ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.20, No.9, September 2009, pp.2417−2425
doi: 10.3724/SP.J.1001.2009.03461

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

基于人脸检测与 SIFT 的播音员镜头检测*

杨武夷+, 曾 智, 张树武, 李和平

(中国科学院 自动化研究所,北京 100190)

Anchor Shot Detection Based on Face Detection and SIFT

YANG Wu-Yi+, ZENG Zhi, ZHANG Shu-Wu, LI He-Ping

(Institute of Automation, The Chinese Academy of Sciences, Beijing 100190, China)

+ Corresponding author: E-mail: eagleywy@126.com

Yany WY, Zeng Z, Zhang SW, Li HP. Anchor shot detection based on face detection and SIFT. *Journal of Software*, 2009,20(9):2417−2425. http://www.jos.org.cn/1000-9825/3461.htm

Abstract: Anchor shot detection is a fundamental step for segmenting news video into stories. In this paper, a algorithm is developed for anchor shot detection based on face detection and SIFT. Face detection is first used to filter out the shots which do not have any face in the special area. Then, color histogram is used to judge whether two shots are similar, and the distinctive image features from scale-invariant key points are detected and matched to find groups of shots which may be anchor shots. Last, anchor shots are identified based on the same prior information of anchor shots. Compared with other algorithms, the significant advantage of the proposed algorithm is that the algorithm is based on neither the templates nor the learned classifiers. The method has been tested on many kinds of news video, which demonstrates its effectiveness.

Key words: face detection; color histogram; SIFT feature points; anchor shot detection; video analysis

摘 要: 播音员镜头的检测是新闻视频结构化的关键步骤之一.提出了一种基于人脸检测与 SIFT 特征点匹配的播音员镜头自动检测算法.该方法首先利用人脸检测器过滤出具有人脸的候选镜头,然后利用颜色直方图判断镜头是否可能相似,再利用 SIFT 特征点匹配从候选镜头关键帧中找出相关的镜头组,最后利用各镜头组的信息判断出哪些是播音员镜头.对比传统的方法,该方法除了训练一个通用的人脸检测器外,不需要模板,也不需要针对某类新闻节目训练特别的分类器,可以直接利用算法对新类型的新闻节目提取播音员镜头.实验结果表明,该算法能够广泛地适应于各种不同种类的新闻节目、不同视觉质量的视频,可以有效地应用于新闻视频分析.

关键词: 人脸检测;颜色直方图;SIFT 特征点;播音员镜头检测;视频分析
中图法分类号: TP391 **文献标识码**: A

计算机及网络技术的迅猛发展,对多媒体资源的制作、管理和传播提出了极大的挑战.目前,每天都有大量的视频资源产生,如何有效管理这些视频资源成为一个非常活跃的研究领域.视频结构化就是一种有效手段,它通过对视频流进行分析处理,将视频组织成具有不同层次的特定信息,如关键帧、镜头、场景、片段和节目等.

* Supported by the National Sciences & Technology Supporting Program of China under Grant Nos.2006BAH02A13, 2006BAH02A03 (国家科技支撑计划)

Received 2007-12-12; Accepted 2008-07-02

这种层次化结构为视频资源的有效管理提供了有利条件.

新闻视频的结构特征比较明显,它是由一系列新闻条目构成.如何准确地定位每个新闻条目的起始时间点即新闻条目分割,是实现新闻自动编目系统的一项重要研究内容.新闻节目的结构在时间序列上较为固定,通常,播音员镜头和新闻内容镜头交替出现,由于新闻内容丰富多样,难以识别新闻内容镜头信息;而播音员镜头内容变化不大,运动较小,且播音员镜头的开始通常就是一个新闻条目的开始,能够作为新闻条目分割的边界.因此,对播音员镜头的检测成为新闻条目分割的重要手段之一.

目前,已经存在的播音员镜头检测方法主要分为基于模板的方法[1−4]和基于聚类的方法[5,6]两种.文献[1]利用关键帧时间和空间结构的先验知识建立播音员镜头模板,把候选播音员镜头和模板进行匹配,然后根据相似度度量决定其是否为播音员镜头,最后根据整段新闻的时间信息确定真正的播音员镜头.文献[2]提出了一种基于知识的二阶段模板匹配法用于新闻节目播音员镜头的检测.文献[3,4]首先利用非监督学习的方法提取出播音员镜头的音视频模板,然后利用模板判断候选镜头其是否为播音员镜头.文献[5]利用颜色特征对镜头聚类,得到候选播音员镜头类,然后利用人脸检测,滤除不存在大尺度人脸的候选镜头类,最后根据镜头时间序列规则得到播音员镜头.文献[6]利用镜头的颜色特征基于自动聚类方法找出候选播音员镜头,然后根据播音员镜头出现的时空特征,用神经网络分类器对候选播音员镜头进行确认.

上述的两类方法都存在一定的不足:基于模板检测播音员镜头的方法缺乏通用性,不同种类的新闻节目需要生成不同的模板,且当播音员在画面中的位置、大小或播音室环境发生变化时,该模板就可能失去作用;基于聚类的算法对于某些在一个新闻节目中出现次数较少的播音员镜头类会出现漏检,而某些非播音员镜头重复出现,其时空特征与播音员镜头相似,造成误检测[6].同时,基于聚类的算法还需要同一种新闻节目的长时间视频作为训练集.例如,文献[6]中的方法需要一种新闻的3个小时的节目作为训练样本.针对上述问题,本文提出了一种基于人脸检测、视觉特征度量与 SIFT 特征点匹配的播音员镜头自动检测算法.该算法不需要模板,也不需要针对某类新闻节目训练特别的分类器,算法能够广泛适应于各种不同种类、不同压缩质量的新闻节目.

本文第 1 节详细描述算法的基本原理和处理流程.第 2 节利用实验验证算法的可靠性.第 3 节为结论.

1 播音员镜头自动检测算法

1.1 基本原理

通过观察不同电视台的不同类型的新闻节目可以发现,新闻视频中播音员镜头具有一些普遍的规律:

(1) 新闻节目有一位或两位播音员,具有同一个播音员的镜头在新闻中会出现多次,且具有同一个播音员的第 1 个镜头和最后一个镜头之间的时间间隔比较大;

(2) 播音员一般都是正面朝向观众,播音员的上半身都在镜头中,不同镜头中播音员的上半身一般只存在一些小的差别;

(3) 在一个新闻节目的播报过程中,播音员的衣着是不变的,但背景可能有较大的变化.

根据上述规律进行播音员镜头检测,首先进行镜头关键帧提取,然后利用人脸检测对提取的镜头关键帧进行过滤,去除检测不到人脸的镜头关键帧,同时,记录镜头关键帧中人脸的个数和区域.对于能检测到多于一个人脸的镜头关键帧,通过人脸之间的空间关系判断是否可能为包含两个播音员的镜头关键帧.对镜头关键帧的某些特定区域提取视觉特征后,利用 SIFT 特征点检测算法在关键帧中通过人脸位置计算得到的播音员所在的大致区域中检测 SIFT 特征点,以某些镜头关键帧为基准,基于视觉特征和 SIFT 特征点匹配,找出能匹配到足够 SIFT 特征点的一组组关键帧作为候选的播音员镜头关键帧组.在进行 SIFT 特征点匹配前,利用基于颜色直方图的图像相似度度量方法判断镜头关键帧是否可能相似,减少 SIFT 特征点检测和匹配的计算量.并且,利用包含同一个播音员的第 1 个镜头和最后一个镜头之间的时间间隔比较大的规律,如果一组关键帧在视频中的时间跨度大于某个阈值,就认为它们是播音员镜头的候选关键帧组;否则,认为它们不是播音员镜头关键帧.最后,综合候选的只包含单个播音员的镜头关键帧组和包含两个播音员的镜头关键帧组的信息,判断哪些最可能为播音员镜头.

图 1 为整个播音员镜头检测算法的流程.下面将对流程中的各个步骤进行详细叙述.

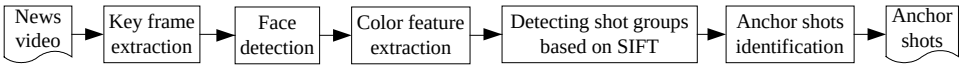

Fig.1 Processes of anchor shot detection

图 1 播音员镜头自动检测流程

1.2 镜头关键帧提取

在进行播音员镜头检测之前,先对视频数据进行一次处理,筛选出可能包含播音员镜头的候选关键帧.然后,在这些候选关键帧的基础上确定最终的播音员镜头关键帧.这样可以减少后续参加播音员镜头检测计算帧的数量,大大降低计算量.

我们对视频数据中的 I 帧进行处理来筛选关键帧.可以观察到,当视频从一般的视频帧(比如新闻内容)切换到播音员镜头关键帧时,存在一些区域的颜色直方图特征应该发生明显的变化.这样,可以通过观察这个关键区域的变化来检测可能的播音员镜头关键帧的出现.

我们取一个 3 帧长的时间窗口在 I 帧序列中滑动,在一个窗口内,利用颜色直方图求交的方法分别计算前后两帧与中间帧关键区域的相似度 $sim(n,n+1)$和 $sim(n+1,n+2)$,根据预先设定的相似度阈值 T_c 来判断前后两帧是否与中间帧相似,进而判断第 $n+1$ 个 I 帧是否应选为候选关键帧.如果 $sim(n,n+1)<T_c$ 并且 $sim(n+1,n+2)>T_c$,即第 n 个 I 帧和第 $n+1$ 个 I 帧不相似而第 $n+1$ 个 I 帧和第 $n+2$ 个 I 帧相似,那么将第 $n+1$ 个 I 帧选为候选关键帧,否则不选第 $n+1$ 个 I 帧为候选关键帧.然后,将窗口向下滑动一帧继续计算.这样,当窗口在整个 I 帧序列中滑过一遍时,就得到了可能包含播音员镜头的候选关键帧序列,并且一次播音员出场只会有一个关键帧.

1.3 镜头关键帧人脸检测

播音员镜头中的播音员一般都是正面朝向观众,所以可以通过检测镜头关键帧中是否存在人脸来过滤出可能存在播音员的镜头关键帧.人脸检测除了可以大部分去除不可能含有播音员的镜头关键帧,还可以通过定位人脸在关键帧中的位置计算出提取视觉特征的区域以及检测 SIFT 特征点的区域.

为了快速地在镜头关键帧中检测出有人脸的镜头关键帧,过滤、去除不存在人脸的镜头关键帧,我们利用了基于 Haar 特征的 AdaBoost 算法[7,8].AdaBoost 算法是可以从基于 Haar 型特征的弱分类器空间中自动地挑选出若干弱分类器组合成强分类器的统计学习方法.利用 AdaBoost 算法学习瀑布型人脸检测器.

图 2 中的 F 区域是利用人脸检测算法检测出的人脸区域,人脸位置用(x,y,w,h)表示.其中,x 和 y 表示人脸区域左上角在关键帧中的坐标,w 和 h 分别表示人脸区域的宽度和高度.利用人脸检测算法,除了可以过滤没有播音员的镜头关键帧,当检测到一个以上的人脸时,还可以利用检测到的人脸位置判断关键帧是否为可能为同时包含两个播音员的镜头关键帧.设在镜头关键帧中检测到的两个人脸的位置分别为(x_1,y_1,w_1,h_1)和(x_2,y_2,w_2,h_2),且 $x_1<x_2$,人脸区域中心点的坐标分别为(x_{c_1},y_{c_1})和(x_{c_2},y_{c_2}),关键帧中心点的坐标为(x_k,y_k).镜头关键帧检测出的两个人脸的位置如果满足下面条件,就判断这个镜头关键帧为候选的同时包含两个播音员的镜头关键帧:

① $(x_1+2w_1<x_k)$ && $(x_2-2w_2>x_k)$;

② $(y_{c_1}<y_k)$ && $(y_{c_2}<y_k)$;

③ $(y_1>y_{c_2}\ y_{c_2})$ && $(y_2>y_{c_1})$;

④ $((fr\times w_1)<w_2)$ && $((fr\times w_2)<w_1)$,其中,fr=0.8;

⑤ $|(x_k-x_1)-(x_2-x_k)|<(w_1+w_2)/2$.

对镜头关键帧进行人脸检测处理后,定义两个集合 S_1 和 S_2,其中,S_1 包含所有剩余的镜头关键帧,S_2 包含所有的可能同时包含两个播音员的镜头关键帧,S_2 是 S_1 的一个子集.

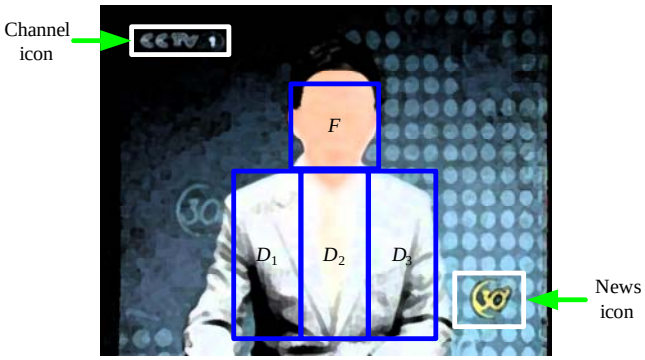

Fig.2 Spatial structure of the key frame of an anchor shot

图 2 播音员镜头关键帧结构

1.4 视觉特征提取与相似度度量

两个镜头关键帧是否相关,是通过匹配它们中的SIFT特征点来确定的.SIFT特征点的提取和匹配的运算量相对其他步骤是比较大的,为了尽量减少运算量,在提取和匹配 SIFT 特征点之前先计算两个镜头关键帧在视觉特征上的相似度,只有相似度高的镜头关键帧才进行 SIFT 特征点的匹配.对于给定的阈值 T_h,若第 i 帧和第 j 帧的视觉特征距离 $d(i,j)<T_h$,则提取和匹配这两个关键帧的 SIFT 特征点,否则认为这两帧不相似.

我们利用颜色直方图作为关键帧的视觉特征.播音员镜头背景可能有较大的变化,利用关键帧图像的全部区域提取颜色直方图作为视觉特征是不一定稳定的.不同人脸的颜色直方图大致相同,人脸区域的颜色直方图没有区分能力,所以也不统计人脸区域的颜色直方图.但在一个新闻节目的播报过程中,播音员的衣着是不变的,所以利用播音员的上半身区域的颜色直方图可以得到较稳定的视觉特征.同时,考虑到播音员的衣服中间部分的颜色和两边的颜色经常是不同的,所以根据人脸位置计算出人体上半身的位置,除头部以外,把上半身的区域分为 3 个同样大小的子区域,每个区域分别提取 HSV 颜色空间的颜色直方图来描述视觉特征.对于图 2 中的人脸区域,D_1,D_2 和 D_3 这 3 个区域就是相应的统计颜色直方图的区域.

设第 j 帧中 $D_n(n=1,2,3)$区域归一化的颜色直方图为 $hist_{j,n}(k),k=1,\ldots,N$,$N$ 为颜色直方图 bin 的个数.利用直方图相交的距离度量方法,则第 i 帧和第 j 帧的视觉特征距离为

$$d(i,j)=1-\frac{1}{3}\sum_{n=1}^{3}\sum_{k=1}^{N}\min(hist_{i,n}(k),hist_{j,n}(k)).$$

由于同时包含两个播音员的镜头中播音员比较小,且这种镜头中播音员的大小可能是变化的,通过比较颜色直方图不能作为判断关键帧是否相似的可靠标准,所以不对候选的同时包含两个播音员的镜头关键帧提取颜色直方图,而直接利用 SIFT 特征点匹配来判断镜头关键帧是否相关.

1.5 基于SIFT特征点匹配的相关镜头关键帧组检测

1.5.1 SIFT 特征点的提取与匹配

SIFT[9]特征描述子对图像的尺度变化和旋转是不变量,而且对光照和摄像机的三维视角变化具有一定的适应性.同时,这些特征描述子还具有较高的区分能力.SIFT 特征描述子的良好性质使得它非常适用于匹配不同图像中相同的物体或场景,基于 SIFT 描述子的匹配算法已经被成功地应用于目标识别、图像拼接等很多领域.

不同播音员镜头的背景可能有较大的变化,但播音员的衣着是不变的,并且在播音员报道新闻时,播音员身体基本保持不动,只有头或肩膀有一些小的运动,这些微小的变化不会影响到 SIFT 特征点的匹配.即便在有两个人播音员的新闻节目中,有些播音员镜头中只有一个播音员,有些播音员镜头中有两个播音员,有两个播音员的镜头中的播音员通常会比只有一个播音员的镜头中的播音员要小,但由于 SIFT 特征描述子对图像的尺度是不变量,也能在这两种镜头关键帧播音员所在的区域中找到相互匹配的 SIFT 特征点.所以,在两个有相同播音

员的镜头关键帧中,可以在播音员的上半身和脸部找到相互匹配的 SIFT 特征点,而不用考虑播音员在不同镜头中的大小.利用 SIFT 特征点匹配可以在一个镜头关键帧序列中找到候选的包含同一个播音员的一组镜头关键帧.

1.5.1.1 SIFT 特征点的检测

SIFT 特征点的检测包括 4 个步骤:① 尺度空间极值点的检测;② 极值点的精确定位;③ 计算特征点的主方向;④ 生成特征点的描述子.DoG 算子为两个不同尺度的高斯核的差分与图像的卷积.利用 DoG 算子对不同尺度大小的原始图像进行处理得到 DoG 图像,DoG 图像中极值点的值要比同一尺度的邻近 8 个像素以及相邻的上下两个尺度对应位置的 9×2 个像素的值都大或都小.对每个极值点处理得到它的精确位置和尺度,并根据稳定性度量标准剔除非稳定的极值点,剩下极值点就为 SIFT 特征点.在以特征点为中心的邻域窗口内采样,计算邻域像素的梯度直方图,直方图的峰值位置则作为该特征点邻域梯度的主方向.最后,每个特征点用 128 维的向量来表示,这样就得到了用于图像特征匹配的 SIFT 特征点描述符.

对只可能包含一个播音员的候选镜头关键帧,则按照上述步骤在关键帧中的人脸区域和由人脸区域计算出的 D_1,D_2 和 D_3 区域中检测 SIFT 特征点,如图 2 所示.对可能同时包含两个播音员的候选镜头关键帧,设人脸的位置分别为(x_1,y_1,w_1,h_1)和(x_2,y_2,w_2,h_2),则在以(x_1-w_1,y_1)为左上角坐标、宽$(x_2+w_2)-(x_1-w_1)$、高 $3\times\min(h_1,h_2)$的长方形区域中检测 SIFT 特征点.电视台台标(见图 2 左上角)和新闻节目图标(见图 2 右下角)会长时间存在不同的镜头中,检测 SIFT 特征点的区域必须排除这两个区域,避免不同镜头关键帧中,这些区域的 SIFT 特征点相互匹配影响相关镜头关键帧组的检测.

1.5.1.2 SIFT 特征点的匹配

通过计算两个特征点的 SIFT 描述符之间的欧氏距离作为 SIFT 特征点的相似度度量,在某候选镜头关键帧的全部 SIFT 特征点中,找出与基准镜头关键帧中某个特征点描述符欧氏距离最近和次近的两个特征点.如果最近的距离 d_0 与次近的距离 d_1 的比值 $r=d_0/d_1$ 小于阈值 T_r,则表示基准镜头关键帧中该特征点与候选镜头关键帧中距离最近的特征点匹配,否则,该特征点在候选关键帧中找不到匹配点.这种匹配方法简单快捷,但可能会产生误匹配.不同镜头中播音员的上半身一般只存在一些小差别,对于两个有相同播音员的镜头关键帧,它们之间能够正确匹配的 SIFT 特征点在这两个镜头关键帧的水平方向和竖直方向上的排列顺序是相同的.所以,为了排除误匹配的特征点,利用这些相互相匹配的 SIFT 特征点在水平方向和竖直方向上的相对位置是否相同来排除误匹配点.利用上述方法,如果两个镜头关键帧匹配的 SIFT 特征点个数小于给定的阈值 T_s,则两个镜头关键帧不相关,认为这两个关键帧不可能有相同的播音员;如果匹配的 SIFT 特征点个数不小于给定的阈值,则认为两个关键帧相关,可能有相同的播音员.

1.5.2 相关镜头关键帧组的自动检测

在第 1.3 节中,人脸检测处理后根据包含人脸的镜头关键帧的性质定义了两个集合 S_1 和 S_2,S_2 是 S_1 的一个子集.我们分别利用基于 SIFT 特征点匹配的相关镜头关键帧组自动检测算法对它们进行处理,分别得到候选的仅仅包含一个播音员的相关镜头关键帧组和同时包含两个播音员的相关镜头关键帧组.提取 S_1 和 S_2 中相关的镜头关键帧组的处理流程大致相同,如图 3 所示,只有几个步骤不同.下面将首先介绍处理的大致流程,然后介绍对 S_1 和 S_2 进行处理的不同方面.

假设要处理的镜头关键帧都预先保存在一个队列 Z_1 中,用$|Z_1|$表示队列中的镜头关键帧数.首先,从队列 Z_1 中取出一个镜头关键帧 k,以镜头关键帧 k 为基准,遍历在队列中的其他镜头关键帧,基于 SIFT 特征点匹配寻找与它相关的镜头关键帧.两个镜头关键帧是否相关,首先计算它们视觉特征的距离 $d(j,k)$,如果 $d(j,k)$小于某个阈值 T_h,则它们可能相关,然后计算它们之间能匹配的 SIFT 点的个数.如果 k,j 中有一个为候选的同时包含两个播音员的镜头关键帧,则不考虑视觉特征的相似度,直接计算它们之间能匹配的 SIFT 点的个数.如果能找到至少 T_s 个匹配的 SIFT 特征点,则认为两个镜头关键帧相关,取 $T_s\geq 3$[9].遍历完在队列 Z_1 中的所有其他关键帧后,队列 Z_3 中保存的镜头关键帧都是相关的镜头关键帧,认为它们是一类镜头,它们可能是播音员镜头.考虑到同一个播音员的镜头在新闻中出现的间隔比一般新闻镜头的时间长,计算队列 Z_3 中两两镜头关键帧间隔的最大值 t,如

果 $t \geq T_t$,则 Z_3 中的镜头关键帧将作为候选的播音员镜头关键帧组;否则认为它们是不可能包含播音员的镜头关键帧,可以丢弃不再考虑.当队列 Z_3 中只有关键帧 k 本身时,我们定义最大间隔 t=0.反复利用上述方法寻找相关的候选播音员镜头关键帧,直到最后队列 Z_1 中剩余的镜头关键帧数不大于 m,m 是根据处理集合的不同而选择的不同阈值.

对 S_1 中的镜头关键帧利用图 3 中的流程进行处理,取 m 为集合 S_2 中的关键帧数,选择的基准镜头关键帧 k 为只可能包含单个播音员的镜头关键帧.在流程的处理步骤 p 中,把队列 Z_3 中不属于 S_2 的镜头关键帧作为一个候选的只包含单个播音员的镜头关键帧组保存下来,而把所有队列 Z_3 中属于 S_2 的镜头关键帧重新保存在队列 Z_1 中.这样处理是为了处理某个播音员的镜头出现次数少的问题.处理后,将得到可能含有单个播音员的镜头关键帧组 SC_i,i=1,…,l.当 S_2 不为空时,对 S_2 中的镜头关键帧利用图 3 中的流程进行处理时,取 m 为 0,T_t=0,在处理步骤 p 中,把队列 Z_3 中所有的镜头关键帧作为可能同时包含两个播音员的镜头关键帧组保存下来,最后得到可能同时含有两个播音员的镜头关键帧组 TC_i,i=1,…,h.

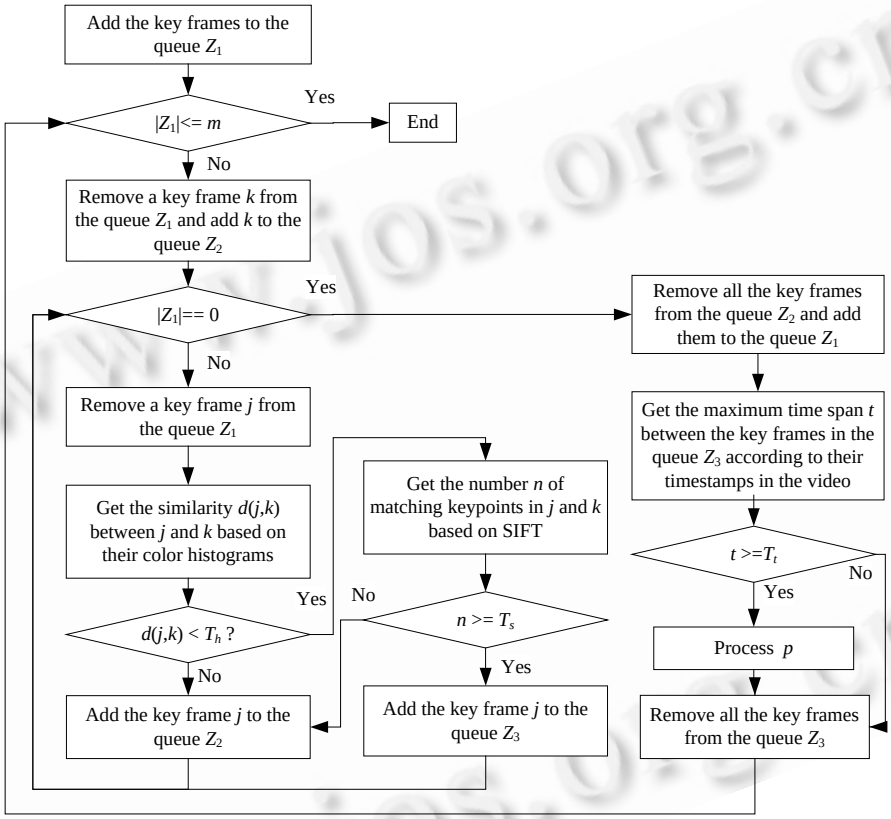

Fig.3 Processes of detecting shot groups based on SIFT

图 3 基于 SIFT 特征点匹配的相关镜头关键帧组检测流程

1.6 播音员镜头的判决

得到各种候选播音员镜头关键帧组 SC_i(i=1,…,l)和 TC_i(i=1,…,h)后,利用下面的方法对关键帧组进行判决:

① 如果 l=0,判断没有检测到播音员镜头;

② 如果 l=1,判断新闻节目中只有一个播音员,SC_1 中的所有镜头关键帧为播音员镜头;

③ 如果 l>1,h=0,则把处理中相互匹配的 SIFT 特征点数最多的那组镜头关键帧判断为播音员镜头;

④ 如果 l=2,h=1,判断新闻节目中有两个播音员,SC_1 和 SC_2 中的所有镜头关键帧为单人播音员镜头关键帧,

TC_1 中所有镜头关键帧为同时包含两个播音员的镜头关键帧;

⑤ 如果 $l>2,h=1$,不断取出 $SC_i(i=1,\ldots,l)$中的一个镜头关键帧 k_i,然后与 TC_1 中的镜头关键帧进行 SIFT 特征点匹配,找到能够匹配的 SIFT 特征点数最少的那个镜头关键帧 k_j,把包含 k_j 的 SC_j 中所有的镜头关键帧判断成为非播音员镜头关键帧,$l=l-1$,直到 $l=2$.之后的处理同情况④;

⑥ 如果 $l>2,h>1$,取出 $TC_i(i=1,\ldots,h)$中的一个镜头关键帧 k_i,然后 k_i 依次与 $SC_i(i=1,\ldots,l)$中的一个镜头关键帧进行 SIFT 特征点匹配,计算相互匹配的特征点数总和,找出总和数最多的那个镜头关键帧 k_j,则把包含 k_j 的 TC_j 中所有的镜头关键帧判断成为含有两个播音员的镜头关键帧.之后的处理同情况⑤.

2 实验结果

为了测试算法的通用性和可靠性,我们选择了 7 种不同种类的 8 个新闻视频对算法进行测试,用于实验的新闻视频,见表 1.表中形式如 n/m 的数据表示单人镜头数和双人镜头数,分别为 n 和 m,播音员镜头关键帧样本如图 4 所示.中央 1 台《新闻联播》的播音员镜头如图 4(d)所示,由于其背景有很多小屏幕不断显示不同的内容而导致背景的视觉特征不断变化.从表 1 中可以发现,中央 1 台《新闻 30 分》和陕西台的《陕西新闻联播》都只有 1 个同时包含两个播音员的镜头.《新闻 30 分》同时包含两个播音员镜头中的播音员比只有一个播音员镜头中的播音员小很多,如图 4(c)和图 4(g)所示.浙江卫视的《浙江新闻联播》中有 2 个同时包含两个播音员的镜头关键帧,如图 4(a)和图 4(b)所示,但镜头中的播音员大小是不同的.《陕西新闻联播》由于视频压缩比大的影响,提取的镜头关键帧整体比较模糊.从图 4 可以看出,不同新闻视频中播音员位置和大小是不同的,一个节目的模板不能对其他新闻节目进行处理,基于模板的方法必须对不同类型的新闻节目提取不同的模板.新闻节目 1 和新闻节目 8 都只有 4 个播音员镜头,对这种播音员镜头数较少的情况,基于聚类的算法将导致漏检[6].

Table 1 Experimental results in detecting anchor shots

表 1 播音员镜头检测结果

Video	Program name	N_t	Length (s)	N_d	N_f	N_m	Recall (%)	Precision (%)	Cost of detection (s)
1	CCTV news	2/2	1 800	2/2	0/0	0/0	100	100	83.773
2	CCTV news	11/2	1 800	10/2	0/0	1/0	92.3	100	128.352
3	News in 30 minutes	15/1	1 600	15/1	0/0	0/0	100	100	75.436
4	Sports news	15/0	1 500	15/0	0/0	0/0	100	100	53.948
5	International news	6/0	1 200	6/0	0/0	0/0	100	100	37.005
6	Zhejiang news	11/2	1 200	11/2	0/0	0/0	100	100	89.591
7	Xiashi news	10/0	930	10/0	0/0	0/0	100	100	60.701
8	Shaanxi news	3/1	1 800	3/1	0/0	0/0	100	100	53.449
Total		73/8	11 830	72/8	0/0	1/0	98.765	100	582.255

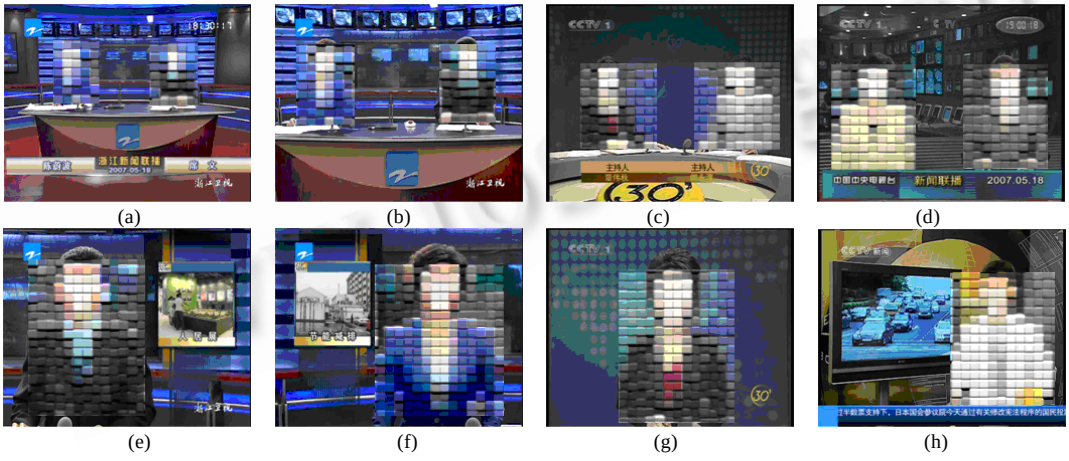

Fig.4 Key frames of anchor shots

图 4 播音员镜头关键帧

用 N_t 表示提取的镜头关键帧中实际播音员镜头数,N_d 表示算法检测到的正确的播音员镜头数,N_f 表示误检数,N_m 表示漏检的播音员镜头数,则召回率=$(N_d/N_t)\times100\%$,准确率=$(N_d/(N_d+N_f))\times100\%$.在实验中,$T_s$=3,$T_h$=0.5,$T_r$=0.54,对 S_1 中的镜头关键帧进行处理时,取 T_t=200s,表 1 为播音员镜头检测的实验结果.

从表 1 可以看出,本文算法具有较高的检测性能.新闻视频 2 漏检一个播音员镜头是因为提取的镜头关键帧中播音员的头扭到一边,只能看见播音员的侧面,人脸检测算法没能检测到人脸.这种情况可以通过训练一个侧面人脸的检测器解决,也可以对一个镜头提取两个不同时间点的关键帧,取出其中能检测到人脸的那个关键帧进行后面的处理,其他处理步骤可以不变.对于新闻视频中播音员镜头较少的情况,基于聚类的方法可能导致漏检,而从实验结果可以看出,本文的方法对播音员镜头的个数有比较强的适应性.对于如中央 1 台新闻联播背景由于有很多小屏幕而导致背景的视觉特征不断变化的情况,也能准确地检测到播音员镜头,算法受播音员背景变化的影响较小.同时,本文的方法能够准确地检测出哪些镜头同时包含两个播音员,哪些镜头只有一个播音员.在机器配置为 3.0GHz P4 CPU 1GB 的电脑处理全部 11 830s 的电视新闻节目,共需约 582.255s(不包括镜头关键帧提取时间),能够应用于实际的视频分析系统中.

表 2 引用了不同论文中播音员镜头检测算法的性能,作为本文算法与其他算法性能比较的一个参考.其中,文献[2]利用通用的模板进行播音员镜头检测.文献[4]中的方法需要训练样本,其通过非监督学习的方法提取出播音员镜头的音视频模板,用于判断候选镜头是否为播音员镜头.文献[6]利用镜头的颜色特征,基于自动聚类方法找出候选播音员镜头,然后根据播音员镜头出现的时空特征,用神经网络分类器对候选播音员镜头进行确认.文献[6]中的方法获得了较高的检测性能,但需要训练样本,训练和测试使用的是同一个电视台的新闻视频.本文方法除了训练一个通用的人脸检测器外,不需要模板,也不需要针对某类新闻节目训练特别的分类器.

Table 2 Performance measures for various anchor shot detection algorithms

表 2 不同播音员镜头检测算法性能的比较

	Proposed algorithm	Ref.[2]	Ref.[4]	Ref.[6]
Recall (%)	98.765	91.3	96.6	99.1
Precision (%)	100	100	95.4	98.2

3 结 论

通过观察不同类型的新闻节目发现,播音员镜头中存在一些普遍的规律,在此基础上提出了一种基于人脸检测与 SIFT 特征点匹配的播音员镜头自动检测算法.算法首先利用人脸检测器过滤出可能包含播音员的候选镜头,然后利用视觉特征判断镜头是否可能相似,再利用 SIFT 特征点匹配找出相关的镜头组,最后利用各镜头组的信息判断出哪些是播音员镜头.本文方法除了训练一个通用的人脸检测器外,不需要模板,也不需要针对某类新闻节目训练特别的分类器.实验结果表明,算法能够较为广泛地适应于各种不同种类的新闻节目、不同视觉质量的视频,得到较满意的结果,可以有效地应用于新闻视频分析.

References:

[1] Zhang HJ, Gong Y, Smoliar SW, Shuang YT. Automatic parsing of news video. In: Proc. of the Int'l Conf. on Multimedia Computing and Systems. Boston: IEEE Computer Society Press, 1994. 45−54.

[2] Ma YF, Bai XS, Xu GY, Shi YC. Research on anchorperson detection method in news video. Journal of Software, 2001,12(3): 377−382 (in Chinese with English abstract). http://www.jos.org.cn/ch/reader/view_abstract.aspx?flag=1&file_no=20010310&journal_id=jos

[3] De Santo M, Percannella G, Sansone C, Vento M. Unsupervised news video segmentation by combined audio-video analysis. In: Yeung DY, *et al*., eds. Proc. of the SSPR & SPR 2006. LNCS 4109, Berlin: Springer-Verlag, 2006. 273−281.

[4] D'Anna L, Marrazzo G, Percannella G, Sansone C, Vento M. A multi-stage approach for anchor shot detection. In: Yeung DY, *et al*., eds. Proc. of the SSPR & SPR 2006. LNCS 4109, Berlin: Springer-Verlag, 2006. 773−782.

[5] Shearer K, Dorai C, Venkatesh S. Incorporating domain knowledge with video and voice data analysis in news broadcasts. In: Proc. of the 1st Int'l Workshop on Multimedia Data Mining MDM/KDD 2000. 2000. 46−53.

[6] Yang N, Luo HZ, Xue XY. A method to detect anchorperson shots for digital TV news. Journal of Software, 2002,13(8): 1559−1567 (in Chinese with English abstract). http://www.jos.org.cn/ch/reader/view_abstract.aspx?flag=1&file_no=20020831&journal_id=jos

[7] Viola P, Jones MJ. Rapid object detection using a boosted cascade of simple features. In: Proc. of the Int'l Conf. on CVPR. IEEE Computer Society Press, 2001. I-511−I-518.

[8] Lienhart R, Maydt J. An extended set of haar-like features for rapid object detection. In: Proc. of the Int'l Conf. on Image Processing. IEEE Computer Society Press, 2002. I-900−I-903.

[9] Lowe DG. Distinctive image features from scale-invariant keypoints. Int'l Journal of Computer Vision, 2004,60(2):91−110.

附中文参考文献:

[2] 马宇飞,白雪生,徐光祐.史元春.新闻视频中口播帧检测方法的研究.软件学报,2001,12(3):377−382. http://www.jos.org.cn/ch/reader/view_abstract.aspx?flag=1&file_no=20010310&journal_id=jos

[6] 杨娜,罗航哉,薛向阳.一种用于电视新闻节目的播音员镜头检测算法.软件学报,2002,13(8):1559−1567. http://www.jos.org.cn/ch/reader/view_abstract.aspx?flag=1&file_no=20020831&journal_id=jos

杨武夷(1982－),男,福建泰宁人,博士,助教,主要研究领域为水声通信,水声信号处理,图像处理,多媒体内容管理,模式识别.

张树武(1964－),男,博士,研究员,博士生导师,主要研究领域为数字内容处理技术,多语言语音识别技术,数字媒体信息挖掘技术,音乐检索技术,音视频分类与识别技术,现代服务科学与技术.

曾智(1981－),男,博士,主要研究领域为多媒体内容管理.

李和平(1978－),男,博士,助理研究员,主要研究领域为模式识别,计算机视觉,多媒体.