中的连续 p_d、p_m、p_a 行分别对应作比较，然后求的 SIFT 特征的相似度 θ_s、运动特征的相似

度 θ_m、音频特征的相似度 θ_a，其计算方法如下：

$$\theta_s = p_s / p_d$$

其中：

$$p_s = \sum_{i=1}^{p_d} \begin{cases} 1 & C_i < 5 \\ 0 & \text{otherwise} \end{cases}$$

$$C_i = \sum_{j=1}^{12} \begin{cases} 1 & |F_s(i,j) - Q_s(g+i,j)| < 5 \\ 0 & \text{otherwise} \end{cases}$$

$$\theta_m = p_c / p_m$$

其中：

$$p_c = \sum_{i=1}^{p_m} \begin{cases} 1 & \sum_{j=1}^{12} |F_m(i,j) - Q_m(g+i,j)| < 5 \\ 0 & \text{otherwise} \end{cases}$$

$$\theta_a = p_e / p_a$$

其中：

$$p_e = \sum_{i=1}^{p_a} \begin{cases} 1 & \sum_{j=1}^{32} |F_a(i,j) - Q_a(G+i,j)| < 13 \\ 0 & \text{otherwise} \end{cases}$$

其中，G 为 $Q_a(P_a,a)$ 在本次匹配前已经滑到的位置；g 为 $Q_s(P_q,a)$、$Q_m(P_m,a)$ 在本次匹配前已经滑到的位置；而 g 和 G 的关系为 $g = round(G \times 0.650\,9)$ 以保证音视频之间的同步。总的相似度 θ 的计算公式如下：

$$\theta = W_s \times \theta_s + W_m \times \theta_m + W_a \times \theta_a$$

其中，W_s、W_m、W_a 分别为经 AdaBoost 训练后得到的 SIFT 特征、运动特征、音频特征的权重。

如果 θ 大于阈值 H 则匹配成功；如果匹配成功则把 $Q_s(P_q,a)$、$Q_m(P_m,a)$、$Q_a(P_a,a)$ 分别向后滑动 p_d、p_m、p_a 行；如果不成功，则 $Q_a(P_a,a)$ 向后滑动一行，$Q_s(P_q,a)$、$Q_m(P_m,a)$ 分别滑动到 $g = round(G \times 0.650\,9)$ 位置。

3 实验结果及分析

本文的实验环境如下：硬件：Pentium(R) D 3.00 GHz CPU，1.00 GB RAM，软件：Windows XP，Matlab 7.1。

在视频指纹库中，包含从 4 s~36 s 之间不同长度的 53 段视频片段，视频图像大小为 352×288，其总长度为 7 min 37 s。待检视频来自 11 个不同电视频道的 28 段视频，总时长为 3 h 1 min 19 s。

视频检索常用的评估参数为查全率 R、查准率 P，以及 F 值 F，其计算方法给出如下：

$$R = \frac{\text{正确检测}}{\text{正确检测} + \text{丢失检测}}$$

$$P = \frac{\text{正确检测}}{\text{正确检测} + \text{错误检测}}$$

$$F = \frac{2RP}{R+P}$$

3.1 准确性分析

本文方法与 3 种单特征方法的查全率和查准率比较如图 2 所示。由图 2 可知，本文方法比单特征方法的查全率和查准率要好。由于音频特征是从音频中提取帧频带之间能量的差异，如果遇到音频中的能量不发生变化或变化很小时(如对白、旁白或安静的时候等“静音”视频)，音频特征就无能为力了，因此尽管音频特征检索的效果只是稍劣于多特征检索，但还是需要融合多特征。而 SIFT 特征是提取视频图像帧的特征、运动特征则是视频特有的运动信息，所以，这 3 种特征缺一不可。

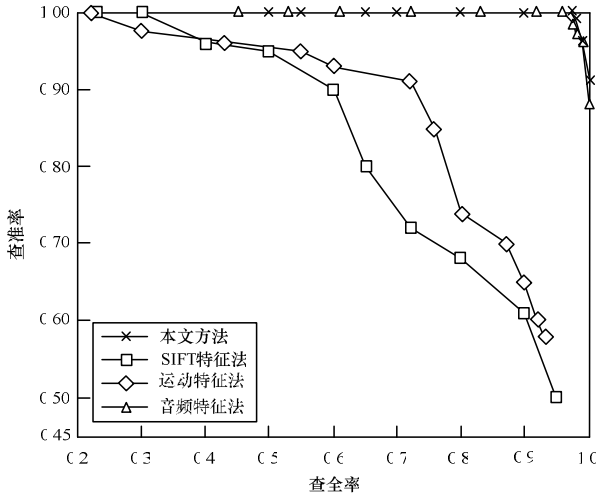


图 2 本文方法与 3 种单特征方法的查全率和查准率比较

文献[4]通过视频顺序特征和颜色特征相结合的方法进行视频检索，以此提高视频检索的准确度，但没有融合音频特征；文献[6]利用概率模型融合多特征以提高多特征之间的耦合度；本文利用 AdaBoost 算法获得 3 种特征的自适应权重，能更好地融合这些特征；本文方法与文献[4]方法、文献[6]方法比较如图 3 所示，实验表明，本文准确性要优于其余 2 种方法。

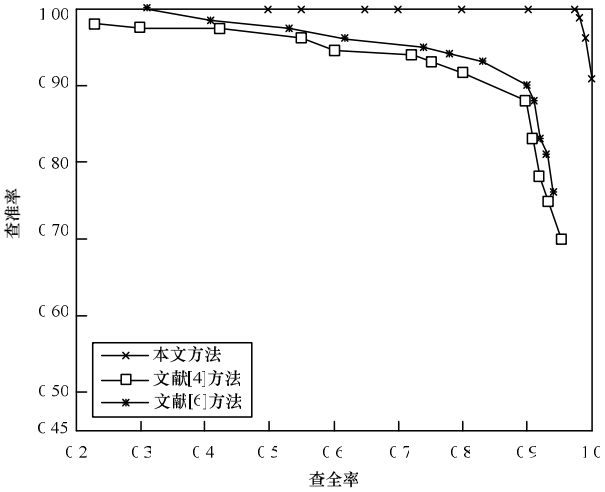


图 3 3 种方法的查全率和查准率比较

3.2 鲁棒性分析

视频文件在上传、传输、下载过程中，可能遭受各种各样的攻击，造成被检视频与原始参考视频存在差异，因此，需要对视频指纹进行鲁棒性研究。本文进行视频亮度(分别增加 30%($A1$ 表示 attack set 1，下同)、20%($A2$)、10%($A3$)，以及减少 10%($A4$)、20%($A5$)、30%($A6$))的攻击实验、尺寸攻击(攻击尺寸分别为 352×240($A7$)、320×240($A8$)、720×576($A9$))实验、均衡噪声($A10$)以及白噪声(白噪声强度分别为 2($A11$)和 3($A12$))实验。实验的具体结果

如表 1 所示。实验结果表明，本文方法对亮度、尺寸攻击鲁棒性非常好，对均衡噪声以及白噪声也比较好。与文献[4]方法、文献[6]方法相比，本文方法在各种攻击下都能获得更好的鲁棒性。

表 1　各种攻击实验的结果　(%)

类型	本文方法			文献[4]方法			文献[6]方法		
	R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
*A*1	100	96.6	98.2	91.3	83.0	86.9	93.9	87.5	90.6
*A*2	100	96.6	98.2	91.5	85.2	88.3	92.9	88.6	90.5
*A*3	100	96.6	98.2	93.8	86.4	89.9	95.2	89.8	92.4
*A*4	100	96.6	98.2	93.9	87.5	90.6	94.1	90.9	92.5
*A*5	100	96.6	98.2	91.5	85.2	88.3	95.2	89.8	92.4
*A*6	100	96.6	98.2	89.2	84.1	86.6	91.7	87.5	89.5
*A*7	100	96.6	98.2	87.7	80.7	84.0	90.4	85.3	87.7
*A*8	100	96.6	98.2	90.2	84.1	87.0	91.6	86.4	88.9
*A*9	100	96.6	98.2	88.8	80.7	84.5	90.2	84.1	87.1
*A*10	100	94.3	97.0	86.5	80.3	83.3	95.2	90.9	93.0
*A*11	100	93.2	96.4	95.1	87.5	90.9	91.4	84.1	87.6
*A*12	100	87.4	93.3	95.1	87.5	90.9	90.0	81.8	85.7

当对视频进行亮度攻击时，SIFT 特征自身对亮度有很强的适应性，而本文只是计算视频帧每个分块的 SIFT 特征点个数，因此，SIFT 特征基本是不变的；运动特征提取的是相邻帧对应分块的像素值差，相邻帧的像素值差是不变的，所以，提取的运动特征也是不变的。对于尺寸攻击，提取 SIFT 特征和运动特征都是先将视频帧分成 $M \times N$ 块，所以，虽然改变了视频尺寸，在每一分块提取的 SIFT 特征点的数量是不变的，同样，相邻帧间对应分块的像素值也是一致的，因此，提取的 SIFT 特征和运动特征都是不变的。而对于均衡噪声和白噪声的攻击鲁棒性则有待进一步提高。

4　结束语

本文提出一种基于 AdaBoost 的多特征融合指纹检索方法。采用图像的 SIFT 特征、帧间差异的运动特征，以及音频特征相结合的方法进行视频检索。利用 AdaBoost 算法训练得到 3 种特征的自适应权重，以更好地融合 3 种特征。实验结果表明，该方法产生的指纹对视频亮度变化、尺寸攻击、均衡噪音，以及音频噪音等有较强的鲁棒性，能达到准确检索视频的要求。下一步的工作是将该视频指纹方法应用到大型视频库，并实现高效的检索。

参考文献

[1] Sunil L, Yoo C D. Video Fingerprinting Based on Centroids of Gradient Orientations[C]//Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing. [S. l.]: IEEE Press, 2006.

[2] Haitsma J, Kalker T, Oostveen J. Robust Audio Hashing for Content Identification[EB/OL]. (2010-11-21). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.16.2893.

[3] Kucuktunc O, Muhammet B, Gudukbay U. Video Copy Detection Using Multiple Visual Cues and MPEG-7 Descriptors[J]. Journal of Visual Communication and Image Representation, 2010, 21(8): 838-849.

[4] Yuan Junsong, Duan Lingyu, Tian Qi. Fast Video Segment Identification from Large Video Collection[EB/OL]. (2010-08-21). http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.106.499.

[5] Bertini M, Bimbo A D, Nunziati W. Video Clip Matching Using MPEG-7 Descriptors and Edit Distance[EB/OL]. (2010-09-07). http://researchr.org/publication/BertiniBN06%3A0/bibtex.

[6] Wu Zhiyong, Cai Lianhong, Meng Helen. Multi-level Fusion of Audio and Visual Features for Speaker Identification[C]//Proc. of International Conference on Advances in Biometrics. Berlin, Germany: Springer-Verlag, 2005.

[7] Gao Changxin, Sang Nong, Tang Qiling. On Selection and Combination of Weak Learners in AdaBoost[J]. Pattern Recognition Letters, 2010, 31(9): 991-1001.

[8] Oostveen J, Kalker T, Haitsma J. Feature Extraction and a Database Strategy for Video Fingerprinting[C]//Proc. of the 5th International Conference on Recent Advances in Visual Information Systems. London, UK: Springer-Verlag, 2002.

[9] 聂　华. 基于音频指纹的广告检测技术研究[D]. 湘潭: 湘潭大学, 2011.

编辑　刘　冰