一种基于多特征的视频人物聚类方法*)

蒋 鹏 秦小麟

(南京航天航空大学计算机科学与技术系 南京 210016)

摘 要 检索一段视频中出现的人物并进行人物归类具有重要的研究意义和实用价值。本文提出一种基于多特征的视频人物检索聚类算法:先用一种结合人脸检测和物体跟踪的算法检测镜头人物,并提取人物衣服区域颜色以及声音作为人物特征,再用一种无监督模糊聚类方法对人物进行聚类,最后利用声音特征对聚类结果进行修正。该方法适用于人物数未知的条件下进行无监督的人物聚类。不同类型视频的试验证明该方法有效而实用。

关键词 人物检索,无监督聚类,人脸检测

Automated Person Indexing in Video

JIANG Peng QIN Xiao-lin

(Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China)

Abstract Person retrieval and indexing in video sequences are challenging task for many multimedia applications. This paper proposes a new method that indexes the person by their similarity. Firstly, the persons in a shot are detected and tracked through face detector and continuously adaptive mean shift algorithm. Then mid-level features such as clothes colors and voice are applied to represent the person. An unsupervised cluster method is performed to cluster the person for further index. At last, the cluster is validated and refined by the use of voice feature. Conducted on different kinds of video, the method has been found to be effective.

Keywords Person retrieve, Unsupervised clustering algorithm, Face detection

1 引言

随着数字电视和多媒体技术的快速发展,如何对海量的视频数据进行有效的索引是目前研究的一个热点问题。早期大量的研究工作集中于利用底层特征(颜色、纹理、运动等)进行视频结构分析。然而人们更倾向于利用高级语义级别的检索方式进行视频检索,例如电影观众趋向于查找某电影中女主角出现的所有镜头,而不是查询包含某种颜色的所有图像帧。如何提取视频中的语义级别特征并有效地进行视频检索是目前研究的一个重要方向。

目前大量的电影、电视剧、谈话等视频都是以人物为重点,观众对该类节目中出现的人物比较关心。如何将视频中的人物检索出来并进行归类具有重要的使用价值。现在已经有较多的工作在视频中进行人物检索并识别,利用视频中的视觉特征和声音特征进行人物识别是常用的方法。在视觉特征方面,将视频看成不相关的一系列图像帧,并将基于图像的人脸识别方法应用到视频帧中进行人脸识别是一种基于图像的方法。一种改进的做法是首先筛选视频中人脸大小合适、光照度良好、较好的帧,再利用在静态图像中常用的主成分份分析(PCA)方法对视频帧进行人脸识别[1]。文献[2]利用嘴巴、鼻子等脸部特征的空间位置关系特征作为人脸描述符进行人物识别。但利用人脸进行人物识别的方法受到光照度、人脸面积和人脸角度变换等因素影响较大,当视频中光照度、人脸角度等变化较大时,直接利用人脸图像进行人物识别的方法用于视频中效果不太理想。借助衣着、头发等非人脸视觉特征进行人物识别是一个重要的人物识别方法,文献[3]提出利用人物头发特征对人物进行识别,该方法提取头发的颜色、发型、长度等特征作为人物特征,从头发特征的区别进行人物识别,但该方法识别率较低。利用说话人识别技术进行人物识别是从人物的语音信号中提取个人特征,从而区分说话人。但利用声音进行人物识别容易受到背景噪声干扰并且无法对于沉默等镜头进行人物识别,具有一定的限制性。结合视觉特征和声音特征对人物进行识别是目前的一个重要趋势。文献[4]提出的人物索引算法首先利用人脸检测技术将视频中具有人物的镜头提取出来,并联合声音和衣服颜色作为底层特征,在距离空间用支撑向量机(SVM)概率输出将人物底层特征的距离映射为语义层人物相似度,最后用无监督聚类算法对于SVM输出的人物相似度进行修正。但当衣服颜色或者声音特征有其一无法提取时,该方法无法正确获取人物相似度,从而导致人物检索错误。文献[5]结合人脸识别和声音特征进行人物检索,该方法利用人脸和声音特征分别进行人物相似性判定,并综合两个特征判定的结果作为最后结果,并证明该方法优于先将人脸特征和声音特征进行混合,再利用混合特征进行人物相似性判定的方法。

考虑到在一段视频中,人物的衣着和声音是不同人物进行区别的重要特征,本文首先采用一种结合人脸检测和物体跟踪的算法检测镜头人物,并提取人物衣服颜色以及人物声音作为特征,再使用一种无监督模糊聚类方法对人物进行聚类,最后利用声音特征对聚类结果进行修正,以便在某些特征无法提取时,依然能够正确地进行人物聚类。图1显示了这

*)国家自然科学基金(60673127)资助。蒋 鹏 博士研究生,研究方向为视频检索和多媒体数据库检索;秦小麟 教授,博士生导师,主要研究领域为安全数据库、时空数据库等。

种结构。

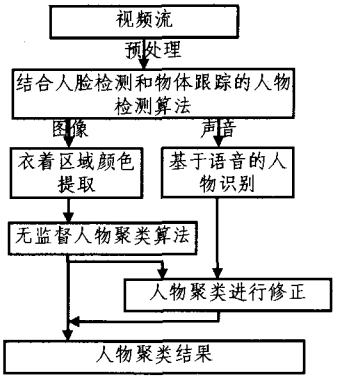


图 1　视频人物检测以及聚类

2　人物检测以及特征提取

2.1　预处理

本文首先利用文献 [6] 的运动查询窗口算法将视频流分割成子镜头级别，每个子镜头内容较为相似，该算法对于镜头切变和渐变的检测都有较高的准确率。由于人脸检测计算量较大，对于每一帧检测人脸效率较低，首先用 K 均值聚类将相似的颜色特征的视频帧进行聚类，并提取利聚类中心最近的帧作为关键帧。试验中根据镜头长度动态设置聚类数目 $k=N$，其中 N 为镜头长度，单位为秒，然后用文献 [7] 中的人脸检测算法在关键帧集合中检测人脸，当大多少帧能检测到人脸时，认为该镜头为包含人物的镜头。

2.2　镜头人物检测

将视频看成一系列图像帧，并检测图像帧中人脸来判断镜头人物是目前较为常用的方法[1,2,4]。但该类方法对于视频的连续性以及上下文关联性考虑得较少。当人脸检测算法在复杂背景等干扰下，可能会发生误检或者漏检，从而直接影响到人物特征提取。考虑到在大部分以人物为主体内容的镜头中，人物的运动应该是连续的和稳定的，当人脸出现时间很短或者快速改变位置，则该人物可能是人脸误检，但是人脸检测计算量较大，对于镜头中每帧图像，检测人脸并比较位置效率较低。本文用一种结合人脸检测和物体跟踪技术来检测镜头人物。首先在关键帧中检测人脸，再利用 CamShift(Continuously Adaptive Mean Shift) 算法进行人物跟踪。CamShift 是一种基于颜色信息的跟踪算法，在跟踪过程中，CamShift 利用目标的颜色直方图模型得到每帧图像的颜色投影图，并根据上一帧跟踪的结果自适应调整搜索窗口的位置和大小，从而得到当前图像中目标的尺寸和位置。首先采用 2.2.1 节提取到的衣服颜色直方图进行颜色跟踪，并根据跟踪窗口的中心点的移动以及跟踪窗口大小来判定人物的稳定性。当跟踪窗口中心点位置或者窗口面积的变化大于一个阈值时，认为跟踪失败，如果人脸帧前后 30 帧内跟踪成功，则成功地检测到镜头人物。

2.2.1　人物衣着颜色特征

文献 [4]也提出利用人物衣着区域颜色信息进行人物识别的算法，该算法首先方法定位人脸位置，并取人脸位置以下固定矩形区域作为衣服区域，提取该区域的颜色等直方图作为特征。考虑到现实情况，人脸区域以下部分包含颈部、手臂等区域，利用人脸以下固定位置作为衣服区域会导致将背景或者手臂等作为衣服区域，从而导致误检。为了减少背景以及颈部、手臂区域对衣着颜色的干扰，本文定义了 3 个区域，分别是人脸区域、衣着区域以及辅助衣着区域，如图 2 所示。设人脸区域位置为 (x,y,w,h)，x,y 为人脸区域坐标，w 和 h 分别为人脸区域的宽度和长度，3 个区域的结构如图 2(b)所示，其中衣着区域计算公式为：

$x1=x-w$；$y1=y+h$；$w1=3*w$；$h1=h$；

辅助衣着区域计算公式为：

$x2=x-1.5*w$；$y2=y+h$；$w2=4*w$；$h2=h$。

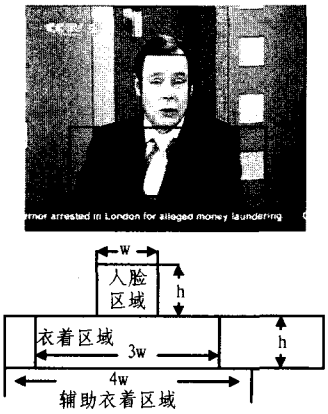


图 2　人脸以及衣着区域

衣着区域是辅助衣着区域的一个子集，目的是将背景等干扰去除。进一步获取 3 个区域的颜色特征，首先分别提取 3 个区域的颜色直方图，将 YUV 颜色空间转换成 HUE 颜色空间，并量化到 180 级(bin)。设人脸区域颜色直方图为 H_1，辅助区域直方图为 H_2，人物衣着区域颜色直方图为 H_3，则衣着区域直方图计算公式为：

$$H'_3(I)=2H_3(I)-H_2(I)-H_1(\max b(I)) \quad (1)$$

其中 $H_1(\max b(I))$为 H_1 中具有最大值的 bin，式(1)首先从衣着区域将背景去除并减去以肤色为主的非衣着区域颜色，$H'_3(I)$为将颈部、手臂和背景等干扰除去的衣服区域颜色直方图，图 3 为图 2 衣着区域提取结果。

图 3　衣着提取结果

为了使颜色直方图不随尺度变化，对直方图 $H(i)$进行归一化处理：

$$H_m=\frac{H(i)}{\sum_m H(i)} \quad (2)$$

人物衣着不变的情况下，可以利用上述方法获得的衣着颜色直方图的相似度来描述人物的相似度。则人物衣着相似度可以由直方图之间的欧式距离表示，表示为 D_H。

2.2.2　声音特征

人物的声音是进行人物区分的另一个重要特征，本文采用文献[8]提出的基于线谱对系数(line spectrum pair，LSP)相关性分析进行镜头人物语音特征提取，首先将镜头人物声音划分成窗长为 3 秒的采样单位，提取 10 级线性预测导频系数(LPC)并转换成 LSP 作为声音数据的特征表示，使用混合高斯模型 GMM 对声音特征进行建模，开始当没有足够的说

话者声音数据时，利用 GMM-1 模型，随着说话者声音数据的增加，GMM 中的概率分布增加到最多为 32 个高斯分量的高斯混合密度。人物声音可以由下列公式表示：

$$V_m=\{\mu_i^m,\sum_i^m\}i=1,\cdots,32 \tag{3}$$

其中 μ_i^m 是混合高斯模型的均值向量，$\sum_i^m$ 为混合高斯模型的协方差矩阵，人物声音的距离采用 K-L 离散度作为距离度量的方法：

$$D_{kl}(i,j)=\frac{tr[(\Sigma_i-\Sigma_j)(\Sigma_i^{-1}-\Sigma_j^{-1})]+tr[(\Sigma_i^{-1}+\Sigma_j^{-1})(\mu_i-\mu_j)(\mu_i-\mu_j)^T]}{2} \tag{4}$$

具体做法参见文献 [8]。

3 人物聚类及其修正算法

当提取到人物衣着颜色特征和声音特征后，一种做法是将两类特征联合成一个特征向量，并对该特征向量进行聚类。但该方法具有两个缺点：(1)利用联合特征向量进行聚类时一般使用相同的距离度量方式例如欧氏距离，但不是所有的特征都适合于同一种距离度量方式。(2) 该方法对于特征的部分缺失比较敏感，例如当出现某些不说话的镜头时，无法提取人物的声音特征，或者人脸太大等因素无法提取到衣着特征时，难以直接将多特征进行混合。本文提出的人物聚类算法首先利用人物衣着颜色特征进行无监督聚类，再在人物基本聚类的条件下利用人物的声音特征对聚类结果进行修正。在人物聚类时，用户常常无法事先知道视频内容，因此视频中出现的人物类别个数往往是未知的，现有的 K 均值和模糊 C 均值(FCM)等聚类算法无法直接使用。由于人物衣着颜色受光照度等多种因素影响，具有一定的不确定性，利用基于隶属度的模糊聚类方式比较适合。本文以衣着区域颜色直方图为特征，提出一种改进的基于 FCM 算法进行人物聚类，首先设置一个较大的聚类数，聚类过程中动态合并相似的类别，最后获得最佳聚类数同时完成聚类。在进行人物聚类前，首先定义两个聚类中心的分散度为：

$$SF_{ij}=\frac{Col_i+Col_j}{Dist_{ij}} \tag{5}$$

其中 Col 为一个类的聚集度，$Dist$ 为两个类之间的距离，其中 $Col(i)=\frac{1}{N(i)}\sum_{x_k\in V_i}|x_k-V_i|$，$Dist(i,j)=|V_i-V_j|$，$N(i)$ 为聚类于 V_i 的人物个数。当 $SF(i,j)<1$，两个聚类分散度较高，即人物的相似性较小，不需要合并。当 $SF(i,j)\geq 1$，将类 i 和类 j 合并成新类 k，合并方法如下：

$$V_k=(V_i+V_j)/2$$
$$u_k=(u_i+u_j)$$
$$c'=c-1$$

V_k 为合并后类的聚类中心，u_k 为新类的隶属度，c' 为合并后类总数。

人物聚类算法如下：

步骤 1：设置一个较大的聚类数初始值 c 并用随机数初始化隶属矩阵 U 以及聚类中心。

步骤 2：利用 FCM 更新规则更新 c 个聚类中心 V 和隶属矩阵 U。

步骤 3：计算 FCM 价值函数 J。如果它大于某个确定的阈值，或它相对上次价值函数值的改变量大于某个阈值，则转至步骤 2。

步骤 4：计算各个类间分散度，并合并相似的类。如果 c 值减小，转到步骤 2，如果 c 值稳定则退出算法。

FCM 更新规则和价值函数 J 参见文献 [9]。经过上述聚类，具有相似衣服颜色的人物被归到同一类别，当有多个人穿相似颜色的衣服或者同一人物更换衣着时，利用上述算法会产生误判，利用人物的声音特征则可以较好地对误判进行修正。在人物基本聚类的条件下，根据人物声音特征 KL 距离和衣着颜色距离，人物之间的相似度利用如下距离形式表示：

$$K_{ij}=w_1D_h+w_2D_{kl} \tag{6}$$

其中 w_1，w_2 分别为 D_h，D_{kl} 权重值，D_h，D_{kl} 均已归一化。通过比较试验，本文分别设为 $w_1=0.4$，$w_2=0.6$，并利用如下两个规则进行聚类修正：

RULE1：
DO
提取聚类在某类别中任意两个样本
IF $K_{ij}>\tau_1$
THEN 分裂该聚类
UNTIL(所有样本均已比较)
RULE2：
DO
各提取两个不同聚类中任一样本
IF $K_{ij}<\tau_2$
THEN 合并两个聚类
UNTIL（聚类无法继续合并）

其中 τ_1，τ_2 为预先设置阈值。

4 实验及其结果

为了验证本文方法的有效性，分别对故事片、新闻、访谈节目等常见类型视频进行了试验。试验采用真实的电影或者电视录制数据，首先利用预处理方法将视频分割成子镜头级别，再对人物镜头进行手工标注。对镜头人物聚类采用聚类准确度表示算法性能：

准确度＝总共正确聚类的人物个数/总共出现人物个数

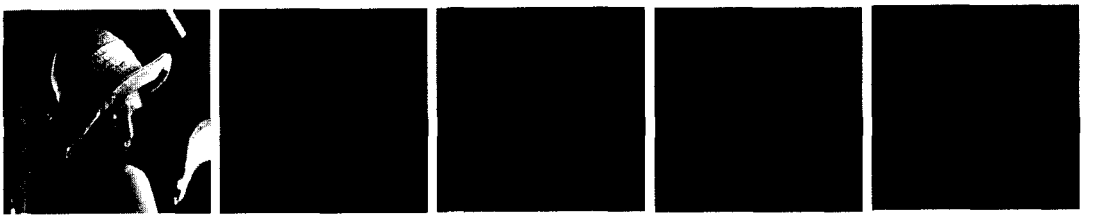
表 1 为试验数据以及试验结果。

表 1 人物聚类试验结果

视频片段	类型	子镜头数目	实际人物类别数	计算获得人物类别数	修正前准确度	修正后准确度
电影《云中漫步》片段	故事片	124	7	5	62.1%	74.2%
CCTV-9 新闻片段	新闻	252	12	11	75.4%	83.3%
CCTV-9 访谈节目片段	谈话	281	6	6	89.3%	91.8%
	平均				75.6%	83.1%

由表 1 可知，经过语音修正后，人物聚类的准确率平均从 75.6%提高到 83.1%，说明语音修正在人物聚类中起到良好效果。新闻和访谈节目效果较好，主要原因是新闻和访谈节目中人物镜头大多是正面或者侧面出现，光照稳定并且背景音较少，便于提取人物的衣着颜色和人物声音特征，因而聚类

（下转第 245 页）



(a) Lena 原始图像　(b) 本文算法　(c) Roberts 算子　(d) Sobel 算子　(e) LoG 算子

图 1　原始图像及边缘检测结果

(a) 噪声图像　(b) 本文算法　(c) 文献[8]　(d) LoG 算子　(e) Canny 算子

图 2　噪声图像的边缘检测结果

参 考 文 献

[1] Gonzalez R C, Woods R E. Digital Image Processing Second Edition[M]. Beijing: Publishing House of Electronics Industry, 2006

[2] Wechsler W. Texture Analysis—a Survey[J]. Signal Processing, 1980(2): 271-280

[3] Furht B, et al. Video and Image Processing in Multimedia Systems[M]. Boston: Kluwer Academic Publishers, 1995: 226-270

[4] Zhou F, Shi J Q. Texture Feature Based on Local Fourier Transform[J]// IEEE International Conference on Image Processing. 2001, 17(2): 610-613

[5] Mallat S, Zhong S. Characterization of Signals from Multi-scale Edges[J]. IEEE Transactions on PAMI, 1992, 14(9):710-732

[6] Tamura H, Mori S, Yamawaki T. Texture Features Corresponding to Visual Perception[J]. IEEE-SMC, 8(6): 460-473

[7] Haralick R M. Statistical and Structural Approaches to Texture [J]. Proceedings of IEEE, 1979, 67(5):786-804

[8] 尚晋，施成湘. 小波多尺度模糊竞争边缘检测[J]. 计算机科学，2005，32(7)：182-184

[9] 王光勇，汪林林，王佐成. 基于纹理分析的边缘检测算法[J]. 计算机科学，2007，34(9)：227-229

(上接第 242 页)

效果较好。该类主要的错误来至于基于声音的聚类修正过程，将不同的人物归类到同一个人。在电影《云中漫步》片段中，人物聚类的错误主要来至于两个方面：(1)人脸区域过大或衣着区域被阻挡时导致衣着区域无法正确提取；(2) 镜头中出现的人物和当时的说话人并不一致情况。虽然语音修正方法提高在该试验中总体提高了准确度，但在镜头人物和说话人不一致时会将人物修正到说话人的类别从而导致聚类错误。

结束语　本文提出一种基于多特征的视频人物索引算法，该算法能够自动检索一段视频中出现的人物并进行分类，由于人物的相似度是基于人的衣着的颜色和声音，因此与人物表情、动作和背景等无关，具有较强的鲁棒性和实用性。试验结果表明，该方法实用而有效，为现有的视频检索方法提供了一种有效的补充，尤其适用于电视剧、谈话节目等视频节目。而对于镜头中存在配音情况或者人物的图像和声音不匹配时，利用本方法无法正确聚类，需要改进；如何将利用人物聚类结果提取更高级语义信息也是今后要深入的问题。

参 考 文 献

[1] Berrani S-A, et al. Enhancing Face Recognition from Video Sequences using Robust Statistics[C] //IEEE International Conference on Video and Signal-Based Surveillance. Italy, 2005:324-329

[2] Sivic J, et al. Person spotting: video shot retrieval for face sets [C]// The 4th International Conference on Image and Video Retrieval (CIVR2005). Singapore, 2005:226-236

[3] Yacoob Y, et al. Detection and Analysis of Hair[J]. IEEE Transaction on Pattern analysis and machine intelligence, 2006, 28(7):1164-1169

[4] 王鹏，马宇飞，等. 一种利用人物相似度的视频索引算法. 电子学报，2004，32(6)：968-972

[5] Albiol A, Torres L, et al. Fully automatic face recognition system using a combined audio-visual approach[J]. Vision, Image and Signal Processing //IEE Proceedings, 2005(6): 318-326

[6] Volkmer T, Tahaghoghi, et al. The moving query window for shot boundary detection at TREC-12[C]// Proceedings of the TRECVID 2003 Workshop. Gaithersburg, Maryland, USA, 2003:147-156

[7] Xiao R, et al. Robust Multi-Pose Face Detection in Images[J] IEEE Transactions on Circuits and Systems for Video Technology, 2004, 14(1):31-41

[8] Lu Lie, Zhang Hong-Jiang. Real-time Unsupervised Speaker Change Detection. [C]// 16th International Conference on Pattern Recognition. Quebec, 2002: 358-361

[9] Bezdek J C. Pattern Recognition with Fuzzy Objective Function Algorithms [M]. New York: Plenum Press, 1981

[10] Castrill'on-Santana M, et al. Real-time Detection of Faces in Video Streams[C]// The Second Canadian Conference on Computer and Robot Vision (CRV'05). Spain, 2005:298-305