一种基于学习的视频字幕验证方法

王 勇[1),2)] 李建彬[2)] 胡德文[1)] 郑 辉[2)]

[1)](国防科技大学机电工程与自动化学院，长沙 410073)

[2)](西南电子电信技术研究所现代信号处理国家重点实验室，成都 610041)

摘 要 视频字幕验证是字幕检测中的重要环节，其目的在于提高检测准确率。当前的验证方法多是依据经验规则。这些方法在图像背景复杂、图像分辨率低以及字幕字体、大小、颜色多变这些条件下，适应性差。为提高验证方法的适应性和准确性，通过将 2 维主成分分析(2DPCA)应用到视频字幕验证中，提出了一种基于 2DPCA 和支撑向量机(SVM)的视频字幕验证方法。该方法分训练和判别两个步骤，即首先采用 2DPCA 方法提取视频图像块特征，然后通过训练 SVM 对图像块进行验证和分类。实验结果表明，在图像背景复杂、图像分辨率低以及字幕字体、大小、颜色多变这些传统验证方法或多或少都存在困难的条件下，该方法不仅具有良好的视频字幕验证能力，而且也能明显降低算法的运行耗时。

关键词 视频字幕 字幕验证 2 维主成分分析 支撑向量机

中图法分类号：TP391.4 **文献标识码**：A **文章编号**：1006-8961(2006)11-1645-05

A Learning Based Approach to Validate Text in Video

WANG Yong[1),2)], LI Jian-bin[2)], HU De-wen[1)], ZHENG Hui[2)]

[1)](*College of Mechatronics and Automation, National University of Defence Technology, Changsha* 410073)

[2)](*National Key Laboratory of Modern Signal Processing, Southwest Institute of Electron & Telecom. Techniques, Chengdu* 610041)

Abstract For improving accuracy, validating text is a key step of detecting text in video. The current approaches mostly based on experiential rules. The approaches are not adaptive, in condition of complex background, low resolution, varied font, size, color of text in video. For improving adaptability and accuracy of validating text, the application of two-dimension principal component analysis(2DPCA) for video frame processing is investigated and a novel 2DPCA and support vector machine(SVM) based approach for validating text in video is proposed. The approach has two steps of training and validating. Firstly, 2DPCA is adopted to get the features of video image patches. Then, SVM is trained to validate and classify video image patches. The experimental results illustrate that the novel approach for validating text in video is more effective and costs less time than the other approaches, in condition of complex background, low resolution, varied font, size, color of text in video.

Keywords text in video, text validation, 2DPCA, support vector machine(SVM)

1 引 言

视频中的字幕(如新闻标题、节目对白等)是视频高层语义内容的重要来源，其对视频高层语义的自动理解、索引和检索非常有价值[1,2]。在视频字幕检测过程中，由于图像背景复杂、图像分辨率低以及字幕字体、大小、颜色多变，不可避免会误把非字幕图像区域判为图像区域。许多检测方法提高查全率的代价就是降低准确率，从而产生大量虚警。为了减少虚警和提高检测准确率，视频字幕验证成为字幕检测中的重要环节。目前的验证方法主要有以

基金项目：国家专项工程项目(613)；国家杰出青年科学基金项目(60225015)

收稿日期：2006-04-18；**改回日期**：2006-08-04

第一作者简介：王勇(1976～)，男。2001 年于国防科技大学机电工程与自动化学院获工学硕士学位，现为国防科技大学与西南电子电信研究所联合培养博士研究生。主要研究方向为智能图像识别、基于内容的多媒体检索。E-mail：cncdwy@yahoo.com.cn

下两种:一是基于单帧图像的启发式经验规则,如字幕行高度、字幕行的宽高比、区域填充率、区域纹理特点;二是基于视频多帧的方法[3,4],由于字幕通常出现在连续的若干帧中,因此通过对若干帧的检测,可利用投票来判定相同区域是否是字幕[5,6]。

在基于单帧图像的启发式经验规则中,往往是根据经验进行约束判断。由于不同视频节目的字幕编辑风格多种多样,因此造成字幕字体、大小、颜色多变,而且由于视频背景复杂,字幕又是嵌入到视频背景中的,因此字幕特征是不可预测和复杂的。由此可见,视频字幕的验证不能只考虑字幕本身的固有特征,还应该考虑一种学习机制去处理这些多变因素。从模式分类的角度来说,基于机器学习的方法使用一种学习机制(如支持向量机等)构造一个分类器,用于在视频帧中对图像块进行字幕与非字幕的两模式分类。

2 维主成分分析(two - dimension principal component analysis, 2DPCA)[7]方法是近两年发展起来的一种新方法。它是在主成分分析(principal component analysis, PCA)的基础上,开发的一种图像特征提取算法。它不需要将图像矩阵变换为相同维数的行向量,而是直接依赖原始图像矩阵构建图像总方差矩阵,然后用该图像总方差矩阵进行主成分分析。因图像总方差矩阵的尺度往往远小于原始图像经行变化获得的矩阵,故 2DPCA 方法比 PCA 方法更节省计算时间。因为这一特点,使 2DPCA 在图像处理领域倍受关注,并且其已经很好地应用于如人脸识别等领域[8]。

本文针对传统视频字幕验证方法中遇到的图像背景复杂、图像分辨率低以及字幕字体、大小、颜色多变的不利条件,探讨了 2DPCA 在视频字幕的检测提取中的应用。

2 2DPCA 与 SVM

2.1 2DPCA 原理

经典的主成分分析方法是基于 1 维向量,而 2 维主成分分析方法直接针对 2 维图像数据[8]。

令 $\boldsymbol{x}$ 为 n 维单位列向量,$\boldsymbol{A}$ 为 $m \times n$ 大小的随机矩阵,即图像矩阵。通过线性变换

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} \tag{1}$$

即可得到 $\boldsymbol{A}$ 映射在 $\boldsymbol{x}$ 上的特征向量。为了度量映射向量 $\boldsymbol{x}$ 的判别力,引入映射样本的总类分散度(scatter)。总类分散度用映射特征向量的协方差矩阵的迹来描述。采用准则

$$J(\boldsymbol{x}) = \mathrm{tr}(\boldsymbol{S}_x) \tag{2}$$

其中,$\boldsymbol{S}_x$ 为训练样本的映射特征向量的协方差矩阵,$\mathrm{tr}(\boldsymbol{S}_x)$ 表示 $\boldsymbol{S}_x$ 的迹。最大化上述准则的意义就是要找到最优映射轴 $\boldsymbol{x}_{\mathrm{opt}}$,使得将所有样本映射该方向后能够使映射样本的总类分散度最大。假设共有 M 个训练样本,将第 j 个样本记作 $m \times n$ 维矩阵 $\boldsymbol{A}_j (j = 1, 2, \cdots, M)$,将所有样本的平均图像记作 $E(\boldsymbol{A})$,其协方差矩阵 $\boldsymbol{S}_x$ 定义为

$$\begin{aligned}\boldsymbol{S}_x &= \frac{1}{M}\sum_{j=1}^{M}(\boldsymbol{y}_j - E(\boldsymbol{y}))(\boldsymbol{y}_j - E(\boldsymbol{y}))^{\mathrm{T}} \\ &= \frac{1}{M}\sum_{j=1}^{M}[(\boldsymbol{A}_j - E(\boldsymbol{A}))\boldsymbol{x}][(\boldsymbol{A}_j - E(\boldsymbol{A}))\boldsymbol{x}]^{\mathrm{T}}\end{aligned} \tag{3}$$

于是

$$\mathrm{tr}(\boldsymbol{S}_x) = \boldsymbol{x}^{\mathrm{T}}\left(\frac{1}{M}\sum_{j=1}^{M}(\boldsymbol{A}_j - E(\boldsymbol{A}))^{\mathrm{T}}(\boldsymbol{A}_j - E(\boldsymbol{A}))\right)\boldsymbol{x} \tag{4}$$

定义图像 $\boldsymbol{A}$ 的协方差矩阵 $\boldsymbol{G}_{\mathrm{t}}$(下角 t 代表 total)为

$$\boldsymbol{G}_{\mathrm{t}} = \frac{1}{M}\sum_{j=1}^{M}(\boldsymbol{A}_j - E(\boldsymbol{A}))^{\mathrm{T}}(\boldsymbol{A}_j - E(\boldsymbol{A})) \tag{5}$$

$\boldsymbol{G}_{\mathrm{t}}$ 又称为图像总方差矩阵,由于从定义很容易证明 $\boldsymbol{G}_{\mathrm{t}}$ 是非负的,而且可以直接从图像训练样本得到,因此(式(2))可改写为

$$J(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{G}_{\mathrm{t}}\boldsymbol{x} \tag{6}$$

最优映射轴 $\boldsymbol{x}_{\mathrm{opt}}$ 是最大化 $J(\boldsymbol{x})$ 的单位向量,也就是 $\boldsymbol{G}_{\mathrm{t}}$ 对应最大特征值的向量。一般来说,一个最优轴向是不够的,通常需要选择映射轴向的一个子集,即最大化 $J(\boldsymbol{x})$ 的一组正交向量 $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_d$。实际上,$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_d$ 就是 $\boldsymbol{G}_{\mathrm{t}}$ 对应前 d 个最大特征值的特征向量。有了最优映射轴,就可以通过将给定图像样本在最优映射轴上投影,来得到图像的特征。

2.2 SVM 原理

SVM(support vector machine)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。SVM 使用结构风险最小化(structural risk minimization, SRM)准则来构造决策超平面,以便使每一类数据之间的分类间隔(margin)最大[9]。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,如果有 $0 \leqslant \eta \leqslant 1$,则满足如下关系

$$R \leqslant R_{emp} + \sqrt{\frac{h(\log(2n/h)+1) - \log(\eta/4)}{n}} \quad (7)$$

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。而

$$\sqrt{\frac{h(\log(2n/h)+1) - \log(\eta/4)}{n}} \quad (8)$$

则叫做“VC 置信值”,n 是训练样本个数,h 是学习机的 VC 维(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 VC 置信值达到一个折中,最终使实际风险 R 变小。

由于视频字幕出现的非确定性和多样性,因此即使提取的特征良好,也不能保证视频字幕和非视频字幕的线性可分,即视频字幕与非视频字幕是非线性可分的。SVM 首先通过核函数把训练样本中的低维数据映射到高维特征空间,然后在高维特征空间构造一个最佳分类平面来进行分类。由于构造的核函数满足 Mercer 条件[10],因此在训练中只需考虑核函数 K,而不必知道低维向高维的映射函数 Φ。在实验中,SVM 只需指定特定核函数,而无需指定原始图像特征到高维特征的映射函数,是一种适合的分类器。

3　基于 2DPCA 与 SVM 的视频字幕验证方法

3.1　特征提取

利用最优映射向量提取图像特征,对于一个给定的图像样本 A,

$$y_k = Ax_k \quad k = 1,2,\cdots,d \quad (9)$$

将这样得到的一组映射特征向量 $y_1,y_2,\cdots,y_d$,称作图像样本的主成分。需要指出的是 2 维主成分分析的每一个主成分都是矢量,而 1 维主成分分析的是标量。

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_d \end{bmatrix} = \begin{bmatrix} Ax_1 \\ Ax_2 \\ \vdots \\ Ax_d \end{bmatrix} \quad (10)$$

$Y=[y_1,y_2,\cdots,y_d]$为图像 A 的特征,用作后续分类器的输入。

3.2　SVM 分类

在实验中,输入数据为每个图像子块按 3.1 小节的方法提取的特征向量和对这个子块是否为字幕的标注(+1 或 −1)。SVM 分类学习机有以下两个功能:一是用非线性映射把输入数据从原始低维空间映射到高维特征空间;二是计算特征向量和支持向量的内积。在实际处理中,这两步是通过核函数一步来实现的,核函数满足

$$K(x,y) = \Phi(x) \cdot \Phi(y)$$

其中,K,Φ 和 · 分别代表核函数、高维非线性映射和内积。输出层将分类结果输出,即输出识别判断结果。

SVM 中研究最多的核函数主要有以下 3 类:多项式、径向基函数(radial base function, RBF)和多层 Sigmoidal 神经网络。实验中使用的是 RBF 核函数,其形式为

$$K(x,y) = \exp\left\{-\frac{|x-y|^2}{\sigma^2}\right\} \quad (11)$$

在实验中将字幕子块定义为 +1,而将非字幕子块定义为 −1,对每个输入块,如果输出为正,则该块被判定为字幕块;否则,为非字幕块。

3.3　算法步骤

(1)训练过程

①将训练图像集标注为字符或非字符;

②用式(5)计算训练集图像总方差阵 G_t;

③用式(6)计算与前 d 个最大特征值对应的特征向量,并由这 d 个特征向量构成子空间,即投影阵;

④将投影阵代入式(9),来计算训练集图像特征;

⑤用字符图像特征及非字符图像特征训练 SVM,来得到训练好的 SVM。

(2)判别过程

①对视频图像做尺寸归一化,通过窗口滑动得到待判别的测试图像;

②据式(9)提取待判别字幕图像特征;

③用训练好的 SVM 判定待判别测试图像是字符或非字符;

④对同属一个候选字幕块的多个测试图像的判别结果进行投票,最终判定该候选字幕块是否字幕。

4　实验结果

为验证该算法的效果,本文在 MATLAB6.5 环境下实现了该算法,并在主频 1.4MHz、内存 256MB 的微机上进行了大量实际测试。实验采用的训练数据

库由 2 000 幅中文字符图像、2 000 幅英文字符图像以及 4 000 幅非字符图像组成，图像分辨率均为 32×28。其直接来源于含字幕的数字视频广播图像中，包含不同字体，不同笔画宽度，不同背景污染程度等因素。字符图像并不就是完整的中文字符或英文字符，而是具有字符特征的图像块（如图 1 所示）。

图 1 训练集样本图像示例

Fig. 1 Samples of train set images

从实际 PAL 制式的电视视频节目（包括新闻、广告、故事片等）中随机地选取了 100 幅图像，图像尺寸为 576×720，其中包括 30 幅含中文字幕的图像、30 幅含英文字幕的图像及 40 幅随机包含中文或英文字幕的图像。利用文献[11]中的字幕检测算法，适当降低阈值，对以上图像进行检测。将其检测得到的虚警非字幕图像块和字幕图像块，作为本文检测算法的实验数据。经人工标注后，从 30 幅含中文字幕的图像中检测得到 37 个字幕块，12 个虚警非字幕块；从 30 幅含英文字幕的图像中检测得到 36 个字幕块，17 个虚警非字幕块；从 40 幅随机包含中文或英文字幕的图像中检测得到 52 个字幕块，21 个虚警非字幕块。通过高度归一化，并用 32×28 的窗口进行滑动抽取，则得到的 3 个测试集合如表 1 所示。

表 1 测试数据集属性

Tab. 1 The attribute of test dataset

中文		英文		中英文混合	
字幕块数	非字幕块数	字幕块数	非字幕块数	字幕块数	非字幕块数
312	73	289	103	337	138

为验证训练样本数及投影矢量个数对检测算法的影响，随机在 2 000 幅中文字符图像及 4 000 幅非字符图像中各抽取 M 个图像作为训练样本集（M = 50，100，300，600，1 000），分别选取 1～10 个不等的投影矢量个数，用本文算法对中文测试集进行验证，判断正确的字幕块个数如表 2 所示。

从表 2 中可以看出，在训练样本各取 600 时，检测算法可整体得到相对较大的正确字幕块数，因此在后续的实验中，训练样本将各取 600 进行比较。

表 2 不同训练样本及不同投影矢量个数的区别

Tab. 2 Compare of varied number of train samples and projecting vectors

训练样本数	不同投影矢量数判断正确的字幕块个数									
	1	2	3	4	5	6	7	8	9	10
50	283	258	251	245	247	240	236	239	240	239
100	308	281	289	282	265	271	258	267	262	261
300	339	325	329	326	325	321	315	316	313	311
600	351	345	348	343	346	343	343	342	340	339
1 000	344	334	335	334	334	332	331	327	324	320

为比较算法对中文、英文及中英文混合字幕的验证效果，实验分 3 组训练及测试样本。中文组的训练集是随机在 2 000 幅中文字符图像及 4 000 幅非字符图像中各抽取 600 个图像组成，英文组的训练集是随机在 2 000 幅英文字符图像及 4 000 幅非字符图像中各抽取 600 个图像组成，中英文混合组的训练集是随机在 2 000 幅中文、英文字符图像中各取 300 个图像及 4 000 幅非字符图像中各抽取 600 个图像组成。测试集由前述的 3 个测试集组成。其验证结果如表 3 所示。

表 3 中文、英文及中英文混合字幕的验证比较

Tab. 3 Compare of validating text in Chinese, English and mix of two

样本	字幕块总数 T_t	非字幕块总数 B_t	正确验证的字幕块个数 $T_{correct}$	正确验证的非字幕块数 $B_{correct}$	准确率 R(%)
中文	37	12	37	10	95.9
英文	36	17	35	12	88.7
中英文	52	21	49	19	93.2

表 3 中，$R=\dfrac{T_{correct}+B_{correct}}{T_t+B_t}\times 100\%$。

从表 3 可以看出，本算法对中文的检测结果明显好于英文，通过对检测结果进行分析发现，因中文笔画复杂，字符特征明显，其在图像复杂背景下，比笔画相对简单的英文更易于与背景相区分。未正确辨别的字幕块主要是因为图像中字幕太小或字符很少，从而被误认为是噪声块而清除。误判的主要原因是由于一部分图像子块体现出了字幕块特征，如比较明显的横竖纹理。因此，英文查全率及准确率均明显较低。

为比较 PCA、ICA（Independent component analysis）

与2DPCA 等不同特征提取方法的效果,随机在2 000幅中文字符图像及4 000 幅非字符图像中各抽取600 个图像组成训练集。PCA 及 ICA 方法可参照文献[12 ~ 14],分别利用这些方法提取 8 个投影矢量。实验结果如表 4 所示。

表 4　PCA、ICA 与 2DPCA 等不同特征提取方法的效果对比

Tab. 4　Compare of methods based on PCA, ICA and 2DPCA

	不同特征提取方法		
	PCA 特征	ICA 特征	2DPCA 特征
准确率(%)	82	92.0	93.2
特征投影矢量提取耗时(s)	34	62	10

由表 4 可以看出,在所有特征提取方法中,2DPCA 方法与 ICA 方法的效果最好,而 2DPCA 方法与 ICA 方法在准确率相当的情况下,2DPCA 方法的特征矢量提取的计算耗时(不包括文件读取耗时)大大低于 ICA 方法。

上述理论分析和实验结果已展示了 2DPCA 在视频字幕验证中的能力,不仅从方法复杂性和结果的准确性综合来看,该方法明显优于传统方法,而且从大量的实验结果来看,2DPCA 方法在图像背景复杂、图像分辨率低以及字幕字体、大小、颜色多变这些传统检测提取方法或多或少都存在困难的条件下,是一种有效的视频字幕验证方法。

5　结　论

在视频字幕的检测提取过程中,由于字幕在字体、大小和排列上多变,以及受视频复杂背景、图像分辨率低的影响,因此无法避免检测虚警。本文提出使用 2DPCA 进行图像特征提取,并采用 SVM 学习分类的方法进行视频字幕验证,在样本不是很多的情况下实现了较高精度、快速的视频字幕验证。同时,在保证查全率的基础上,提高了检测准确率。相对于现存的基于启发式经验规则的视频字幕验证方法,本文方法适应能力强。大量实验结果表明,该方法是有效的。本文在训练样本数量及投影矢量个数的选取上带有实验性质,而是否这两个参数的选取有必然的规则,还需要进一步研究,这是下一步的工作方向。

参考文献(References)

1　Rainer Lienhart. Video OCR: a survey and practitioner's guide[R]. 95052-8119, Intel Corporation, Microprocessor Research Labs, Santa Clara, CA, USA, 2003.

2　LU Han-qing, KONG Wei-xin, LIAO Ming, *et al.* A review of content-based parsing and retrieving for image and video[J]. ACTA Automatic SINICA, 2001, **27**(1): 56 ~ 59. [卢汉清,孔维新,廖明等. 基于内容的视频信号与图像库检索中的图像技术[J]. 自动化学报, 2001, **27**(1): 56 ~ 59.]

3　Rainer Lienhart, Axel Wernicke. Localizing and segm-enting text in images and videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2002, **12**(4): 256 ~ 268.

4　Sato T, Kanade T, Hughes E, *et al.* Video OCR: indexing digital news libraries by recognition of superimposed caption[J]. Multimedia System, 1999, **7**(5): 385 ~ 395.

5　Li H, Doermann D, Kia O. Automatic text detection and tracking in digital video[J]. IEEE Transcations on Image Processing, 2000, **9**(1):147 ~ 156.

6　Li H, Doermann D. Text enhancement in digital video using multiple frame integration[A]. In: ACM Multimedia[C], Orlando, Florida, USA, 1999: 65 ~ 71.

7　Yang Jian, Yang Jing-yu. From image vector to matrix: a straightforward image projection technique-IMPCA vs. PCA[J]. Pattern Recognition, 2002, **35**(9):1997 ~ 1999.

8　Yang Jian, Zhang David, Frangi Alejandro F, *et al.* Two-dimensional PCA: A new approach to appearance-based face representation and recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2004, **26**(1): 131 ~ 137.

9　Vapnik V. Statistical learning theory[M]. New York: Springer-Verlag, 1998.

10　Zhang Li, Zhou Wei-da, Jiao Li-cheng. A new kind of support vector machine kernel[J]. Journal of Software, 2002, **13**(4): 713 ~ 718. [张莉,周伟达,焦李成. 一类新的支撑矢量机核[J]. 软件学报, 2002, **13**(4): 713 ~ 718.]

11　Wang Yong, Yan Ji-kun, Zheng Hui. An adaptive method for detecting and location text in video frame[J]. Computer Applications, 2004, **24**(1): 134 ~ 135. [王勇,燕继坤,郑辉. 一种自适应的视频帧中字幕检测定位方法[J]. 计算机应用, 2004, **24**(1): 134 ~ 135.]

12　Hyvarinen A, Oja E. Independent component analysis: a tutorial [J]. Neural Networks, 2000, **13**(4-5): 411 ~ 430.

13　Yuen P C, Lai J H. Face representation using independent component analysis[J]. Pattern Recognition, 2001, **34**(3): 545 ~ 553.

14　Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis[J]. IEEE Transactions on Neural Networks, 1999, **10**(3): 626 ~ 634.