一种基于语义的视频场景分割算法

曹建荣
(山东建筑大学信息与电气工程学院,济南 250101)

摘 要 针对如何在镜头基础上进行聚类,以得到更高层次的场景问题,提出了一个基于语义的场景分割算法。该算法首先将视频分割为镜头,并提取镜头的关键帧,然后计算关键帧的颜色直方图和 MPEG-7 边缘直方图,以形成关键帧的特征;接着利用镜头关键帧的颜色和纹理特征对支持向量机(SVM)进行训练来构造 7 个基于 SVM 对应不同语义概念的分类器,并利用它们对要进行场景分割的视频镜头关键帧进行分类,以得到关键帧的语义,并根据关键帧包含的语义概念形成了其语义概念矢量,最后根据语义概念矢量通过对镜头关键帧进行聚类来得到场景。另外,为提取场景关键帧,还构建了镜头选择函数,并根据该函数值的大小来选择场景的关键帧。实验结果表明,该场景分割算法与 Hanjalic 的方法相比,查准率和查全率分别提高了 34.7% 和 9.1%。

关键词 场景 支持向量机 语义 视频
中图法分类号: TN941 **文献标识码**: A **文章编号**: 1006-8961(2006)11-1657-04

An Algorithm of Video Scene Segmentation Based on Semantics

CAO Jian-rong
(*School of Information and Electrical Engineering, ShanDong Jianzhu University, Jinan* 250101)

Abstract The scene segmentation is a high-level temporal video segment. This paper presents a method of scene segmentation based on semantics. At first, the video clips are segmented into shots and the shot key frames are extracted. Then the features of color histogram and MPEG-7 edge histogram of each key frame are computed and the feature vectors of shot key frames are formed. The support vector machines (SVM) are trained by these feature vectors and 7 binary classifiers in accordance with difference semantic concepts are constructed. These binary classifiers are used to classify the shot key frames of the video clips based on the features of the color and the texture and the semantics concepts of shot key frames can be obtained. The semantic concept vectors of shot key frames are formed by the semantic concepts contained in the key frames. The shot key frames are clustered by the semantic concept vectors and the video scene can be constructed. The shot select function is defined to extract the scene key frame based on the value of function. The experimental results shown that the recall and the precision of this algorithm are higher than those of the Hanjalic's method about 34.7% and 9.1%, respectively.

Keywords scene, support vector machine(SVM), semantic, video

1 前 言

随着数字视频的大量出现,如何有效地管理和处理大量视频数据已成为很重要的问题,而基于内容的视频浏览和检索则是解决这一问题的有效方法,这种基于内容的视频浏览提供了一个快速和有效地查看视频内容的工具。视频浏览可以在镜头的层次上进行,也可以在场景的层次上进行,而后者则是更高水平意义上的视频浏览,因为它是对视频内容进行更加压缩和概括基础上的浏览。所谓场景是指发生在同一时间和/或同一地点描述一个事件或多个并行事件的一系列镜头的组合。为了在场景层次上对视频浏览,首先要把视频分割为镜头,并在镜头的基础上将相似的镜头聚类构成场景,然后提取基于场景的关键帧,即可得到浏览用的视频帧。近

基金项目:国家自然科学基金项目(60462001)
收稿日期:2006-06-28;**改回日期**:2006-08-05

第一作者简介:曹建荣(1965 ~),男,副教授。2006 年获北京邮电大学博士学位。主要研究方向是图像处理和视频检索。E-mail: jrcao65@sohu.com

十几年来，在这方面已经做了很多的研究[1-3]。

大多数场景分割算法都是以镜头关键帧为基础按相似度进行聚类或分割，聚类或分割中的阈值大多根据经验设定，但阈值设定得不恰当都会直接影响到场景构成，并会造成很多错误。为了解决这些问题，本文针对风光记录片这类视频，提出了一个基于语义的视频场景分割算法，即首先将视频分割为镜头，然后提取镜头的关键帧，并利用由支持向量机(support vector machine，SVM)构成的分类器来对镜头关键帧分类，并提取镜头关键帧的语义概念。本文定义了一个语义矢量，并对每个关键帧求取其语义矢量，然后根据语义矢量通过场景分割算法对视频进行场景分割。另外，还提出了一个提取场景关键帧的方法。本文方法没有阈值选择的问题，实验结果表明，本文算法优于文献[2]的方法。

2 支持向量机的基本原理

设一组训练数据 $(\boldsymbol{x}_i, y_i)_{1\leqslant i\leqslant N}$，其中，$\boldsymbol{x}_i \in \boldsymbol{R}^n$，$y_i \in \{-1,1\}$ 是 $\boldsymbol{x}_i$ 所属类的标号。若训练数据可以被一个超平面

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) > 0,\ i = 1,\cdots,N \tag{1}$$

分开，则训练数据是线性可分的。此时，总能通过调节向量 $\boldsymbol{w}$ 和标量 b，使下式成立

$$\min_{1\leqslant i\leqslant N} y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geqslant 1,\ i = 1,\cdots,N \tag{2}$$

这样，离超平面最近的点到超平面的距离是 $1/\|\boldsymbol{w}\|$。于是式(1)就变成

$$y_i(\boldsymbol{w}\cdot\boldsymbol{x}_i + b) \geqslant 1 \tag{3}$$

如果满足式(3)的向量集合能被超平面正确分开，并且离超平面最近的向量与超平面之间的距离是最大的，则这个超平面被称为最优超平面。寻找最优超平面就等价于在式(3)的条件下最小化 $\|\boldsymbol{w}\|$。

由于 $\|\boldsymbol{w}\|^2$ 是凸的，在线性约束条件(式(3))下最小化 $\|\boldsymbol{w}\|$ 可以通过 Lagrange 乘子得到。记 $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_N)$ 是 N 个非负的，且与约束(式(3))有关的 Lagrange 乘子，则最优化问题就归为最大化下式

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i\cdot\boldsymbol{x}_j \tag{4}$$

约束条件为：$\alpha_i \geqslant 0$，$\sum_{i=1}^{N} y_i\alpha_i = 0$

对非线性支持向量机，首先通过一些非线性映射将数据映射到高维特征空间，在这个特征空间中构造最优超平面。设将 $\boldsymbol{x}$ 通过函数 $\varphi(\boldsymbol{x})$ 映射到特征空间，则式(4)变为

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j \varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j) \tag{5}$$

令 $K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \varphi(\boldsymbol{x}_i)\cdot\varphi(\boldsymbol{x}_j)$，可称为核函数，若该函数对称且满足 Mercer 条件，则有

$$W(\boldsymbol{\alpha}) = \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j K(\boldsymbol{x}_i,\boldsymbol{x}_j) \tag{6}$$

判决函数为

$$f(\boldsymbol{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N}\alpha_i y_i K(\boldsymbol{x}_i,\boldsymbol{x}) + b_0\right) \tag{7}$$

对 SVM 更加详细的介绍可参阅文献[4]。

3 视频关键帧语义提取

为提取视频关键帧，首先应将视频分割为镜头，本文采用文献[5]方法分割视频，并提取镜头关键帧。

3.1 特征的选取

3.1.1 颜色特征

颜色特征采用颜色直方图表示，在颜色直方图中，为了降低维数，可将每个颜色分量的 bin 数固定为 16，则每个直方图的维数是 $16^3 = 4\,096$。已经证明这种方法对用 SVM 进行分类是合适的[6]。

3.1.2 边缘纹理特征

由于颜色直方图失去了图像的空间关系，因此为了得到图像的空间关系，本文选择了图像的 MPEG-7 边缘直方图特征来描述。MPEG-7 边缘直方图的具体求法可参阅文献[7]。

3.2 支持向量机的构造和训练

本文中的 SVM 的实现是基于 libsvm[8] 的，使用的是 C-SVM[4]，选用的核函数是径向基函数(radial basis function，RBF)，其表达式为

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = \exp(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2),\gamma > 0 \tag{8}$$

SVM 训练的视频素材选自风光记录片《西部采风》中的 4 集片段。训练前，首先对这些视频片段进行了镜头分割，并提取了镜头关键帧，总数是 435 帧，都是 88×72 像素的 DC 图；同时对每个镜头关键帧都计算了颜色直方图和 MPEG-7 的边缘直方图；并将所有的镜头关键帧根据其图像中的主要物体手工地分成 7 类，即建筑、山、沙漠、树、草原、湖和溪流 7 类，如果关键帧图像中的主要物体不好确定，这个关键帧图像就从实验数据库中删除；然后将每个关键帧的颜色和纹理特征值归一化后组合在一

起，作为这一帧的特征数据，即 SVM 的训练数据。训练中构造了 7 个两类分类器，每个分类器只针对一个语义概念，然后将 SVM 训练数据针对不同的分类器构成 7 个训练数据集，每个训练集训练一个分类器，使每个训练数据集的特征数据一样，只是前面的类的标号不同，如山的训练集只是有山的关键帧标号为 1，其余为 0；最后用每个训练数据集对其相对应语义概念的 SVM 进行交叉对比训练，来得到最优的参数 C 和 γ，即可完成对每个 SVM 的构造。

3.3　关键帧语义提取

本文选择了 12 个 MPEG 风光纪录片视频片段，这些片段选自《埃及》、《美国冰河公园》和《黄石公园》。关键帧语义提取时，首先对这些视频片段进行镜头分割，并提取镜头关键帧的 DC 图；然后计算每个关键帧的颜色和纹理特征，归一化后即形成关键帧序列的 SVM 测试数据；最后用训练好的 7 个 SVM 分类器分别对这些测试数据进行分类，这样根据分类的结果就得到了关键帧的语义概念。

4　视频场景分割

4.1　语义概念矢量

为在语义层次上分割场景，本文定义了一个视频关键帧的语义概念矢量。语义概念矢量 $V_S=[C_1, C_2, \cdots, C_N]$（下角 S 代表 semantic，下同），其中每个 C_i 与一个语义概念相对应，$C_i(i=1,\cdots,N)$ 的值是语义概念出现的次数，N 是语义概念的数量，这里 $N=7$。语义概念矢量可以根据 SVM 分类后得到的语义概念构成。例如，C_2 是与山这个语义概念对应，如果一个关键帧经过山的分类器后的分类结果为 1，则说明该关键帧图像是山的图像，C_2 的值是 1，否则是 0。

根据 7 个分类器分类的结果，就可以得到每个镜头关键帧的语义矢量，然后根据语义矢量就可以在语义层次上对视频进行场景分割。

4.2　场景分割算法

场景分割算法如下：

输入：视频片段的镜头关键帧数量 n、每个镜头关键帧的语义矢量、前向搜索的范围 F、后向搜索的范围 R，$F \geqslant R$

输出：场景边界（开始帧号和结束帧号）

注：*Preshot*：从当前镜头关键帧到前面的镜头关键帧的数目

Furshot：从当前镜头关键帧到视频片段最后一个镜头关键帧的数目

开始：

设当前镜头关键帧 S_i 为视频片段的第 1 个镜头关键帧

步骤 1：/* 向前 */

如果（$f=\min(F, Furshot)$ and $f\neq 0$）

{若 $Sim(S_i, S_{i+m})=0, m=1,2,\cdots,M$，则将 S_i 和 S_{i+m} 放到同一个场景中，$S_i=S_{i+m}$，更新 Preshot 和 Furshot 的值，回到步骤 1。}

步骤 2：/* 向后 */

$r=\min(R, Preshot)$，如果（S_i 在 r 范围内）

{若 $Sim(S_{i-r}, S_{i-r+m})=0, m=1,2,\cdots,M, i-r+m\neq i$，则将 S_{i-r} 和 S_{i-r+m} 放到同一个场景中，$S_i=S_{i-r+m}$，更新 Preshot 和 Furshot 的值，回到步骤 1。}

步骤 3：/* 场景边界 */

一个场景分割完成，同时记录场景的起始和结束帧号，更新 *Preshot* 和 *Furshot* 值。若 $Furshot\neq 0$，则回到步骤 1 开始下一个场景的搜索，否则场景分割完毕。

算法中若 V_i^s 中的 C_p 有一个等于 V_j^s 中的 C_p，$p=1,2,\cdots,7$，则 $Sim(S_i, S_j)=0$；若 V_i^s 中的 C_p 都不等于 V_j^s 中的 $C_p, p=1,2,\cdots,7$，则 $Sim(S_i, S_j)=1$。

上式说明当两个镜头关键帧语义矢量中有相同的语义概念时，则两个关键帧是相似（匹配）的，即 $Sim(S_i, S_j)=0$，否则是不相似的。

5　场景关键帧的提取

对风光记录片，考虑其拍摄特点，在提取场景关键帧时，应考虑以下几个原则：①在场景镜头中选择与其他镜头最相似的镜头；②镜头持续的时间长；③镜头内容的活动度小。根据这 3 条本文引入一个选择函数，计算该函数值，将其中大的作为备选镜头，并将该镜头的关键帧作为场景的关键帧。定义选择函数为

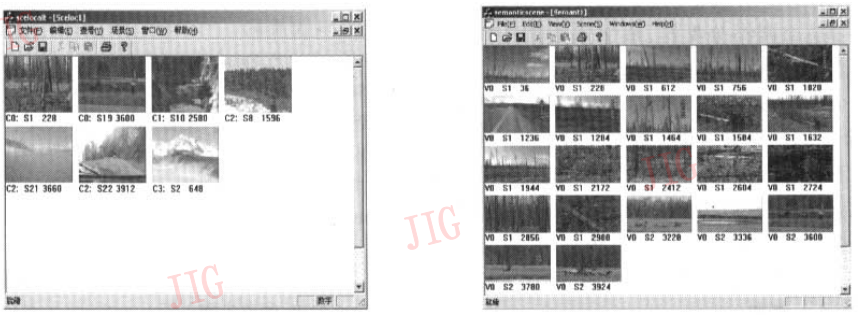
$$T(i)=\frac{L_i}{\sqrt{C(i)}\,e^{M_i}} \tag{9}$$

其中，L_i 是第 i 个镜头的长度；M_i 是第 i 个镜头的活动度；$C(S_i)$ 是场景中第 i 个镜头与场景中其他镜头的相似度的和，即

$$C(i)=\sum_{j\in S} Sim(S_i, S_j) \tag{10}$$

其中，$Sim(S_i, S_j)$ 是场景中第 i 个镜头和第 j 个镜头的相似度，可由镜头关键帧颜色直方图的欧氏距离计算。

图 1（a）是一个根据此函数选择的场景关键帧的例子。现以第 1 张图为例来说明图下标注的含义，C_0 是指第 0 个片段，S_1 是指第 1 个镜头，228 是指第 1 个镜头关键帧帧号。第 1 个片段所有镜头的关键帧可参见图 1（b）。



(a) (b)

图 1 提取的场景关键帧和正确分割的场景示例

Fig. 1 The example of extracted scene key frames and scene partitioned by our method

6 实验结果及讨论

为了对场景分割的结果进行评价,采用了与镜头边界检测相似的指标,即查全率(recall)和查准率(precision)。设 N_c(下角 c 代表 correct)和 N_t(下角 t 代表 total)分别是正确分割的场景边界和分割得到的全部场景边界(包括正确分割和错误分割的场景边界)数,N_a 是人工分割得到的全部的场景边界(作为正确的场景边界)数,则查全率 $R_{recall} = N_c/N_a$,查准率 $R_{precision} = N_c/N_t$。

实验中通过选择关键帧语义提取中用的 12 个 MPEG 风光纪录片视频片段来验证本文的场景分割算法。为了了解算法性能的优劣,实验中还将用本文算法与比较经典的文献[2]算法(简称为 SIM)进行的视频片段场景分割结果进行了比较。实验结果如表 1 所示。

表 1 实验结果

Tab. 1 The results of experiments

算法	镜头数	手工分割场景数	算法分割总场景数	算法分割正确的数	查全率(%)	查准率(%)
本文	174	33	27	25	75.8	92.6
SIM	174	33	38	22	66.7	57.9

从表 1 可以看出,本文算法的性能(查全率、查准率)优于 SIM 算法,主要原因是在风光记录片中,由于有很多摄像机的运动,使同一个场景中镜头的颜色变化很大,因此使用颜色等这些低层特征分割场景容易产生很多误判,从而会分割出比真实的场景多的场景。而本文算法由于提取了镜头关键帧的语义概念,因此,镜头颜色和纹理的变化并不直接影响分割的结果,其产生错误分割的原因是二元分类器错误分类导致语义概念提取错误,但这种情况只有场景边界的镜头关键帧语义提取错误时才会发生,因为本文算法是在一定范围内向前向后搜索比对,所以可以避免场景中间镜头关键帧个别语义提取错误带来的影响。

图 1(b)是本文算法正确分割的一个结果示例。这个片段由两个场景构成,分别标为 S_1 和 S_2,每个帧都是一个镜头的关键帧。

参考文献(References)

1 Rui Y, Huang T S, Mehrotra S. Constructing table-of-content for videos [J]. Journal of ACM Multimedia System, Special Issue Multimedia Systems on Video Libraries, 1999, 7(5): 359 ~ 368.

2 Hanjalic A, Lagendijk R L, Biemond J. Automated high-level movie segmentation for advanced video-retrieval systems [J]. IEEE Transactions on Circuits System and Video Technology, 1999, 9(4): 580 ~ 588.

3 Rasheed Z, Shah M. Detection and representation of scenes in videos [J]. IEEE Transactions on Multimedia, 2005, 7(6): 1097 ~ 1105.

4 Vpanik V. The nature of statistical learning theory[M]. New York: Springer Verlag, 1995.

5 Zhang Y, Rui Y, Huang T S, *et al.* Adaptive key frame extraction using unsupervised clustering[A]. In: Proceedings of International Conference on Image Processing[C]. Chicago, IL, USA, 1998: 886 ~ 870.

6 Chapelle O, Haffner P, Vapnik V N. Support vector machines for histogram-based image classification [J]. IEEE Transactions on Neural Network, 1999, **10**(5): 1055 ~ 1064.

7 Cao Jian-rong, Cai An-ni. A method for classification of scenery documentary using MPEG-7 edge histogram descriptor [A]. In: Proceedings of IEEE International Workshop on VLSI Design and Video Technology[C], Suzhou, China, 2005: 105 ~ 108.

8 Chang C C, Lin C J. LIBSVM: a library for support vector machine [EB/OL]. http://www.csie.ntu.edu.tw/~cjlin/libsvm. 2004