# 基于语义人脸的视频新闻标注

姚 青 吴 飞
(浙江大学人工智能研究所 杭州310027)

**摘 要** 视频和图像中的人脸蕴涵了丰富的语义信息,可以使用人脸对视频内容进行分析与标注,尤其是视频新闻节目。而要达到这样的目的,就必须先将对视频新闻具有语义价值的人脸从视频流中检测出来。本文提出基于语义人脸检测的视频新闻语义聚类与标注算法:在这个算法中,首先使用肤色模型检测人脸可能出现区域,然后提取人脸可能区域的独立成分特征,用训练好的支持向量机检测出所有人脸,套用语义人脸模板过滤出最终的语义人脸集合,最后通过高斯混合聚类,将视频新闻标注为主持人镜头、访谈类新闻镜头和其他新闻故事镜头三类。实验表明,该算法在视频新闻结构化中可以得到较好的应用。

**关键词** 语义人脸,视频新闻结构化,支持向量机,独立成分分析,语义聚类

## Video News Indexing Using Semantic-Face

YAO Qing WU Fei
(Institute of Artificial Intelligence, Zhejiang University, Hangzhou 310027)

**Abstract** The human faces in video and image imply lost of semantic contents, thus we can use faces to index and analyze video contents, especially for video news. In order to realize such goal, semantic human faces must be detected and recognized from video stream. This paper presents a new algorithm for semantic clustering and indexing of video news based on semantic-face: in this algorithm, complexion model is first used to detect possible face area; then pre-trained face/non-face support vector machine is used coarse-grained to recognize face and non-face respectively based on face independent component features from possible face area; third, the semantic-face template is used to filter out non-semantic-faces and we get legible and obverse semantic-faces; in the end, video news is segmented and classified into anchorperson shot, interview shot and other news story shot through mixture Gaussian clustering of semantic-faces. Structure used to index and explorer the video news is established. Experiment shows this algorithm works well for video news indexing.

**Keywords** Semantic-face, Video news structure, Support vector machine, Independent component analysis, Content clustering

## 1 简介

人物是视频新闻中最重要的语义成分,而人脸是表征人物身份的最主要的特征。因此利用视频新闻中出现的人脸信息来对视频新闻进行结构化是非常有意义的。

在传统的视频新闻结构化里[1],一般是分析每一个镜头中的关键帧的视觉特性,然后根据这些视觉特征来对镜头进行分类,得出主持人镜头。也有的文献中利用了视频新闻中的人脸信息来标注视频新闻[2],但是事实上并不是视频新闻中出现的所有人脸都是具有重要的语义价值的。基于此,本文要解决的问题就是从视频新闻所检测出的所有人脸中筛选提取出重要的语义人脸,然后基于这些语义人脸对视频新闻进行标注和结构化。

为了达到这样的目的,本文第2节介绍语义人脸的定义和检测方法;第3节介绍基于语义人脸的视频新闻结构化;实验数据和分析在第4节给出。

## 2 语义人脸的定义和检测

### 2.1 语义人脸的定义和分类

视频新闻为了将它所要表达的内容更加直接明白地告知观众,就必须保证其中出现的重要人物的脸部正面清晰地呈现给观众,同时也就意味这重要人脸的脸部器官(包括嘴巴,鼻子,左眼,右眼)也都是清晰可见的。

在视频新闻中存在着大量的人脸图像,主要包括主持人人脸、被访问人人脸、新闻人物人脸和环境人脸四类,其中主持人人脸、被访问人人脸和新闻人物人脸对视频新闻故事的语义内容具有十分直接的指示作用。例如主持人人脸的出现可以用来切分新闻故事;被访问人人脸的出现则标志着新闻中的热点的出现;而重要新闻人物的人脸的出现也就直接暗示了相关新闻的内容分类(例如观众看到哈里波特的脸就知道是娱乐新闻,看到乔丹的脸就知道是体育新闻)。如果将视频新闻中的所有具有语义信息的人脸提取出来,对其完成分类,并且利用每一个视频镜头所含有的语义人脸信息对该镜头进行标注,那么观众就可以方便而且快速地根据每一段新闻中出现的人脸来对该视频新闻进行检索和浏览。

基于如上分析,本文定义视频新闻中的语义人脸(semantic-face)如下:语义人脸是指视频新闻中出现的蕴涵大量语义信息的重要人脸;语义人脸对视频新闻的内容具有指示作用;语义人脸的特点是正面朝向,并且清晰;视频新闻中出现的语义人脸主要包括主持人人脸,访谈新闻中被访问人人脸,一般新闻中的主要人物人脸;语义人脸一般是所在画面的焦点所在。

### 2.2 语义人脸分层检测算法

人脸检测的基本思想是用统计或知识的方法对人脸建模,然后比较所有可能的待检测区域与已建立模型的相似度,从而得到人脸存在的可能区域。基于统计的方法从人脸图像中提取特征向量,把人脸检测问题转化为信号分布概率的检测,比如神经网络[3]和特征脸[4]方法。而基于知识的方法则利用先验知识为人脸建立若干规则,通过规则指导完成人脸检测,这里的规则包括人脸器官对称性和非对称性分布、轮廓形

状、颜色明暗、纹理粗细、运动方向等[5]。其中特征脸和人脸肤色模型是较为广泛使用的两种人脸检测技术。

但是上述传统的人脸检测方法都只是在一定精度内完成对一幅图像中是否存在人脸的判断，将人脸区域从背景图像中分离出来。由于并不是视频新闻中出现的所有人脸都是语义人脸，因此传统的人脸检测方法必须经过改进后，才能完成语义人脸的检测工作。

针对语义人脸正面朝向，并且清晰的特点，本文提出包含三个步骤的复合分层语义人脸检测算法：首先在 YCbCr 图像空间进行人脸肤色粗筛；然后使用支持向量机与独立分量分析(SVM/ICA)结构识别人脸块；最后套用语义人脸模板完成语义人脸的最终检测(如图1所示)。

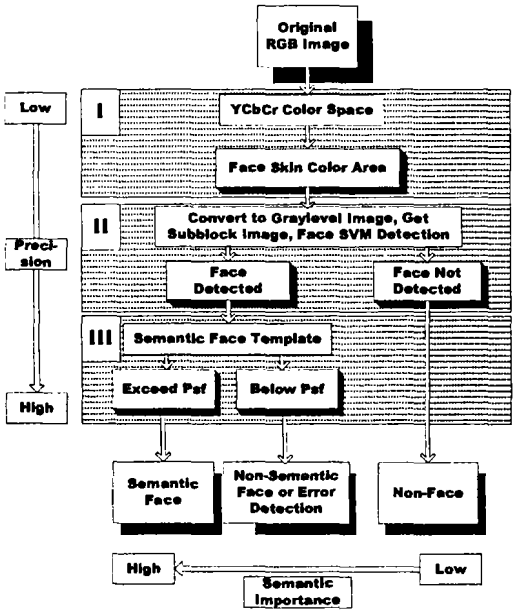


图1　语义人脸分层检测算法

*第一层：人脸肤色模型粗筛*　RGB 是一种被广泛采用的颜色模型，然而 RGB 却不是表征肤色的最佳模型，这是因为 RGB 颜色同时反映了颜色和亮度因素，而在人脸检测中，图像中的亮度变化对检测效果会造成很大影响。因此一般的肤色模型都将 RGB 颜色空间的图像转换到 HSV 颜色空间或者 YCbCr 颜色空间进行处理。本文采用的是 YCbCr 颜色空间中的 Cb 和 Cr 两个色度分量，用这两个色度的二维高斯分布来表示肤色模型。

肤色模型过滤出来的肤色部分还需要进行后期处理过程，后期处理的过程包括相连区域的合并、空洞的填充和去除小区域。

*第二层：基于 SVM/ICA 结构的人脸检测算法*　支持向量机(Support Vector Machine，SVM)起源于统计学习理论，它研究如何构造学习机，实现模式分类问题[6,7]。支持向量机使用结构风险最小化(Structural Risk Minimization，准则)原理构造决策超平面，使每一类数据之间的分类间隔(Margin)最大。SRM 准则认为：学习机对未知数据分类所产生的实际风险是由两部分组成的，并且满足如下关系：$R \leqslant R_{emp} + \sqrt{\frac{h(\log(2n/h)+1)-\log(\eta/4)}{n}}$，其中，$R$ 是实际风险，不等式的右边叫做风险边界，$R_{emp}$ 称为经验风险，$\sqrt{\frac{h(\log(2n/h)+1)-\log(\eta/4)}{n}}$ 叫做“VC 置信值”，$n$ 是训练样本个数，$h$ 是学习机的 VC 维($h$ 反映了学习机的复杂程度)。SVM 的思想就是在样本数目适宜的前提下，选取比较好的 VC 维 $h$，使经验风险 $R_{emp}$ 和置信值达到一个折中，最终使实际风险 $R$ 变小。这样，SVM 能够在样本数目适宜的前提下，取得实际最好分类效果，它近来已经在话者识别[8]、文本分类[9]、音频分类[10]和人脸识别[11]等模式分类领域取得了很大成功。

独立分量分析(Independent Component Analysis，ICA)起源于盲源分离问题(Blind Source Separation，BSS)，它与主分量分析和奇异值分解(Singular Value Decomposition，SVD)均属于线性变换技术，但是后两者只能按能量大小对数据进行分解，消除数据之间的二阶相关性，而 ICA 能够消除输入数据的高阶相关性[12]。在图像、视频和声音识别分类等应用中，可以提取的特征很多，特征之间存在相关性，并且特征重要特性一般隐藏在高阶统计特性中，因此使用 ICA 方法能够约减特征维数，并且使特征保持高阶相互独立，比只是消除2阶相关性的 PCA 和 SVD 方法更能提高识别正确率。Bartlett[12]使用了两种不同的方法(ICA1：独立图像基方法；ICA2：独立系数方法)来提取人脸 ICA 特征进行人脸识别，取得了比 PCA 更高的识别正确率。

图像和视频中的模式识别问题有许多都是基于图像子块来进行的。在这些问题中，正样本在纹理或者颜色等视觉特征上总体呈现某一种模式，形成一个闭集；而负样本则包括所有的反例，体现为一个开集。对于这种基于图像子块的二类识别，SVM/ICA 结构提供了一种很好的解决方法。其具体步骤如下：

1)选取具有代表性的正样本和负样本生成训练样本库；

2)根据具体的问题对样本数据进行预处理，例如去噪等；

3)使用独立成分分析算法得到所有正例的 ICA 图像基，图像基可以采取上述的任何一种方法生成；

4)根据提取出的 ICA 图像基计算训练样本中所有样本的 ICA 特征；

5)使用这些特征训练 SVM 分类器，应用训练好的 SVM 分类器对测试样本进行分类；

6)由于在人脸识别等许多模式识别问题中，负例样本不能用一个通用的模式进行描述，因此 SVM 分类器还需要不停地通过修正来完善，这一逐步过程(bootstrapping)通过在训练样本中添加误判结果来完成[13]。

本文的人脸检测算法就是基于上面 SVM/ICA 结构。由于在上一层人脸肤色粗筛中人脸的颜色特性已经被利用了，因此在这一层首先将上一步中得到的人脸可能区域图像块转换为灰度图像。接着对于每一个灰度图像子块都采用如下方法来确定人脸可能区域是否是真正人脸以及人脸具体位置。

根据统计可知，人脸的高宽比例大致为4/3。因此对于待测图块，按照4/3的高宽比例，以一定步长遍历，就可以得到所有的人脸可能子块。然后对每一个子块，提取其 ICA 特征，输入预先训练好的人脸识别 SVM 分类机，就可以判断出人脸是否存在。同时，如果存在人脸，那么人脸的具体位置也就可以确定了。如果有许多子块都被判断为人脸块，那么可以通过 Sigmoid 函数将 SVM 的分类输出结果转化为概率结果[14]，选取其中概率最大的子块作为检测出的人脸子块。

需要注意的是，由于背景色的干扰，在许多情况下，通过肤色模型筛选出的肤色块包括了人脸以及其周围的较大区域，因此只在原始尺度下按照长宽比提取子块很可能无法得到正确的结果。所以，对某一待测图块，如果在原始尺度下没有检测到人脸，那么就需要定义更小的子块高和宽，重复子块的遍历和人脸的检测过程，直到子块的大小达到一定的下限

或者检测到了人脸的存在。

*第三层：基于语义人脸模板的语义人脸检测* 在上一步中已经检测出了图像中的人脸，并且确定了人脸的所在区域。但是暂时只能作出人脸在哪里的判断，而无法具体知道人脸区域是否清晰，同时也可能存在误判人脸。本文对候选人脸通过套用语义人脸模板来完成语义人脸的检测。

在上面的分析中已经指出，语义人脸最主要的视觉特征就是正面而且清晰。而无论是衡量一幅人脸图像是不是正面或者是不是清晰，最主要的依据就是看其人脸器官是否都是清晰可见的。对于正面清晰的语义人脸，绝大多数情况下，嘴巴，鼻子，左眼和右眼这四个脸部器官应该都是清晰可见的。

因此，本文提出了语义人脸模板（semantic-face template）用于检测视频新闻中的语义人脸。

语义人脸模板包括两层子模板，人脸器官分布概率模板（facial organ distribution probability template，FODPT）和人脸器官相似概率模板（facial organ likely probability template，FOLPT）。人脸器官分布概率模板是通过对大量语义人脸的分析之后得出的各个脸部器官在二维人脸平面上的分布概率；人脸器官相似概率模板则是利用和人脸检测相类似的基于SVM/ICA的检测机制得出的图像子块成为人脸器官的可能性概率。这两层模板都包括嘴巴，鼻子，左眼，右眼四个人脸器官。

语义人脸模板的构造相应也包括两个部分：

1）分布概率模板：首先对训练库中的所有人脸图像均缩放到同一尺度（$W_t \times H_t$），在 $W_t \times H_t$ 的二维平面上统计得出每一个器官质心所在位置的高斯分布 $Dist_{organ}$，同时也得到了在这一标准尺度下每一脸部器官的平均尺寸 $Width_{organ}$ 和 $Height_{organ}$；

2）相似概率模板：对训练库中的每一类器官，计算提取出其ICA图像基，训练每一器官的SVM分类器，$SVM_{organ}$。

语义人脸模板的应用过程实际上就是求出每一个待测人脸为语义人脸的概率 $P_{sf}$。对于从第二层得到的每个人脸子块，为了判断其是否为语义人脸，计算这个子块的 $P_{sf}$ 值，其求解过程包括以下步骤：（图2为语义人脸模板示意图）

1）得出待测人脸图像块的实际宽度 $W_a$ 和高度 $H_a$，得到实际尺寸和模板尺寸的比例系数 $RW = W_a/W_t$，$RH = H_a/H_t$；

2）对于每一器官，计算其估计大致宽度 $W_{actualorgan}$ 和估计大致高度 $H_{actualorgan}$，$W_{actualorgan} = RW \times Width_{organ}$，$H_{actualorgan} = RH \times Height_{organ}$；

3）然后根据得到的人脸器官的大致实际尺寸 $W_{actualorgan}$ 和 $H_{actualorgan}$，在待测人脸图像中按照一定步长遍历，得到所有的器官子块（其中左眼的检测只遍历图像左半部分，右眼的检测只遍历图像的右半部分）；

4）接着计算所提取器官子块的ICA特征，分别输入到对应的 $SVM_{organ}$，得到这一子块识别为相应器官的概率 $P_{SVMorgan}$；

5）同时对于每一器官子块，计算出其质心坐标，然后映射到原始器官分布概率模板中求得 $P_{DISTorgan}$；

6）最后对每一类器官取 $P_{ORGAN}$（$P_{ORGAN} = P_{SVMorgan} \times P_{DISTorgan}$）为最大的那个子块作为该器官的最终检测结果；

7）计算 $P_{sf}$，$P_{sf}$ 为一个人脸区域中所有器官检测概率之和：

$$P_{sf} = SUM(P_{ORGAN}),$$

$$ORGAN = \{mouth, nose, lefteye, righteye\};$$

8）如果 $P_{sf}$ 超过一定的阈值，则认为待测人脸为语义人脸，完成了语义人脸的检测；否则人物待测人脸不是语义人脸。

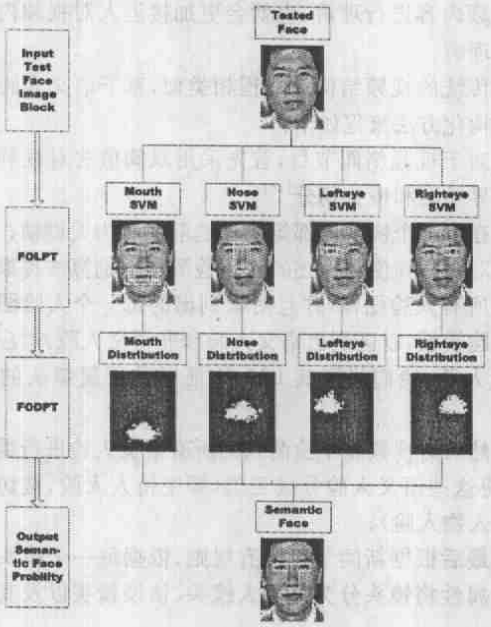


图2 语义人脸模板

## 3 基于语义人脸的视频新闻检索

文[15]将一段视频按照语义等级不同，结构化为视频-场景-镜头组-镜头-关键帧。其中关键帧是在一个镜头中所选取的具有代表性的一帧或多帧图像。按照这一结构对视频的内容进行结构化得到的视频结构称为视频内容表（TOC）。可以说，TOC是对通用的视频数据流所采取的结构化手段。目前对视频新闻这个特殊数据流，进行内容分析的目的就是把切分出来的镜头组合成一个个独立新闻故事，使观众可以单独快速了解感兴趣的独立新闻事件，而不需要把整个时段所有新闻节目都看下来。因此视频新闻节目结构化分析就是要得到视频新闻的每个场景（独立新闻故事），从而得到视频新闻的结构。

为了得到视频新闻TOC，需要经过如下三步：首先将新闻视频流分割成一个个的镜头单元；然后根据事先定义的镜头模型将这些切分出来的各类镜头归类；最后是把分类后的镜头单元组合成独立新闻故事（场景）。在这三步中，将镜头进行分类，归属到每个镜头组是最重要一步，传统的做法都是提取每个镜头关键帧所包含的低层视觉特征，如颜色直方图和运动信息等，作为某一镜头的特征，然后使用这个特征，通过聚类来将相似镜头归属到同一集合[1]。由于从关键帧中所提取的低层视觉特征并不包含有高层语义信息，因此如果一个独立新闻故事中，同一类镜头包含较多的剧烈视觉变化，就很难通过低层视觉特征聚类方式将这些镜头聚类成一个镜头集合。

### 3.1 基于语义人脸的视频新闻TOC

对于主持人新闻和访谈类新闻等视频新闻节目，其中出现的人脸蕴涵了丰富语义信息。因此如果对这些节目的关键帧使用上述语义人脸识别方法，检测出其中所包含的语义人脸，然后对整段视频中出现的所有语义人脸图像通过聚类进行分类，得到每一语义人脸图像块的分类信息以及每一类语义人脸的时间分布信息，最后对每一个镜头，使用其关键帧中所包含的语义人脸信息作为特征，就可以实现视频新闻节目

语义聚类与分析，从而对新闻节目结构化和内容标注。这种基于语义人脸的新闻节目语义聚类和标注方法是在较高语义级上对视频内容进行理解，也就会更加接近人对视频内容的主观语义理解。

与传统的视频结构化过程相类似，基于语义人脸的视频新闻结构化方法流程如下：

1)对于视频新闻节目，首先采用双阈值法对视频流进行镜头边界检测和镜头切分[16]；

2)在每一个镜头内部等步长选取帧作为关键帧；

3)对每一帧使用上述的人脸检测和识别算法提取其中所包含的所有人脸图像，并且对识别出的每一个人脸图像套用语义人脸模板，以区别出语义人脸和非语义人脸。对应检测出的语义人脸，我们提取其 ICA 特征以及所属镜头的时间特征；

4)然后对视频流中检测到的所有语义人脸进行混合高斯聚类，将这些语义人脸分成三类(即主持人人脸、被访者人脸和新闻人物人脸)；

5)最后根据新闻节目特有规则，依据每一个镜头中出现人脸的属性将镜头分为主持人镜头、访谈镜头以及其他新闻镜头；

6)按照一定的规则，通过相应镜头组合实现新闻故事提取。

于是，视频新闻 TOC 可以表示为：视频新闻-场景(新闻故事)-镜头组-镜头-关键帧-语义人脸。

### 3.2 基于高斯混合聚类的语义人脸分类

上面已经谈到，为了实现视频新闻的结构化，最重要的就是完成镜头的分类，实现主持人镜头的识别。在较高的语义层，通过分析每个镜头所包含的语义人脸的特征和时间分布就可以完成镜头的分类。在这里，主要基于以下两条前提假设将视频新闻中出现的所有语义人脸划分为三大类：

① 同一视频新闻中出现的同一个人的人脸外貌特征不可能有显著的变化。首先通过递规二类混合高斯聚类将视频新闻中检测出的所有语义人脸划分为相互独立的子集，具体递规过程按照如下约束进行：如果聚类得到的任一子类的方差大于阈值 $T_{std}$，则对该子类继续递规地进行二类混合高斯聚类分类，直至每一集中的语义人脸的方差均小于 $T_{std}$。经过划分之后，可以得出这样的结果：同一子集中仅包含相似的人脸，而同一个人的人脸不会被划分到不同的子集。

②每一类语义人脸在视频新闻中时间分布各不相同。根据对大量视频新闻的统计可知，主持人人脸、被访者人脸和新闻人物人脸在视频新闻中的出现时间有着各自的特点。通过分析上一步中划分出来的每一子类中的语义人脸的时间分布属性，就可以完成对这三大类语义人脸的划分了。

主持人人脸在视频新闻中会频繁出现，而且一般在新闻的开始和结束时都会出现。因此如果某一子类中的语义人脸在视频新闻中多次有规律地出现，那么就认为这一子类语义人脸为主持人人脸(A)。

被访者人脸在视频新闻中也会多次出现，但是出现的时间会局限在一个时间窗口区域内，而在其他时刻不会再次出现。因此如果某一子类中的语义人脸在一定的时刻内多次出现，则认为是被访者人脸(V)。

新闻人物人脸在视频新闻中的出现是随机的，而且某一人脸一般不会多次出现。因此，所有仅包含极少数人脸的子类以及包含的语义人脸随机出现的子类均认为是新闻人物人脸(O)。

### 3.3 镜头分类

对每一个镜头，通过分析其中包含的所有语义人脸的分类属性，应用一定的规则就可以确定出该镜头类型。实验中定义如下规则：

如果属于某一镜头的人脸绝大多数属于主持人人脸 A，而且连续出现，那么该镜头为主持人镜头(AS)；

如果某一镜头中出现较多主持人人脸 A 和较多被访谈者人脸 V，而且 V 和 A 相继出现，则认为是访谈镜头(VS)；

如果某一镜头中只有新闻人物人脸 O，或者不存在人脸，则定义为其他新闻镜头(OS)。

最后我们定义新闻故事如下：从一个 AS 开始，经过若干个 VS 或 OS 的组合，到达其后续的 AS 的镜头。这样，就可以将连续视频新闻内容切分为新闻访谈场景和其他新闻场景。图3是对一段视频新闻结构化的结果。

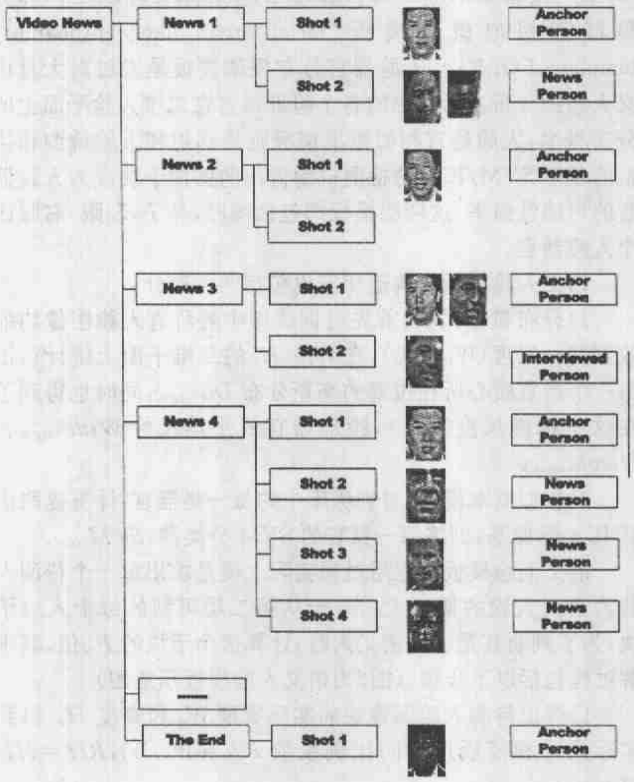


图3 基于语义人脸的视频新闻结构化

## 4 实验数据和分析

### 4.1 人脸检测实验数据和分析

实验中使用的是 Olivetti Research Laboratory 的人脸库。这个人脸库包括40个不同肤色、种族和性别的人，在不同的时间和不同的光照条件下拍摄的脸部灰度图像各10幅。其中若干人脸图像存在部分脸部遮挡(戴眼镜)和脸部有不同侧向。在 Olivetti Research Laboratory 人脸库的基础上加入了部分从 CCTV 新闻中提取的人脸灰度图像100幅。这样，总计500幅人脸图像作为训练样本。

测试样本使用了500幅 RGB 图像，包括照片，CCTV 新闻中提取的图像以及从网络随机获取的图像。测试样本包括不含人脸的图像、含清晰人脸的图像以及包含不清晰人脸的图像。

为了对 ICA/SVM 识别算法的效果进行对比，在使用 SVM 进行人脸检测时，分别提取 PCA 特征和两种 ICA 特征：ICA1(独立图像基)和 ICA2(独立系数)，均为36维特征。

图4是按照 PCA 得到的特征人脸，图5和图6是两种不同 ICA 算法得到的人脸基。

图4 PCA 特征人脸　　图5 ICA1图像基　　图6 ICA2图像基

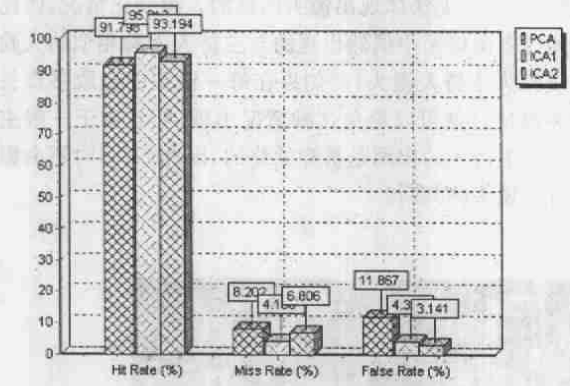


图7 人脸检测中使用 PCA，ICA1，ICA2特征结构比较

对测试样本使用肤色模型得到人脸可能区域，使用三种不同的特征训练了各自的 SVM 人脸分类器，实现人脸识别，其识别结果如图7所示。

总体上来说，在相同维数下，ICA 特征取得了比 PCA 特征更好的识别效果，因为 SVM 通过核函数把低维特征向高维空间映射，而 ICA 特征在高阶统计上保持独立。使用 ICA1 方法时的部分检测结果如图8所示。

### 4.2 语义人脸检测实验数据分析

对于测试样本中使用 ICA1特征识别出的总计574幅人脸图块，选取100幅进行语义人脸检测测试。首先人为手工分为三类，语义人脸（人脸图像清晰正面，对图像有重要语义内容），非语义人脸（人脸图像小或者模糊，不包含语义内容）以及非人脸（误判人脸）。然后进一步套用语义人脸双层模板。其中四个脸部器官在92×112的标准人脸模板二维平面的分布如图9所示；对于每一类人脸器官的 SVM/ICA 检测，均提取36维 ICA1特征。

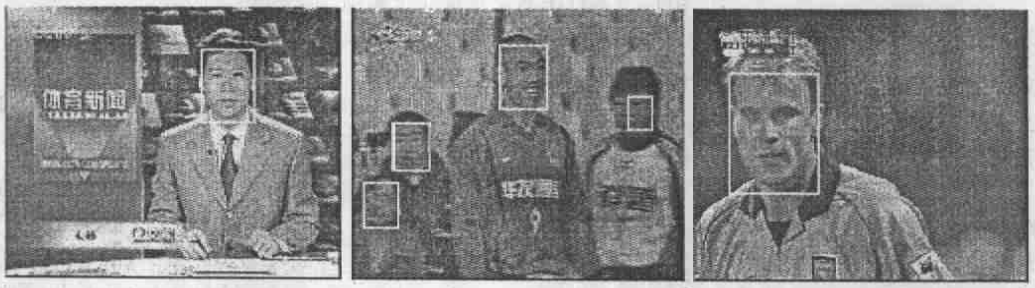

图8 人脸检测结果

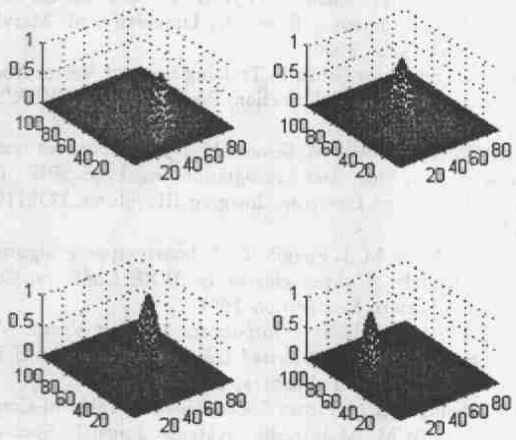

图9 人脸器官分布概率模板（左上：嘴巴，右上：鼻子，左下：右眼，右下：左眼）

$P_{sf}$阈值对于语义人脸的检测具有十分重要的意义。如果$P_{sf}$取值过低，虽然可以保证绝大部分的语义人脸不被误判，但会导致许多并不清晰的人脸或者局部类似人脸的非人脸图块被误判为语义人脸；而由于在人脸检测中检测出的一些清晰人脸存在中心偏移的问题，如果$P_{sf}$取值过高，虽然

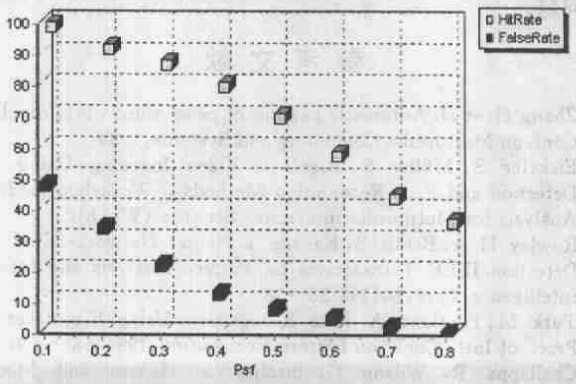


图10 $P_{sf}$取值对语义人脸检测的影响

可以降低误判率，但是会导致这些语义人脸的漏判。$P_{sf}$的取值与语义人脸检测的准确率，漏判率的关系由图10表示，从中可以得出，$P_{sf}$取0.3-0.4综合检测效果较好。

基于语义人脸的视频新闻检索实验数据分析：实验中对只有一个主持人的CCTV体育新闻视频进行主持人镜头检测测试(测试视频总计2小时长)。由于主持人镜头检测是视频新闻结构化中最重要的工作，因此本算法的主持人检测的效果和传统基于颜色直方图方法的效果对比如表1所示，其中N取值均为1。当每一个镜头中选取关键帧数N分别取1，3，5时，其实验结果如表2所示。

表1 主持人镜头检测结果比较

| | 准确数 | 错误数 | 遗漏数 |
|---|---|---|---|
| 基于语义人脸的方法 | 68 | 2 | 0 |
| 基于颜色直方图的方法 | 66 | 5 | 2 |

表2 主持人镜头检测结果和N的关系

| | 准确数 | 错误数 | 遗漏数 |
|---|---|---|---|
| $N=1$ | 68 | 2 | 0 |
| $N=3$ | 70 | 0 | 0 |
| $N=5$ | 70 | 0 | 0 |

表1的结果表明了基于较高层语义的语义人脸方法由于利用了视频新闻中的比较稳定的且携带较多语义信息的人脸特征，因此取得了更好的效果。

传统的主持人镜头检测方法中由于使用的是图像帧的低层视觉特征，因此要求主持人帧的背景不能发生显著变化，否则少数与其他帧背景不同的主持人帧会被漏判。而使用了主持人人脸信息之后，就不再受背景变化的影响。例如图11中的第一幅这一类主持人帧只在测试视频中出现1次，如果按照低层视觉特征聚类，必然会和其他主持人帧分离，而在本算法中，由于这一帧中包含的人脸信息和其他帧相似，因而被正确地分类为主持人帧。

从表2结果中可以看出，当仅从每一镜头中提取一帧作为关键帧时，由于无法体现出镜头内部的人脸变化情况，因此如果新闻内容关键帧中偶然出现的与主持人人脸相似的人脸就会被误判为主持人镜头，而如果在每一镜头中提取多步长帧作为关键帧时就可以避免这种情况出现。而且由于一般主持人镜头中主持人的出现是贯穿始终的，因此N>1时不会影响到主持人镜头的判断。

图11 本算法不受背景变化影响

**结论** 本文使用独立成分分析与支持向量机方法实现视频新闻中的语义人脸的识别方法；并且通过聚类实现了基于语义人脸的视频新闻的分类标注，在实验中取得了良好的效果。

今后的工作集中在三个方面：(1)人脸识别只是基于人脸的视频(图像)内容分析的第一步，人脸本身还蕴涵了许多信息，还有大量的工作可以开展，例如表情检测，注视方向检测等等；(2)本文对语义人脸的定义还是在较低的视觉特征层次，如何在较高的语义层来实现语义人脸的定义和检测，是需要研究的另外一个问题，并且，将本文中的算法应用到大测试样本中去，也是值得研究的；(3)SVM分类器的产生本质上是基于监督方式，在这个过程中，需要人为对输入数据进行初始标注。如何在非监督或半监督方式下训练SVM是值得研究的问题。

## 参考文献


1 Zhang H, et al. Automatic parsing of news video. In: Proc. IEEE Conf. on Multimedia Computing and Systems, 1994
2 Eickeler S, Müller S, Rigoll G. Video Indexing Using Face Detection and Face Recognition Methods. Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)
3 Rowley H A, Baluja S, Kanade T. Neural Network-Based Face Detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(1): 23～38
4 Turk M, Pentland A. Face Recognition Using Eigenfaces. In: Proc. of Intl. Conf. on Pattern Recognition, 1991. 586～591
5 Chellappa R, Wilson C, Sirohey S. Human and Machine Recognition of Faces: A Survey. In: Proc. of IEEE, vol. 83, May 1995
6 Vapnik V. The Nature of Statistical Learning Theory. Springer, New York, 1995
7 Burges C J C. A tutorial on support vector machines for pattern recognition. Knowledge Discovery and Data Mining, 1998, 2(2)
8 Schmidt M. Identifying speaker with support vector machine. In Interface'96, Sydney, 1996
9 Joachims T. Text categorization with support vector machines: Learning with many relevant features. In: European Conf. on Machine Learning (ECML-98), 1998
10 Learning Algorithms for Video and Audio Processing: Independent Component Analysis and Support Vector Machine based Approaches, Yuan Qi, LAMP-TR-056 (CAR-TR-951), Center for Automation Research, University of Maryland at College Park, Aug. 2000
11 Osuna E, Freund R, Girosi F. Training Support Vector Machines: an Application to Face Detection. Prodessings of CVPR'97, June 17-19, Puerto Rico
12 Bartlett M S, Lades H M, Sejnowski T J. Independent component representations for face recognition. In: Proc. SPIE Conf. on Human Vision and Electronic Imaging III, volume 3299, 1998. 528～539
13 Vetter T, Jones M J, Poggio T. A boostraapping algorithm for learning models of object classes. In: IEEE Conf. on Computer Vision and Pattern Recognition, 1997
14 John C. Platt Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. U Fayyad, ed. Kluwer Academic Publishers, Boston, 1999
15 Rui Y, Huang T S, Mehrotra S. Constructing Table-of-Content for Videos. in ACM Multimedia Systems Journal, Special Issue Multimedia Systems on Video Libraries, Sept. 1999
16 Zhang H J, Kankanhalli A, Smoliar S W. Automatic Partitioning of Full-Motion Video. ACM/Springer Multimedia Systems, 1993, 1(1): 10～28