低质量汉字的分块搜索两级识别法

刘 毅[1,2)]，毛震东[1,2)]，张冬明[1)]，张勇东[1)]，林守勋[1)]

[1)](中国科学院计算技术研究所前瞻研究实验室 北京 100190)

[2)](中国科学院研究生院 北京 100049)

(liuyi@ict.ac.cn)

摘 要：由于汉字笔画复杂，从视频中提取的汉字图像质量往往较差，采用传统光学字符识别(OCR)的结果不理想.为了解决低质量汉字图像的识别问题，提出一种基于分块搜索的两级识别方法.首先建立汉字图像的分块结构并模仿低质量汉字生成训练集，然后对训练集中各分块图像应用主成分分析提取特征并建立索引.待识别图像应用分块搜索和投票的方式从索引中获取候选汉字集合(一级识别)，再根据投票结果的显著性辅以全局结构特征匹配识别汉字(二级识别).实验结果证明，该方法对于低质量汉字图像比普通的 OCR 方法具有更高的识别率.

关键词：光学字符识别；低质量汉字识别；分块搜索；多级识别

中图法分类号：TP391

A Two-Stage Scheme Based on Block Search for Low-Quality Chinese Character

Liu Yi[1,2)], Mao Zhendong[1,2)], Zhang Dongming[1)], Zhang Yongdong[1)], and Lin Shouxun[1)]

[1)](*Center for Advanced Computing Research, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100190)

[2)](*Graduate University of Chinese Academy of Sciences, Beijing* 100049)

Abstract: Due to the complex character strokes, the quality of video-extracted Chinese character images is often poor, for which traditional optical character recognition (OCR) could not get desired results. To address this problem, this paper presents a two-stage scheme for low-quality Chinese character recognition based on block search. The block structure of the Chinese character image is built, along with a training set, imitating low-quality Chinese characters. And an index is generated after extracting the features from each block of the training set by applying principle component analysis. This scheme retrieves candidate character set from the index by block search and voting (1st stage), and then recognizes the character according to the salience of voting result assisted with global structural feature match (2nd stage). The experimental results have demonstrated that this scheme has better recognition rate for low-quality Chinese character images compared with traditional OCR method.

Key words: optical character recognition; low-quality Chinese character recognition; block search; multi-stage recognition

视频中内嵌文字的识别是当前计算机视觉领域的研究热点之一，它在图像视频检索、地理信息标注以及城市环境下的机器人导航等方面有着广泛的应用前景.目前，相关的研究多集中于文本的检测和

收稿日期：2011-05-05；修回日期：2011-08-26.**基金项目**：国家“九七三”重点基础研究发展计划项目(2007CB311100)；国家“八六三”高技术研究发展计划(2009AA01A403)；国家自然科学基金(60802028)；北京市科技新星计划项目(2007B071)；北京市教育委员会共建项目专项.
刘 毅(1985—)，男，博士研究生，主要研究方向为图像内容分析；**毛震东**(1984—)，男，博士研究生，主要研究方向为多媒体检索；**张冬明**(1977—)，男，博士，副研究员，硕士生导师，CCF 会员，主要研究方向为多媒体计算；**张勇东**(1973—)，男，博士，研究员，博士生导师，CCF 高级会员，主要研究方向为视频编解码、视频分析；**林守勋**(1948—)，男，博士，研究员，博士生导师，CCF 高级会员，主要研究方向为多媒体技术及应用系统.

提取[1-2]，其识别过程则由光学字符识别（optical character recognition，OCR）软件完成. 然而，从视频中提取的文本图像大都含有错误分割的背景，导致最终生成的单个文字图像质量较差，经常伴随如下情形：

1）部分背景掺杂在图像中；

2）文字笔画缺失；

3）文字的笔画不规则或不均匀.

本文把具有上述特征之一的文字图像统称为低质量的文字图像. 虽然低质量英文图像在现代 OCR 软件中已能获取 95%以上的识别率[3]，甚至已应用到了验证码的识别，但是低质量的汉字图像的识别率却相对较低.

汉字 OCR 主要分为预处理、特征提取、分类识别和后处理 4 个步骤. 传统方法侧重于特征提取，研究如何提取辨别能力强的特征；现代方法则基于传统方法的特征实现多级分类和多分类器组合，借以提高非清晰汉字的识别率. 手写体汉字识别主要针对清晰图像，研究重点在于特征提取和后处理，而低质量的手写体汉字识别难度很大，尚未成为研究主线.

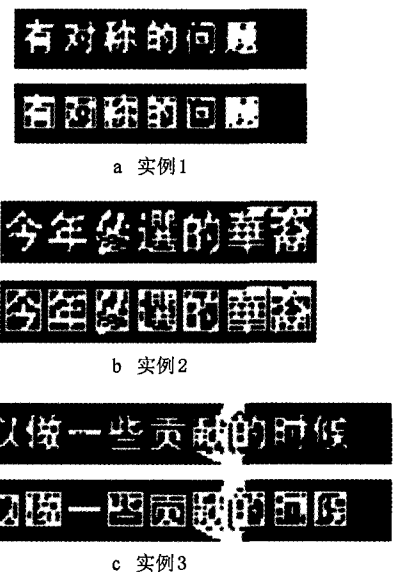
传统方法中，汉字图像的低质量部分会引起全局特征的变化，导致基于全局特征的方法[4]识别率很低. 而现代方法，如文献[5]中的多级分类法的粗分类将多个局部特征合并为一个特征向量，所以局部变化会反映到该特征向量上，一旦汉字图像多个局部产生变化或者某个局部的变化很大，粗分类准确率就会降低，从而影响后续识别. 文献[6]从信息熵的角度分析了多分类器集成的重要性，提出一种用于汉字识别的多分类器集成的方法，但是其特征提取步骤依然沿用了全局特征和局部特征融合的特征向量，仅适用于局部变化轻微的文字图像. 文献[7]利用改进的局域二值模式（local binary patterns，LBP）识别低质量的车牌汉字，由于车牌汉字的类别极少，该方法效果很好，但是 LBP 的特征描述能力尚不足以覆盖一般的汉字识别应用.

本文针对从视频中提取的低质量汉字图像的特点，突破传统单纯应用全局特征或局部融合特征的模式，提出一种基于分块搜索的两级识别方法. 该方法将汉字图像局部的缺陷约束在各个分块内，避免扩散到全局特征，然后各个分块对索引字库中结构相似的汉字进行投票，表决出备选汉字集合. 如果某汉字具有显著的票数，则确定结果为该汉字，否则在备选汉字集合中采用全局结构特征进行更精确的匹配，确定出最终结果.

1　分块结构及特征提取

1.1　分块结构的构建

在介绍分块结构之前，先给出按照文献[8]方法从视频中提取并分割出单字的汉字文本图像实例，如图 1 所示. 其中，每组图像的第一行是字幕提取的结果，第二行是单字分割的结果. 3 类“低质量”的情形在分割单字中均有出现：图 1b，1c 中存在背景掺杂，图 1c 中存在笔画缺失，图 1a～1c 中均存在笔画不均匀.



a 实例1

b 实例2

c 实例3

图 1　文本图像实例

考虑到视频中的文字大多数都比较小，例如视频字幕，本文采用 32×32 像素尺寸来规范化待识别的汉字图像，一方面减少了由尺寸大小带来的运算量，另一方面避免了因图像放大倍数过大造成失真.

我们通过大量实验发现，汉字图像中质量较差的位置主要在边界，因此本文沿边界将汉字图像分割为左上、右上、左下、右下以及居中 5 个等面积、部分重叠的正方形分块，每个分块由 16×16 个像素组成，分块结构如图 2 所示.

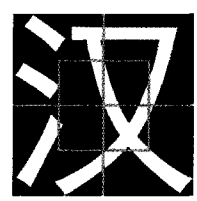


图 2　分块结构示意图

1.2 分块特征提取

在分块结构中，每个分块的尺寸较小，所含笔画数目少，有多种方式可以有效地提取出特征，本文选用较为鲁棒的主成分分析提取图像的统计特征.

主成分分析是统计理论中的基本方法之一，它于 20 世纪 80 年代末被引入到图像表示领域，并因为理论成熟而获得广泛的应用. 其基本思想就是试图找到一组最能反映某分布特点的基向量，用这组基向量去表述子空间. 其数学形式为

$$\boldsymbol{y}=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{x};$$

其中，$\boldsymbol{x}\in\mathbb{R}^n$，$\boldsymbol{y}\in\mathbb{R}^m$，$n>m$，$\boldsymbol{W}$ 代表投影矩阵. 即用线性变换把高维空间中的向量 $\boldsymbol{x}$ 投影到子空间中的特征向量 $\boldsymbol{y}$，从而实现降维和提取特征的目的.

令 $\boldsymbol{\mu}$ 表示 n 维向量 $\boldsymbol{x}_i(i=1,2,\cdots,l)$ 的均值，这些向量的协方差矩阵(代表了向量集的离散状况) $\boldsymbol{S}_t$ 可以表示为

$$\boldsymbol{S}_t=\frac{1}{l}\sum_{i=1}^{l}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathrm{T}};$$

主成分分析投影矩阵的列向量则由 $\boldsymbol{S}_t$ 对应于前 k 个最大特征值 $\lambda_i(i=1,2,\cdots,k)$ 的特征向量组成.

根据上述结论，给定合适的 m 值(见第 5.1 节)对训练集按照分块结构进行主成分分析，即可获得 5 个相互独立的投影矩阵，从而实现分块特征的提取. 其中，高维空间中的向量 $\boldsymbol{x}$ 表示汉字图像一个分块的像素值按行排列形成的一维向量.

2 训练集建立

经文本提取获得的汉字图像极可能质量较差，由于掌握了汉字图像低质量的常见情形，通过人工制造一些包含更多信息的训练数据可以较好地模拟实际视频中的汉字，获得更高的召回率.

本文选择了视频中常见的黑体字体，按照每个汉字生成 5 个样本(1 个原始样本，4 个带噪声样本)的方法建立训练集. 该训练集基本覆盖了低质量图像中笔画粗、细、不均匀以及背景掺杂等特征，如图 3 所示.

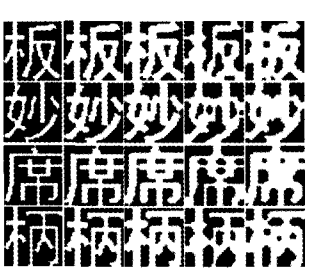


图 3 训练样本

训练样本生成算法中，“降低质量”操作是用线性插值将 32×32 尺寸的原图像缩小到某一尺寸，再用最邻近插值恢复到原尺寸，最后以灰度值 128 为界把恢复后的图像二值化. 为简洁起见，该算法中此操作仅列举缩小尺寸参数，步骤如下：

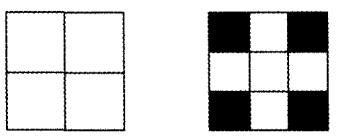
Step1. 直接使用黑体字的原型作为一个样本.

Step2. 如图 4a 所示，用 2×2 尺寸、锚点位于(1,1)的算子对图像进行膨胀运算，得到一个样本.

Step3. 如图 4b 所示，用 3×3 尺寸、锚点位于(1,1)的算子对图像进行膨胀运算，得到一个样本.

Step4. 在 Step2 的基础上降低质量，其中缩小尺寸参数为 16×16.

Step5. 在 Step3 的基础上降低质量，其中缩小尺寸参数为 13×13.



a 2×2膨胀算子 b 3×3膨胀算子

图 4 膨胀算子

经过上述算法即可构造出完整的训练集，图 3 中每列图像对应该算法每个步骤的结果.

汉字图像的识别通常在某一固定尺度下进行，规范化的过程可能会降低图像质量，常见的情形有笔画不规则和不均匀. 本文通过“降低质量”操作来模拟尺度变换造成的影响，从而将先验知识嵌入到训练数据中. 同时，当该操作的缩小参数足够小时，恢复到原始尺寸的图像会产生背景掺杂的效果，生成的训练集就能更好地模拟低质量的汉字图像.

3 特征索引

本文使用的识别方法基于分块结构，每个分块都要搜索与之结构相似的一组结果，分块的搜索速度将直接影响汉字识别速度. 普通的 K 最近邻(k-nearest neighbor, KNN)是线性搜索算法，时间复杂度为 $O(n)$，这在实时 OCR 应用中无疑会成为效率瓶颈. 另一方面，单个分块对整个汉字的识别并非起决定作用，并且个别分块还可能因质量低下产生误识. 所以，精确的近邻搜索不是必须的，只要保证有一个分块的搜索结果含有目标汉字，就有正确识别的可能.

LSH(locality sensitive hashing)[9]是目前近似 KNN 索引算法中的热门技术之一，它将高维空间中的向量按照多个哈希结构有序组织，通过散列实

现高效搜索. 然而，LSH 查询时的正确率需要哈希表的数量来保证，所以这是典型的以空间换时间的方法. 为了弥补传统 LSH 算法中内存消耗过大的缺陷，文献[10]中提出了优化的 LSH 算法——MP-LSH(multi-probe LSH)，该方法通过探测邻近箱格中的向量增大在一次散列过程中找到目标向量的概率，从而锐减了哈希表的数量，提高了内存的使用率.

综上分析，本文选用了 MP-LSH 算法对每个分块建立索引，K 值的选取可以参考第 5.1 节. 当然，在实时性要求不高的场合，采用线性 KNN 算法能够获得更高的识别率.

4　汉字识别

根据上述的分块结构、特征提取以及索引方式即可完成字库的训练，汉字图像的识别同样基于分块结构进行.

得益于本文引入的分块结构，低质量的汉字图像可以通过结构完善的分块进行逆推、各个分块投票表决来确定输入图像的汉字. 这就类似于 5 个基于分块的弱分类器组合构成的一个强分类器，用于一级识别.

如果票选结果不够显著，即没有任何汉字拥有足够多的票数，则在票选出的汉字集合中查找与输入汉字图像全局结构特征最匹配的汉字作为二级识别结果.

4.1　一级识别——投票表决

投票表决按照如下步骤进行：

Step1. 输入汉字图像依据分块独立搜索 K 个近邻，每个分块的 K 近邻滤掉重复汉字后剩下 p 个结果.

Step2. 每个分块给相应的 p 个搜索结果各投一票.

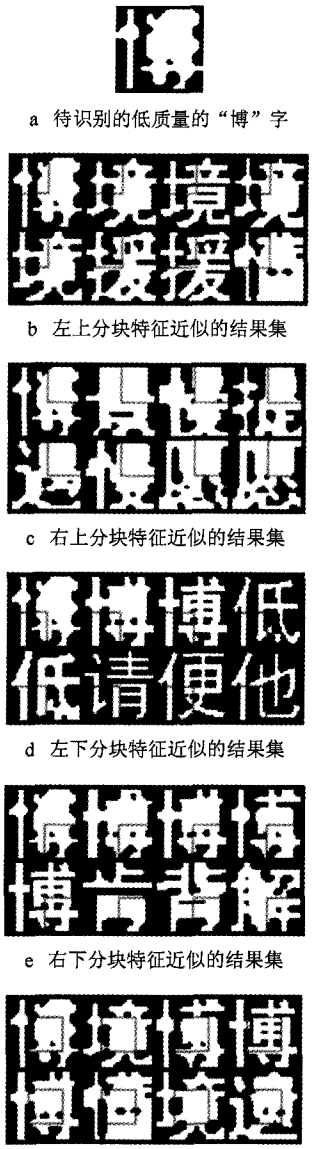
Step3. 最终的表决结果分 3 种情况考虑(票数阈值 $t=4$)：

Step3.1. 某个汉字获得最多的票数 n，$n>t$，那么识别结果即为这个汉字；

Step3.2. 有多个汉字获得最多的票数 n，$n>t$，则在这些汉字生成的所有样本中按全局结构特征进行匹配，确定最终结果；

Step3.3. 如果最多票数 $n\leqslant t$，则合并各分块的搜索结果，再按全局结构特征进行匹配，确定最终结果.

图 5 所示为分块搜索结果示意图. 可以看到，图 5 a 所展示输入汉字图像的笔画细节基本已经丢失，但根据该汉字的偏旁部首以及其他的清晰部分仍然能够正确地辨认. 基于分块结构的投票表决就是由这个思路演变而来的. 下面以实际提取的质量较差的“博”字为例，对投票表决方法进行说明.



a 待识别的低质量的“博”字

b 左上分块特征近似的结果集

c 右上分块特征近似的结果集

d 左下分块特征近似的结果集

e 右下分块特征近似的结果集

f 居中分块特征近似的结果集

图 5　分块搜索结果示意图

图 5 中给出了输入的低质量汉字图像以及对每个分块搜索出的 7 近邻，为了便于比较，每个结果集中展示的第一幅图像均为输入图像. 可以看出，利用分块特征获得的搜索结果集在该分块内都有结构相似的特点，并且对笔画腐蚀、丢失和扭曲具有一定的鲁棒性. 其中，图 5 b，5 c 2 组结果相应分块内笔画质量低下，前 7 个搜索结果均未出现目标汉字，但是，得益于剩下 3 个分块内的图像都能很好地描述输入汉字的结构特征，利用这 3 个分块都搜索出了目标汉字. 这样，在 7 近邻的投票表决结果中，“博”字总共获得 3 票，多于“境”和“懂”字的 2 票，但总票数小于

阈值 t,满足 Step3.3,因此将示意图中搜索出的结果集合并,按全局结构特征进行匹配,最后确定出识别结果"博"字.

在实际实验中,K 值的选取往往比 7 大.选取一个合适的 K 值后,"博"字在左上分块搜索结果中排第 9,在右上分块搜索结果中排第 8,因此该汉字的累计票数为 5,符合 Step3.1,这样,仅利用投票表决方式就可以识别出正确结果.

4.2 二级识别——全局结构特征匹配

如果投票表决不能确定待识别的汉字,则需要在一级识别筛选出的备选汉字集合内进行全局结构特征匹配.

本文选用文献[4]提出的部分特征,如表 1 所示.

表 1 全局结构特征

特征名称	维数
水平、竖直投影直方图	32
灰度值跃变总次数	2
笔画强度	16
两级外围特征	64
局部方向平均长度	64
局部方向像素数	64
笔画比例	32
子向量灰度值跃变比例	16
总特征维数	290

基于全局结构特征匹配速度虽然较慢,但每个分块的票数固定,候选汉字集合的大小也就固定了,所以匹配速度不会受到字库大小的影响.同时,票选出的汉字数量较少,如果票选集合中包含目标汉字,则在这个小型汉字集合中利用全局特征匹配更容易获得正确结果.

5 实验及结果分析

本文建立了一个 1 565 字的字库,该字库囊括了互联网视频中的常见汉字,并基于该字库进行了 2 个实验:实验 1 通过比较实验结果确定出合理的参数;实验 2 在第 5.1 节给定参数的基础上对比测试了汉字图像的识别率.

5.1 m 值和 K 值的选择

第 1.2 节和第 3 节引入了 2 个参数:主成分分析特征提取的维数 m 和搜索近邻数 K.对这 2 个参数的组合实验结果如图 6 所示,当 m 值比较大时,分块特征具有更准确的结构表达能力,于是,在较少的 KNN 中即可查找到目标结果;而当 m 值较小时,分块特征表达能力减弱,为了在近邻中查找到目标结果,就必须相应地增大 K 值.当 m 值和 K 值同时增大时,识别率会先降低后增加,并在 $(m,K)=(50,45)$ 附近重新恢复到较高值,但在 m 值和 K 值增加的同时计算量也会增加,所以针对相等的识别率,优先选择较小的取值组合.

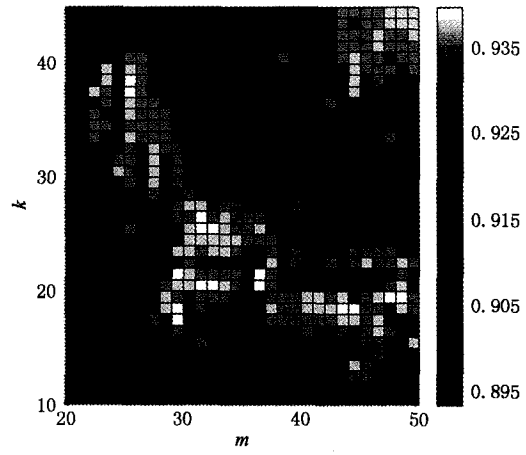


图 6 m 值和 K 值组合的实验结果

综上,本文选取 $(m,K)=(33,25)$,即在保证较高识别率的同时,选取适中的 m 值和 K 值.

5.2 识别率测试

本文根据文献[8]方法从 68 个在线视频片段中提取并筛选出清晰汉字图像 5 000 个,以及人眼能够辨识的低质量汉字图像 7 000 个作为测试集,部分低质量测试数据如图 7 所示;然后利用第 5.1 节确定的 m 值和 K 值与 OCR 软件——汉王文豪 7600 进行对比实验,结果如表 2 所示.其中,"投票候选正识率"指经投票表决生成的候选汉字集合中包含正确结果的比率.

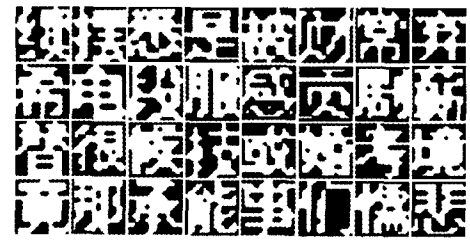


图 7 低质量汉字图像样例

表 2　汉字识别率　　%

图像质量	本文方法		汉王文豪 7600
	投票候选正识率	总识别率	
清晰	100	99.32	99.40
低质量	98.87	91.15	86.41

汉王文豪 7600 是目前最优秀的汉字 OCR 软件之一，虽然它对清晰汉字图像的识别率高达 99.40%，但是对低质量汉字图像仅有 86.41%. 与之相比，本文方法对清晰汉字图像的识别率稍差，但也基本持平，而且对低质量汉字图像取得了 91.15%的识别率，提高了近 5%. 由此证明了本文方法识别低质量汉字图像的有效性.

6　结语与展望

本文提出了一种融合结构和统计信息的分块特征，即把汉字图像划分为部分重叠的等面积分块并使用主成分分析对其进行降维，再从字库中根据分块提取全部特征，并利用快速的近似 KNN 算法对分块特征进行索引，最后采用分块投票表决并辅以全局结构特征匹配的方式确定输入汉字. 实验结果表明，对于低质量汉字的识别本文方法优于当前优秀的 OCR 软件.

我们将在今后的工作中着眼文本检测算法，利用视频帧间相似性等先验知识优化汉字图像的生成质量，这也有利于汉字识别率的提高.

参考文献(References)：

[1] Anthimopoulos M, Gatos B, Pratikakis I. A two-stage scheme for text detection in video images [J]. Image and Vision Computing, 2010, 28(9): 1413-1426

[2] Fu Hui, Liu Xiabi, Jia Yunde. Edge-pixels clustering for text area extraction [J]. Journal of Computer-Aided Design & Computer Graphics, 2006, 18(5): 729-734 (in Chinese)
(付　慧，刘峡壁，贾云得. 用于文本区域提取的边缘像素聚类方法[J]. 计算机辅助设计与图形学学报，2006，18(5)：729-734)

[3] Yan J, Ahmad A S E. A low-cost attack on a Microsoft captcha [C] //Proceedings of the 15th ACM Conference on Computer and Communications Security. New York: ACM Press, 2008: 543-554

[4] Romero R D, Touretzky D S, Thibadeau R H. Optical Chinese character recognition using probabilistic neural networks [J]. Pattern Recognition, 1997, 30(8): 1279-1292

[5] Gao Tao, Li Mingjing, Li Zhifeng. A multi-font huge-set character recognition system [J]. Journal of Chinese Information Processing, 2000, 14(2): 31-36 (in Chinese)
(高　涛，李明敬，李志峰. 一种多字体特大字符集字符识别系统[J]. 中文信息学报，2000，14(2)：31-36)

[6] Guo Hong, Ding Xiaoqing, Guo Fanxia, *et al*. New method of combining multiple classifiers for Chinese character recognition [J]. Journal of Tsinghua University: Science and Technology, 1997, 37(10): 91-94 (in Chinese)
(郭　宏，丁晓青，郭繁夏，等. 汉字识别多分类器集成的新方法[J]. 清华大学学报：自然科学版，1997，37(10)：91-94)

[7] Gao Yanyu, Yang Yang. A survey of off-line handwritten Chinese character recognition [J]. Computer Engineering and Applications, 2004, 40(7): 74-77 (in Chinese)
(高彦宇，杨　扬. 脱机手写体汉字识别研究综述[J]. 计算机工程与应用，2004，40(7)：74-77)

[8] Song Yan, Liu Anan, Zhang Yongdong, *et al*. Video text extraction method based on clustering [J]. Journal on Communications, 2009, 30(2): 136-140 (in Chinese)
(宋　砚，刘安安，张勇东，等. 基于聚类的视频字幕提取方法[J]. 通信学报，2009，30(2)：136-140)

[9] Andoni A, Indyk P. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions [C] // Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science. Washington D C: IEEE Computer Society Press, 2006: 117-122

[10] Joly A, Buisson O. A posteriori multi-probe locality sensitive hashing [C] //Proceedings of the 16th ACM International Conference on Multimedia. New York: ACM Press, 2008: 209-218