JIG
JOURNAL OF IMAGE AND GRAPHICS

中国图象图形学报

主办：中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

2014
05
VOL.19

ISSN1006-8961
CN11-3758/TB

流体模拟 *P781*

中国图象图形学报

刊名题字：宋健 | 月刊（1996年创刊）

第19卷第5期（总第217期）
2014年5月16日

中国精品科技期刊
中国国际影响力优秀学术期刊
中国科技核心期刊

主管单位 中国科学院
主办单位 中国科学院遥感与数字地球研究所
中国图象图形学学会
北京应用物理与计算数学研究所

主　　编 李小文
编辑出版 《中国图象图形学报》编辑出版委员会
邮政信箱 北京9718信箱
邮　　编 100101
电子信箱 jig@irsa.ac.cn
电　　话 010-64807995
网　　址 www.cjig.cn

广告经营许可证 京朝工商广字第0361号
总 发 行 北京报刊发行局
订　　购 全国各地邮局
海外发行 中国国际图书贸易集团有限公司
（邮政信箱：北京399信箱　邮编：100048）
印刷装订 北京科信印刷有限公司

Journal of Image and Graphics
Title inscription: Song Jian | Monthly, Started in 1996

Superintended by Chinese Academy of Sciences
Sponsored by Institute of Remote Sensing and Digital Earth, CAS
China Society of Image and Graphics
Institute of Applied Physics and Computational Mathematics

Editor-in-Chief LI Xiaowen
Editor, Publisher Editorial and Publishing Board of Journal of Image and Graphics
P.O.Box 9718, Beijing, P.R.China
Zip code 100101
E-mail jig@irsa.ac.cn
Telephone 010-64807995
Website www.cjig.cn

Distributed by Beijing Bureau for Distribution of Newspapers and Journals
Domestic All Local Post Offices in China
Overseas China International Book Trading Corporation
(P.O.Box 399, Beijing 100048,P.R.China))
Printed by Beijing Kexin Printing Co., Ltd.

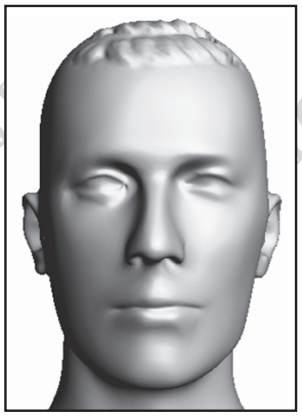
CN 11-3758/TB
ISSN 1006-8961
CODEN ZTTXFZ

国外发行代号 M1406
国内邮发代号 82-831
国内定价 50.00元

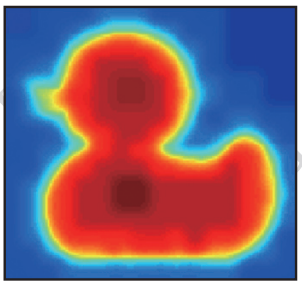
2014年5月 第19卷 第5期（总第217期）
中国图象图形学报
Zhongguo Tuxiang Tuxing Xuebao

目次

处理异常值的相机抖动模糊图像复原(第677页)



鲁棒的网格实时几何编辑算法(第764页)

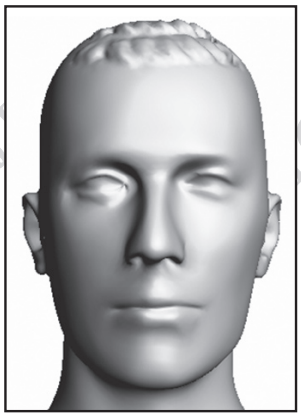


从表面重构的体数据实现3维网格剖分(第771页)

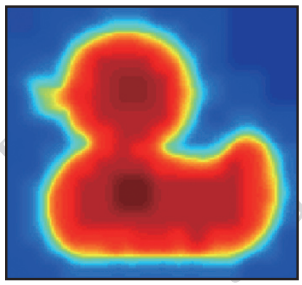
Volume 19, Number 5
Published May 16, 2014

CONTENTS
JOURNAL OF IMAGE AND GRAPHICS

Disposing of outliers in camera-shake blurred images restoration(P677)



Robust algorithm for real-time mesh geometric editing(P764)



Three-de mesh generation through volume data reconstructed from surface(P771)

E-mail: jig@irsa.ac.cn
Website: www.cjig.cn
Tel: 010-64807995

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号: TP391.4　　**文献标识码**: A　　**文章编号**: 1006-8961(2014)05-0683-10

论文引用格式: Li Q R, Lyu X Q, Li Z, Liu K. Logistic model for video caption enhancement[J]. Journal of Image and Graphics, 2014, 19(5): 683-692.
[李钦瑞, 吕学强, 李卓, 刘坤. Logistic 视频字幕增强模型[J]. 中国图象图形学报, 2014, 19(5): 683-692.][DOI: 10.11834/jig.20140505]

Logistic 视频字幕增强模型

李钦瑞[1], 吕学强[1], 李卓[1], 刘坤[2]

1. 北京信息科技大学网络文化与数字传播北京市重点实验室, 北京　100101;
2. 北京拓尔思信息技术股份有限公司, 北京　100101

摘　要: 目的　为提高复杂背景下的视频字幕在光学字符识别(OCR)中的识别率,需要对提取的视频字幕进行有效地字幕增强。首次将 Logistic 模型应用到视频字幕增强中,提出了基于 Logistic 模型的融合多帧信息的视频字幕增强方法。**方法**　对字幕进行检测与跟踪,将出现在连续多帧中的同一字幕片段进行对齐;通过分析字幕片段在多帧中信息,提出字幕背景在时域上的变化特征、背景和字幕文本的固有特征,并将 3 个特征进行量化与融合,构建适用于字幕增强的 Logistic 模型,实现对视频字幕的增强。**结果**　对含阴影或描边效果的特殊复杂背景字幕、普通复杂背景字幕、单一背景字幕分别进行实验,增强后的字幕在 OCR 软件中的识别正确率分别为 81.76%、97.13%、98.19%,与对比方法比较均有一定的提高。**结论**　实验结果表明,本文方法既可以降低字幕背景的复杂度,又可以提高字幕背景与文本的对比度,从而可以对复杂背景和单一背景下的视频字幕进行有效地增强。

关键词: 复杂背景;字幕增强;Logistic 模型;字幕检测与跟踪;时域特征

Logistic model for video caption enhancement

Li Qinrui[1], Lyu Xueqiang[1], Li Zhuo[1], Liu Kun[2]

1. *Beijing Key Laboratory of Internet Culture and Digital Dissemination Research, Beijing Information Science and Technology University, Beijing* 100101, *China*;
2. *Beijing TRS Information Technology Co., Ltd, Beijing* 100101, *China*

Abstract: Objective Video caption contains abundant information related to the video content. Recognizing text in images is the premise of making full use of this information. Although optical character recogonition (OCR) software recognition accuracy has been improved, the video captions with complex backgrounds still cannot be recognized well. Therefore, in order to improve the recognition accuracy, the extracted caption shall be enhanced which can reduce the complexity of caption background and improve the contrast between background and text. In this paper, we propose a method of fusing multi-frame information to realize caption enhancement based on the Logistic model. **Method** Logistic curve is a common form of an S-type curve, which either end or converge to a constant. By counting and analyzing distribution proportion of different pixel values in a single background caption, we establish a proper Logistic model whose output can be used as the enhanced caption's pixel values and their distribution proportion shall generally be kept consistent with the single background caption. According to the convergence of the Logistic model, the majority of pixel values can be assigned to 0 or 255, and a small quantity of gray points can be taken as transitions of black points and white points. Therefore, the enhanced caption

收稿日期: 2013-09-18; **修回日期**: 2013-11-07

基金项目: 国家自然科学基金项目(61171159, 61271304);北京市教委科技发展计划重点项目暨北京市自然科学基金 B 类重点项目(KZ201311232037);北京市属高等学校创新团队建设与教师职业发展计划项目(IDHT20130519);北京信息科技大学网络文化与数字传播北京市重点实验室开放课题(ICDD201303)

第一作者简介: 李钦瑞(1990—　),女,北京信息科技大学计算机应用技术专业在读硕士研究生,主要研究方向为中文与多媒体信息处理。E-mail: ginkgo_cool@126.com

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 5, May 2014

image not only keeps the continuity of the pixel values but also improves the contrast between background and text. Then we detect and track the video caption, and align the same segments of caption, which appears in consecutive frames to obtain multi-frame information of the pixel. In order to reduce the complexity of the background, we analyze the characteristics of the background changing in the time domain as well as, the inherent characteristics of the background and text. Furthermore, we take the fusion of them as the characteristics of Logistic model. Normalizing the characteristic of the model based on caption blocks which is the unit of enhancement, we take the result as the input parameter of the Logistic model. **Result** We select 60 videos with caption from the Paike column of Youku and divide caption into three categories: the special complex background caption containing shadows or stroke effects, the common complex background caption, and the single background caption. We respectively implement four caption enhancement methods: OTSU with adaptive threshold method based on single frame, multiple frame averaging method, minimum pixel value search method and the method proposed in this paper, we use these four methods for each kind of caption for our caption enhancement experiment. We use Hanwang OCR to recognize the enhanced caption and take the recognition accuracy as the evaluation instance of the caption enhancement effect. Experimental results show that the recognition accuracy of the three kinds of caption are 81.76%, 97.13%, and 81.76% respectively after enhanced by adopting the method in this paper. Comparing with the best results of the other three methods, the accuracy respectively increased by 24.35%, 2.70% and 2.70%. Thus, the method in this paper can adapt to both complex background and single background caption. Especially, the enhanced effect of complex background caption containing shadows or stoke show a significant improvement. **Conclusion** In this paper, we propose a method of fusing multi-frame information to realize caption enhancement based on the Logistic model. This method can reduce the complexity of the caption background and improve the contrast between the background and the text as well. Furthermore, the enhanced caption can be recognized well by OCR software. However, the parameters of the Logistic model are static values acquired by artificial parameter adjustment. If we can dynamically adjust parameters according to the characteristics of different video caption, the recognition accuracy will be further improved.

Key words: complex background; caption enhancement; Logistic model; caption detection and tracking; time domain feature

0 引 言

视频字幕中蕴含着丰富信息,与视频表达的内容紧密相关,为视频检索及相关内容研究提供了必要的数据来源。要想充分利用视频字幕信息,关键是将字幕图像识别成文本。但是,与文档图像中的文本识别相比,视频字幕的识别有以下难点:1)受视频的存储与传输的影响,视频字幕的分辨率一般较低;2)视频字幕常嵌于视频的场景之中,复杂的背景和较低的对比度,使得字幕与背景分离困难。

尽管目前光学字符识别(OCR)软件的识别精度已在提高,但对复杂背景的视频字幕图像,仍然不能达到较高识别率。因此,在输入 OCR 识别软件之前,对定位到的字幕区域,需进行字幕增强处理,将文本与背景分离,从而提高字幕识别率。在视频处理过程中,为了有效突出视频字幕的效果,常采用白色作为字幕颜色,因此,本文将研究对象限制在仅含白色字幕的视频上。

1 相关研究

在现有的字幕增强方法中,图像二值化可以直观地将视频字幕区域分成两部分。目前,关于图像的二值化已有许多相关研究[1-3],但是这些方法对于复杂背景的字幕增强效果较差。如图 1 是一段视频字幕原图、采用自适应阈值的最大类间方差(OTSU)算法及 Niblack 算法进行图像二值化的结果。从图中可以看出,图像二值化无法直接对复杂背景的视频字幕进行增强。

(a) 视频字幕原图

(b) 采用OTSU自适应阈值算法的二值化结果

(c) 采用Niblack算法的二值化结果

图 1 视频字幕图像及其二值化结果

Fig. 1 Binary image of video caption

关于字幕增强的研究，主要分为基于单帧的字幕区域增强[4]与基于多帧的字幕区域增强[5-9]两类方法。基于单帧的字幕区域增强方法，主要利用文本的连通性及水平分布特征，采用插值法或直方图拉伸法提高字幕区域的分辨率及对比度。该类方法在简单背景下，可以较好地分离字幕区的文本与背景；但对于具有复杂背景的字幕区域，其效果并不理想。基于多帧的字幕区域增强方法，主要有多帧平均法[5-7]和最小像素搜索法[7-9]。多帧平均法能够降低背景的复杂度，但不能够提高背景与文本的对比度。最小像素搜索法可以提高背景与字幕的对比度，但容易受噪声影响。本文通过分析视频字幕的时域特征，提出基于 Logistic 模型的视频字幕增强方法，在降低字幕背景的复杂度的同时，提高背景与文本的对比度。

视频的时域特征主要有：1）字幕中同一文本串持续出现在几十帧甚至几百帧中；2）在文本串的生存期里，文本串的颜色基本保持不变（除一些噪声影响），而背景的颜色可能会有很大的变化；3）当视频字幕运动时，呈现水平或垂直的线性运动。因此，通过融合多帧中的字幕信息，可以有效降低背景的复杂度，提高字幕区域的对比度，从而实现视频字幕的增强。

本文对视频帧序列进行字幕检测与跟踪，获取连续帧中的字幕区域，并确定相邻帧中字幕的偏移值；对获得的字幕条进行等距离抽取，并根据连续多帧中出现的相同字幕区域进行块分割；建立 Logistic 模型，以字幕块为基本单元，提取字幕背景变化特征及背景与文本的固有特征，对其量化及归一化后，输入 Logistic 模型，得到最终的字幕增强图像。

2　字幕检测与跟踪

利用视频时域特征进行字幕增强的关键是获取出现在连续帧中的同一字幕片段。因此，对视频字幕的检测与跟踪是字幕增强的基本前提。由于视频中非滚动字幕的同一字幕片段在连续帧中位置基本保持不变，而滚动字幕在连续帧中运动偏移更具有一般性，因此，以含水平滚动字幕的视频作为研究对象，通过分析字幕特征，有效地对其进行检测与跟踪。

设一段视频经解帧后，得到的视频帧集合为 $\boldsymbol{C}=\{f_1,f_2,\cdots,f_n\}$，其中 n 表示视频中包含的帧数。滚动字幕的基本特征有：

1）视频字幕呈自右向左的线性运动，从视频的全局来看，视频字幕区域的上边界 *Top* 及下边界 *Bot* 保持不变；

2）字幕滚动具有连续性，即一条完整滚动字幕将出现在连续帧 $f_k,f_{k+1},\cdots,f_{k+t-1}$ 中，且在含滚动字幕的相邻视频帧 f_i、f_{i+1} 中，字幕区域所包含的文本片段 $\boldsymbol{Seg}_i$、$\boldsymbol{Seg}_{i+1}$ 存在交集，即 $\boldsymbol{Seg}_i\cap\boldsymbol{Seg}_{i+1}\neq\varnothing$；

3）滚动字幕在刚进入视频画面中时，由于字幕区长度较短，无法立即检测到视频字幕；

4）字幕的滚动速度非严格匀速，即相邻帧的字幕偏移量不完全相等。

2.1　字幕检测

目前，关于视频字幕检测的研究，大致可分为基于区域的方法[10]、基于边缘和梯度的方法[11-13]、基于纹理的方法[14]及基于学习的方法[15]。这 4 类方法有各自的优缺点及适用范围。结合视频中滚动字幕的特征，从视频全局出发，通过定位字幕区域的上下边界、左右边界以及字幕的起始帧和结束帧，实现滚动字幕的检测。而其中对上下边界的定位是字幕检测的关键，具体采用以下方法：

选择 0°、45°、90°、135°方向的 Sobel 算子对视频帧进行边缘检测；对相邻帧的边缘图像作差分后，统计水平方向的边缘点个数，作为相邻帧的水平边缘特征；对水平边缘特征设定静态阈值，提取满足条件的候选上下边界值，并连同出现频数添加至上下边界集合 $\boldsymbol{CT}$、$\boldsymbol{CB}$ 中，最终得到

$$\boldsymbol{CT}=\{(top_i,c_{1i})\mid 1\leqslant i\leqslant M\}$$

$$\boldsymbol{CB}=\{(bot_j,c_{2j})\mid 1\leqslant j\leqslant N\}$$

c_{1i}、c_{2j} 分别为候选上边界 top_i、候选下边界 bot_j 出现的频数，M、N 分别为候选上边界、候选下边界的数目；分别选取集合 $\boldsymbol{CT}$ 和 $\boldsymbol{CB}$ 中的出现频数最高的候选上下边界 top_i、bot_j 作为滚动字幕区域的上下边界 *Top*、*Bot*。

采用类似方法，从全局上定位滚动字幕区域的左右边界；并在滚动字幕区域内，检测字幕的起始帧和结束帧。根据滚动字幕的连续性，依次提取视频帧中滚动字幕 R_1、R_2、…、R_m，其中 m 为视频帧集 $\boldsymbol{C}$ 中包含的滚动字幕条数。

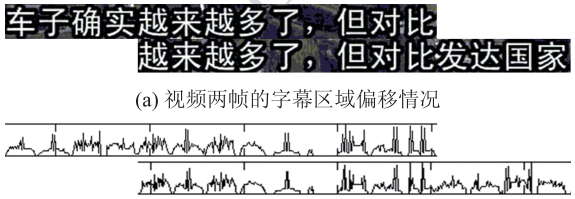
2.2　字幕跟踪

关于视频对象的跟踪，已有很多研究成果[16]，但很少研究涉及视频字幕的跟踪。文献[5]提出一种利用帧间差平方和（SSD）的字幕跟踪方法，但

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 5, May 2014

该方法在复杂背景的视频中,其准确率难以保证。结合视频滚动字幕的特征,采用基于垂直边缘特征的最小差平方和的方法对滚动字幕进行有效地跟踪。

对字幕条 $R_k(1\leqslant k\leqslant m)$ 的边缘特征图像,在垂直方向上进行边缘点个数统计,作为字幕的垂直边缘特征,用向量 $\boldsymbol{VE}^{(k)}=(Vn_1^{(k)},Vn_2^{(k)},\cdots,Vn_{l_k}^{(k)})$ 表示,其中 $Vn_i^{(k)}(1\leqslant i\leqslant l_k)$ 为字幕条 R_k 中第 i 列的边缘点个数,l_k 为字幕条 R_k 的长度。相邻帧的字幕区域偏差与垂直边缘特征图如图 2 所示。



(a) 视频两帧的字幕区域偏移情况

(b) 字幕区域的垂直边缘特征图

图 2　相邻帧的字幕区域及 Sobel 边缘特征图像

Fig. 2　Caption area and edge feature image of adjacent frame

取字幕条 R_{k+1} 的 $\left[0,\frac{l_k}{2}\right]$ 片段的垂直边缘特征 $Vn_j^{(k+1)}\left(1\leqslant j\leqslant\frac{l_k}{2}\right)$,与字幕条 R_k 的 $\left[p,p+\frac{l_k}{2}\right]\left(0\leqslant p\leqslant\frac{l_k}{2}\right)$ 区间的垂直边缘特征 $Vn_j^{(k)}\left(p\leqslant j\leqslant p+\frac{l_k}{2}\right)$,依次求其差平方和 SS;当 SS 达到最小时,p 的取值即为字幕条 R_{k+1} 相对字幕条 R_k 的偏移值。

3　字幕增强

对字幕进行跟踪操作后,合并各帧中的字幕区域 $R_i(1\leqslant i\leqslant m)$,$m$ 表示提取到的字幕条数,得到一条完整的视频字幕条 Cap。将完整字幕条 Cap 分割成 p 个字幕块 $block_j(1\leqslant j\leqslant p)$,使得字幕块所对应的区间内仅包含出现在多个连续帧中的同一字幕片段,如图 3 所示。

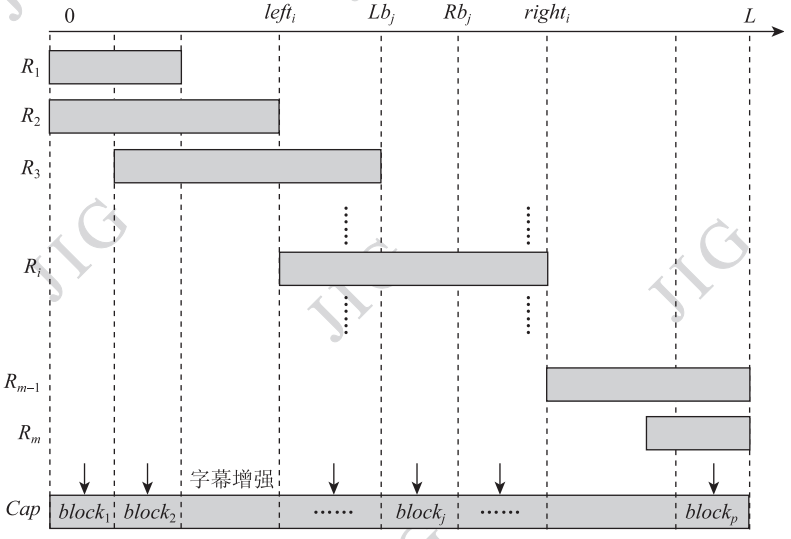


图 3　字幕跟踪及分割示意图

Fig. 3　Caption tracking and block segment

完整视频字幕条 Cap 的长度为 L;以字幕条 Cap 作为位置参考,其左边界对应坐标原点,右边界对应坐标点 L;分割得到的字幕块 $block_j$ 对应区间为 $[Lb_j,Rb_j]$;在区间 $[Lb_j,Rb_j]$ 上对应着多个字幕条 $(R_i,R_{i+1},\cdots,R_{i+Num-1})$ 的字幕片段,其中 Num 表示该区域对应到的字幕条数;字幕条 R_i 所在区间为 $[left_i,right_i]$,长度为 L_i。

以单个字幕块作为基本单元,建立 Logistic 模型;提取字幕背景的变化特征及字幕背景及文本固有特征,并对其进行融合处理后,输入 Logistic 模型,从而实现视频字幕的增强。

3.1　Logistic 模型

Logistic 曲线是一种常见的 S 型曲线,起源于生态学,描述意义在于:当生物在受环境容量限制时,生物数量随时间 t 的变化呈现出 S 型的增长模式,增长率由小变大,再由大变小。其一般函数式为

$$y=\frac{K}{1+\mathrm{e}^{a-bt}} \tag{1}$$

$$K>0,a>0,b>0$$

式中，K 为环境的最大容量，t 为时间变量，b 为增长率，a 为常数。从函数式(1)中可以得出

$$\lim_{t\to+\infty}\frac{K}{1+\mathrm{e}^{a-bt}}=K \tag{2}$$

$$\lim_{t\to-\infty}\frac{K}{1+\mathrm{e}^{a-bt}}=0 \tag{3}$$

根据 Logistic 曲线的特征，当增长率 b 较大时，曲线在 $+\infty$ 方向上可以较快地收敛至值域的最小上界 K；同时，曲线在 $-\infty$ 方向上亦可以较快地收敛至值域的最大下界 0。令 $K=255$，则曲线的值域为(0,255)。用 Logistic 曲线的输出值(四舍五入后的整数)，作为最终字幕增强后像素点 P 的灰度值。此时，在增强后的字幕图像中，像素值较多地分布在白色点及黑色点上，而中间较少的灰度点可以作为白色向黑色过渡的像素点。这样，既保持了图像灰度值的连续性，又提高了字幕图像的对比度。

为了减少字幕中背景像素点之间的差异，降低字幕背景的复杂度，选取字幕背景在多帧中的变化特征及背景像素与文本像素的特征，并对其进行量化及归一化，作为 Logistic 模型的输入参数。

在 Logistic 曲线的基础上，提出 Logistic 模型：将像素点 P 在多帧中的字幕特征 X 作为模型输入参数，将字幕增强后像素点 P 的灰度值 Y 作为模型的输出参数。此时，该模型可表示为

$$Y=\left\lfloor\frac{255}{1+\mathrm{e}^{a-b\times X}}+0.5\right\rfloor \tag{4}$$

式中，$X\in[0,se]$，$se>0$，$a>0$，$b>0$，$Y\in[0,255]$ 且 $Y\in\mathbf{N}$。

3.2　字幕条抽取

以字幕块 $block_j$ 作为一个研究单元，即 Logistic 模型的输入参数是以字幕块为单位进行归一化操作得到的。因此，任意一个字幕块，在进行字幕增强之后，其所有像素点的灰度值应在[0,255]区间内，且必然存在灰度值为 0 及 255 的像素点。为了保证视频的观看效果，视频字幕的滚动速度通常不会太快，相邻帧的字幕偏移量较小。这将导致分割的字幕块中会出现仅包含背景像素的块，从而使 Logistic 模型的输出值失效。因此，需要等距离地对字幕条 $R_i(1\leqslant i\leqslant m)$ 进行抽取，增加新字幕条序列中相邻字幕条间的偏移值。

虽然滚动字幕的运动并非严格匀速，在跟踪字幕时无法使用等距离的字幕偏移值，但在进行字幕条抽取时，可以忽略字幕运动速度上的微小差异。设 Δs 表示相邻字幕条 R_k、R_{k+1} 之间的运动偏移值，其可以使用字幕条 R_1、R_2 之间的偏移值近似表示，即

$$\Delta s\approx left_2-left_1 \tag{5}$$

式中，$left_1$、$left_2$ 分别表示字幕条 R_1、R_2 的左边界。设 Δd 表示对字幕条 $R_i(1\leqslant i\leqslant m)$ 进行抽取的间隔大小，其取值依据为：通过每隔 Δd 个帧对字幕条 $R_i(1\leqslant i\leqslant m)$ 进行抽取，得到新的字幕条序列 $R'_i(1\leqslant i\leqslant t)$，其中 t 表示新序列的字幕条数目；在序列 $R'_i(1\leqslant i\leqslant t)$ 中，相邻帧之间的偏移值至少超过视频帧宽度的$\frac{1}{6}$，从而保证在字幕块分割后，每个字幕块中均含有多个字幕文本。Δd 取值可表示为

$$\Delta d=\left\lfloor\frac{width'}{6\times\Delta s}\right\rfloor \tag{6}$$

式中，$width'$ 表示视频帧的宽度，Δs 表示相邻字幕条间的运动偏移值。

3.3　特征选取

字幕块 $block_j$ 所对应的区间为 $[Lb_j,Rb_j]$，包含的字幕条有 $R'_i,R'_{i+1},\cdots,R'_{i+n-1}$，其中 n 表示包含的字幕条数，特征提取示意图如图 4 所示。

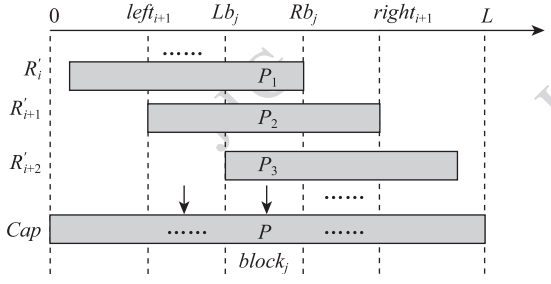


图 4　特征提取示意图

Fig. 4　Feature extraction

设 $BL_1,BL_2,\cdots,BL_n$ 分别为 $R'_i,R'_{i+1},\cdots,R'_{i+n-1}$ 在区间 $[Lb_j,Rb_j]$ 上的字幕片段，与字幕块 $block_j$ 中任意一个像素点 $P(R,G,B)$ 相对应的点集为 $\boldsymbol{CP}$，则

$$\boldsymbol{CP}=\{P_i(R_i,G_i,B_i)\mid 1\leqslant i\leqslant n\}$$

式中，$P_i(R_i,G_i,B_i)$ 为字幕片段 BL_i 上的像素点，R_i,G_i,B_i 分别为 P_i 的 RGB 分量值，n 为区间内字幕条数。根据字幕背景在多帧中的变化特点、字幕背景与字幕文本的固有特点，提取并量化字幕对应点间颜色差异、背景彩色度、文本灰度距离 3 个特征，以区分字幕条中背景与文本像素点，同时削减背景

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 5, May 2014

点之间的差异,降低背景的复杂度。

特征 1　对应点间颜色差异。同一字幕片段在连续多帧中,字幕文本的颜色基本保持不变,而其背景颜色可能会有较大变化。因此,采用字幕片段中相应点之间的颜色差异 DC_P 作为第 1 特征值。DC_P 的计算表示为

$$DC_P = \sqrt{\frac{1}{3}(PR^2 + PG^2 + PB^2)} \tag{7}$$

$$PR = \max_{1\leqslant i\leqslant n} R_i - \min_{1\leqslant i\leqslant n} R_i \tag{8}$$

$$PG = \max_{1\leqslant i\leqslant n} G_i - \min_{1\leqslant i\leqslant n} G_i \tag{9}$$

$$PB = \max_{1\leqslant i\leqslant n} B_i - \min_{1\leqslant i\leqslant n} B_i \tag{10}$$

式中,PR、PG、PB 分别为点集 $\boldsymbol{CP}$ 包含的像素点的 R 值极差、G 值极差、B 值极差,n 为区间内字幕条数,R_i, G_i, B_i 分别为 P_i 的 RGB 分量值,$P_i(R_i, G_i, B_i) \in \boldsymbol{CP}$。

特征 2　背景彩色度。为了增强字幕的显示效果,复杂背景中的滚动视频字幕,常采用白色作为字幕颜色。而字幕区域中出现的彩色点,则应被视为背景点。因此,用像素点的彩色度 CL_P 作为第 2 特征值,抑制彩色的像素点。衡量像素点彩色度

$$CL_P = \frac{1}{n}\sum_{i=1}^{n} PCL_i \tag{11}$$

$$PCL_i = \max(R_i, G_i, B_i) - \min(R_i, G_i, B_i) \tag{12}$$

式中,PCL_i 为点集 $\boldsymbol{CP}$ 中像素点 P_i 的 RGB 分量的极差,$\max(R_i, G_i, B_i)$,$\min(R_i, G_i, B_i)$ 分别表示点 $P_i(R_i, G_i, B_i)$ 的 RGB 分量的最大值、最小值。

特征 3　文本灰度距离。为了达到更好的显示效果,视频字幕的颜色一般较背景颜色更亮。因此通过计算像素点的 RGB 值与白色点的距离 DW_P,来衡量像素点的灰色程度。把 DW_P 作为第 3 特征值,来抑制深色像素点,其计算方法为

$$DW_P = \frac{1}{n}\sum_{i=1}^{n} \sqrt{\frac{1}{3}(DR_i^2 + DG_i^2 + DB_i^2)} \tag{13}$$

$$DR_i = 255 - R_i \tag{14}$$

$$DG_i = 255 - G_i \tag{15}$$

$$DB_i = 255 - B_i \tag{16}$$

式中,R_i, G_i, B_i 分别为 P_i 的 RGB 分量值,$P_i(R_i, G_i, B_i) \in \boldsymbol{CP}$,$DR_i$、$DG_i$、$DB_i$ 分别为 RGB 分量的灰度距离。

3.4　字幕增强

使用特征 1、特征 2 及特征 3 对字幕块 $block_j$ 中像素点 P 在多帧中的信息进行量化提取。特征 1 描述在多帧中对应像素点的颜色差异程度;其值越大,视为背景点的可能性越大。特征 2 描述像素点的彩色程度;其值越高,视为背景点的可能性大。特征 3 描述像素点的灰度距离;其值越大,视为背景点的可能性越大。同时,当非彩色的像素点像素值较小,且其在多帧中变化也较小时,特征 3 将对点像素点的判别起主要作用。特征 1、特征 2 及特征 3 分别从像素点的 3 个独立方面进行描述其在多帧中的特征,且在对 3 个特征进行量化的过程中,已经确保 3 个特征处于同一个取值空间[0,255]。因此,直接对以上 3 个特征进行加权融合,得到像素点 P 在 Logistic 模型中的综合特征值

$$CH_P = \alpha \times DC_P + \beta \times CL_P + \gamma \times DW_P \tag{17}$$

式中,$\alpha + \beta + \gamma = 1$。

在字幕块 $block_j$ 所对应的区域内,将像素点 P 的综合特征值 CH_p 归一化至区间[0,se]内,并作为 Logistic 模型的输入参数

$$X = \frac{-se \times (CH_P - \min\limits_{1\leqslant i\leqslant sum} CH_{Q_i})}{\max\limits_{1\leqslant i\leqslant sum} CH_{Q_i} - \min\limits_{1\leqslant i\leqslant sum} CH_{Q_i}} \tag{18}$$

式中,Q_i 表示字幕块 $block_j$ 中的像素点,CH_{Q_i} 为 Q_i 的综合特征值,sum 为字幕块 $block_j$ 的面积,即

$$sum = h' \times (Rb_j - Lb_j) \tag{19}$$

式中,h'表示字幕区域的高度,Lb_j、Rb_j 分别表示字幕块 $block_j$ 的左、右边界。字幕增强后,点 $P(R, G, B)$ 的 RGB 分量值为

$$R = G = B = \left\lfloor \frac{255}{1 + e^{a - b \times X}} + 0.5 \right\rfloor \tag{20}$$

式中,a、b 为 Logistic 模型的参数。对完整字幕 Cap 中所有字幕块 $block_j(1 \leqslant j \leqslant p)$ 的像素点,以其所在的字幕块为单位,利用式(20)进行像素值变换,最终得到字幕增强图像。

4　实验结果

4.1　参数选取

为了使字幕增强后的字幕图像有较高的对比度和较低的背景复杂度,将把复杂背景字幕转化为纯色背景的字幕作为增强目标。因此,通过对 413 个无重复的纯色背景下的视频字幕条进行灰度值分布统计,来确定 Logistic 模型中的参数 a、b、se。统计结果如下:

1)在同一个字幕条中,灰度值在区间[0,50]上的像素点个数约占该字幕条总像素点个数的 60%;

2)灰度值在区间[51,100]、[101,150]、[151,200]、[201,255]上的像素点个数均占其所在字幕条总像素个数的 10% 左右。

为了使字幕增强的效果更明显,增加白色像素点的比例,并减少灰色过渡区的区间长度,通过调整 Logistic 模型的参数值及输入变量的定义域[0,*se*],使像素值 *Y* 落在区间[0,50]上的变量 *X* 取值范围占定义域的 45%,落在区间[201,255]上的变量 *X* 取值范围占定义域的 45%,落在中间灰色过渡区的变量 *X* 取值范围占定义域的 10%。经过调整参数得 $a=20$,$b=2$,$se=20$,即 Logistic 模型表达为

$$Y=\left\lfloor \frac{255}{1+e^{20-2\times X}}+0.5 \right\rfloor, X\in[0,20] \quad (21)$$

此时,Logistic 曲线图像如图 5 所示。

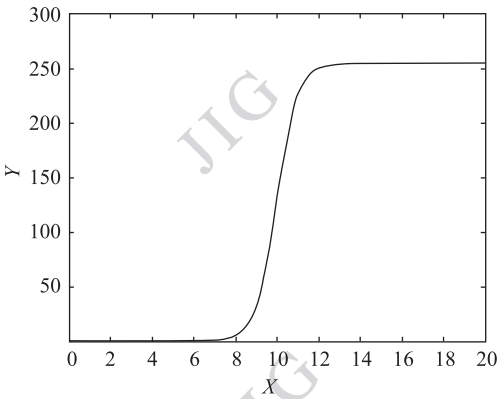


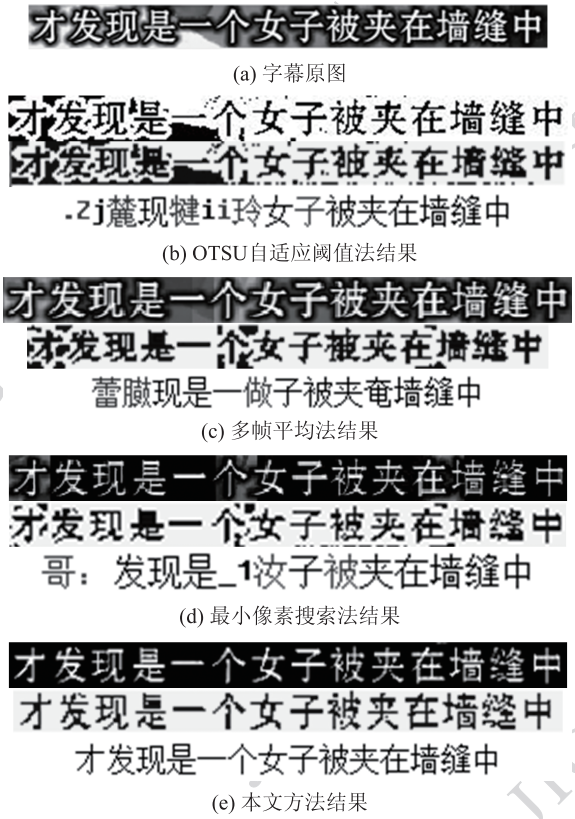
图 5　Logistic 曲线

Fig. 5　Logistic curve

对于式(17)中 3 个特征进行融合时权重大小的选取,实验中分别取 α、β、γ 值为 0.25、0.35、0.4。通过多次实验调参确定的参数值,仅可以使实验结果达到一个较优值,而无法确定是否为最优值。

4.2　实验结果

为了评价本文方法的有效性,从优酷网的拍客栏目中选取 60 个含滚动字幕的视频。提取视频中滚动字幕,并分别采用基于单帧的 OTSU 自适应阈值法、多帧平均法、最小像素搜索法及本文方法进行字幕增强实验。由于汉王 OCR 是目前比较成熟的字符识别软件,且在识别过程中包含对图像的二值化处理,因此,可以将字幕增强后的图像直接输入汉王 OCR 软件进行识别,并将识别正确率作为字幕增强效果的评价指标。字幕增强后的图像、OCR 二值化结果及 OCR 识别结果如图 6 所示。



(a) 字幕原图

(b) OTSU自适应阈值法结果

(c) 多帧平均法结果

(d) 最小像素搜索法结果

(e) 本文方法结果

图 6　各方法字幕增强、二值化、OCR 识别结果

Fig. 6　The results of enhancement, OCR binary and recognition

基于单帧的 OTSU 自适应阈值法是经典的图像二值化算法,该方法可以很好地适应一般图像的二值化,但不能够满足视频字幕的二值化需求;同时,由于该方法仅依赖于单帧信息,从而损失过多视频字幕在时域上的特征。多帧平均法,能够降低背景的复杂度,但不能够提高背景与文本的对比度。最小像素搜索法,可以提高背景与字幕的对比度,但极容易受噪声影响。本文方法充分融合多帧间的字幕信息,通过特征选取降低背景的复杂度,并利用 Logistic 模型提高背景与字幕的对比度。

为了保证字幕与背景的对比度,便于观众阅读,许多视频字幕的文本边缘通常会添加高对比度的阴影或描边效果,而该类视频字幕在一定条件下将会影响字幕增强的效果。因此,实验将 60 个视频的字幕分为 3 类:特殊复杂背景字幕、普通复杂背景字幕、单一背景字幕。特殊复杂背景字幕是指含有阴影或描边效果的复杂背景下的视频字幕,实验中有 32 个该类视频,包含 2 905 个字符;普通复杂背景字幕是指不含阴影及描边效果的复杂背景下的字幕,实验中有 15 个该类视频,

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 5, May 2014

包含 1 568 个字符;单一背景字幕是指出现在画面外的纯色背景(非复杂背景)下的视频字幕,实验中有 13 个该类视频,包含 1 325 个字符。对 3 类视频分别进行字幕增强实验,其 OCR 识别结果如表 1 所示。

按照表 1 的统计结果,对不同类型的字幕在各个方法下的 OCR 识别正确率进行比较,结果如图 7 所示。

表 1　各方法的 OCR 识别正确率

Table 1　OCR recognition accuracy

视频字幕类型	字符数	OTSU 自适应阈值法/%	多帧平均法/%	最小像素搜索法/%	本文方法/%
特殊复杂背景字幕	2 905	45. 47	44. 10	65. 75	81. 76
普通复杂背景字幕	1 568	84. 50	93. 75	94. 58	97. 13
单一背景字幕	1 325	96. 91	97. 81	97. 89	98. 19
平均值	5 798	75. 63	78. 55	86. 07	92. 36

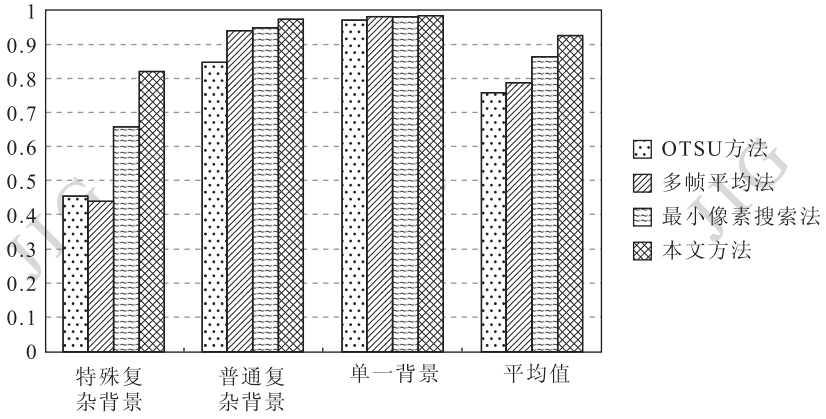


图 7　各方法下的 OCR 识别正确率比较

Fig. 7　Comparison of OCR recognition accuracy

从表 1 中数据和图 7 可以看出,本文方法不仅适应于复杂背景下的滚动字幕,对于单一背景下的滚动字幕同样可以达到较高的识别率,且在 3 类视频字幕中,单一背景字幕的增强效果最好;本文方法对 3 种类型字幕的增强效果均优于其他方法,且与其他方法最好的实验结果相比较,字幕增强后的 OCR 识别率分别提高了 24. 35% 、2. 70% 、0. 31% 。

本文字幕增强方法,最大的优势是对特殊复杂背景字幕的增强,且当视频字幕背景点与文本边缘点对比度较高时,其字幕增强的效果较其他方法更加明显。图 8 为含特殊复杂背景视频在连续两帧中出现的字幕片段,使用 4 种字幕增强方法进行实验,增强结果及 OCR 识别结果如表 2 所示。

对于该类特殊复杂背景字幕,由于无法找到合适的阈值将其二值化,且多帧平均法或最小像素搜索法无法将字幕的边缘与背景像素区分或合并,从而使得这些方法失效。而本文方法可以在弱化背景

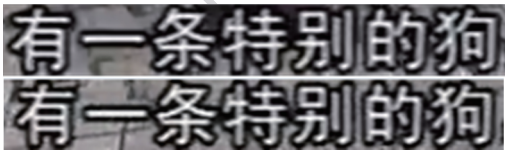


图 8　特殊复杂背景字幕片段

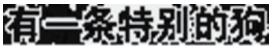
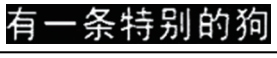
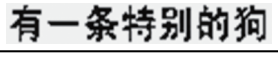
Fig. 8　The segment of special caption with complex background

复杂度的同时,提高背景与字幕的对比度,即将背景像素赋予较低的像素值;对于字幕边缘,由于其灰度差异,同样可以被归属至背景像素,从而使背景像素与字幕边缘融合在一起,达到字幕增强的效果。因此,本文方法对特殊复杂背景字幕的增强效果,对比其他方法有较大的提高。

以含滚动字幕的视频作为研究对象,是为了使字幕跟踪过程更具有一般性,且能够在字幕跟踪的基础上,更清晰地表达字幕增强方法。而单独从基于 Logistic 模型的字幕增强方法的理论分析,字幕增

表 2　特殊复杂背景字幕的增强及 OCR 识别

Table 2　The result of enhancement and OCR recognition of special caption

方法	字幕增强效果	OCR 二值化处理	OCR 识别结果
OTSU 自适应阈值法			哥哥臻蕊聚圈曰面
多帧平均法			寓音国国团翰国
最小像素搜索法			隋嚣舞渍锅啦镶
本文方法			有一条特别的狗

强的关键点在于对 Logistic 模型输入参数的构建，即字幕在多帧之间信息特征的选取。该方法所提取的多帧间对应点颜色差异、背景彩色度及文本灰度距离 3 个特征，分别是依据字幕在多帧间的变化特点、字幕背景及字幕文本的固有特点。这些特征仅与字幕背景在视频时域上的变化有关，即与字幕背景的运动情况有关，而与字幕是否为滚动字幕无关。同时，实验结果表明，当静态的字幕背景为纯色时，尽管特征一并未起到作用，但在特征 2、特征 3 的作用下，该方法依然有效。因此，所提的基于 Logistic 模型的字幕增强方法，其适用的字幕类型可分为以下 4 类：运动背景、运动文本；运动背景、静态文本；静态背景、运动文本；静态且纯色背景、静态文本。而视频中对于静态文本处于静态复杂背景中的字幕类型并不常见，这与观众观看视频的大体感受是保持一致的。

5　结　论

尽管目前 OCR 字符识别软件愈发成熟，但对于复杂背景中的字幕图像，若直接送入识别依然很难得到较好的识别效果。因此，提出一种基于 Logistic 模型的融合多帧信息的字幕增强方法。其优点如下：

1）提取并量化视频多帧之间的字幕变化特征、背景及文本固有特征，作为 Logistic 模型的输入值，可降低字幕背景的复杂度；

2）利用 Logistic 模型，将字幕区域的像素点根据其在多帧中的特征值进行像素值映射，可提高背景与字幕的对比度，达到较好的字幕增强效果；

3）对于视频字幕含有阴影或描边的特殊类型，本文方法的字幕增强效果比其他方法具有更强的适用性；

4）本文方法适用的字幕类型较广泛，包括：运动背景、运动文本，运动背景、静态文本，静态背景、运动文本，静态且纯色背景、静态文本。

本文方法在字幕增强方面，既降低了字幕背景的复杂度，又提高了背景与文本之间的对比度，且增强后字幕图像可以很好地被 OCR 软件所识别。但是，本文依然有需要改进的地方：

1）在提取字幕 3 个特征值时，是以字幕的颜色为非彩色，且对比颜色偏亮色为前提，若可以采用基于颜色的聚类方法，提取视频字幕的颜色值，从而动态设定文中出现的 3 个特征值，使该方法有更广泛的适应范围。

2）在参数选择上，主要通过多次实验得到调整参数，属于静态值。如果可以根据不同视频字幕的特征，动态调整参数，识别效果将会有进一步的提高。

参考文献（References）

[1] Otsu N. A threshold selection method from grey-level histograms [J]. IEEE Transactions on Systems, Man, and Cybernetics, 1979, 9(1): 377-393.

[2] Niblack W. An Introduction to Digital Image Processing [M]. Denmark: Strandberg Publishing Company Birkeroed, 1986, 115-116.

[3] Liu K. Extraction and recognition of video caption[D]. Beijing: Beijing Information Science and Technology University, 2009. [刘坤. 视频字幕的提取与识别研究[D]. 北京：北京信息科技大学，2009.]

[4] Wang Y D, Jiang X S. News video text segmentation algorithm based on gradient reinforcement [J]. Jounal of Computer-aided Design & Computer Graphics, 2009, 21(8): 1170-1174. [王一丁，蒋小森. 基于梯度增强的新闻字幕分割算法[J]. 计算机辅助设计与图形学学报，2009，21(8)：1170-1174.]

[5] Li H, Doermann D. Text enhancement in digital video using multiple frame integration[C]//Proceedings of the seventh ACM international conference on Multimedia(Part 1). New York, USA: ACM, 1999: 19-22. [DOI: 10.1145/319463.319466]

中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS
Vol. 19, No. 5, May 2014

[6] Yi J, Peng Y X, Xiao J G. Recognition of text in video based on color clustering and multiple frame integration [J]. Journal of Software, 2011, 22(12): 2919-2933. [易剑, 彭宇新, 肖建国. 基于颜色聚类和多帧融合的视频文字识别方法[J]. 软件学报, 2011, 22(12): 2919-2933.]

[7] Zhu C J, Li C, Xue L, et al. Video text enhancement using multiple frame information [J]. Journal of Image and Graphics, 2008, 13(9): 1667-1672. [朱成军, 李超, 薛玲, 等. 一种基于多帧视频的文本图像质量增强方法[J]. 中国图象图形学报, 2008, 13(9): 1667-1672.] [DOI: 10.11834/jig.20080907]

[8] Wang R R, Jing W J, Wu L D. A novel video caption detection approach using multi-frame integration [J]. Journal of Computer Research and Development, 2005, 42(7): 1191-1197. [王蓉蓉, 金万军, 吴立德. 一种新的利用多帧结合检测视频标题文字的算法[J]. 计算机研究与发展, 2005, 42(7): 1191-1197].

[9] Mi C J, Li Y, Xue X Y. Video texts tracking and segmentation based on multiple frames[J]. Journal of Computer Research and Development, 2006, 43(9): 1523-1529. [密聪杰, 刘洋, 薛向阳. 基于多帧图像的视频文字跟踪和分割算法[J]. 计算机研究与发展, 2006, 43(9): 1523-1529.]

[10] Shivakumara P, Dutta A, Tan C L, et al. Multi-oriented scene text detection in video based on wavelet and angle projection boundary growing [J]. Multimedia Tools and Applications, 2013, 2: 1-25. [DOI: 10.1007/s11042-013-1385-0]

[11] Cao X X, Liu J, Yang X D, et al. A novel algorithm for the video caption extraction [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2013, 49(2): 197-202. [曹喜信, 刘京, 杨旭东, 等. 一种新的视频字幕提取算法[J]. 北京大学学报: 自然科学版, 2013, 49(2): 197-202.]

[12] Yang Z, Shi P. Caption detection and text recognition in news video [C]//Proceedings of the 5th International Congress on Image and Signal Processing. Washington DC, USA: IEEE Computer Society, 2012: 188-191. [DOI: 10.1109/CISP.2012.6469754]

[13] Sang L. Rolling and non-rolling news subtitle location and segmentation [D]. Shanghai: Shanghai Jiao Tong University, 2012. [桑亮. 滚动与非滚动新闻字幕的定位与分割[D]. 上海: 上海交通大学, 2012.]

[14] Bouaziz B, Mahdi W, Ardabilain M, et al. A new approach for texture features extraction: Application for text localization in video images [C]//Proceedings of the IEEE International Conference on Multimedia and Expo. Washington DC, USA: IEEE Computer Society, 2006: 1737-1740. [DOI: 10.1109/ICME.2006.262886]

[15] Zhou C J. Video caption detection algorithm based on multiple instance learning [D]. Heilongjiang: Harbin Engineering University, 2012. [周长建. 基于多示例学习的视频字幕提取算法研究[D]. 黑龙江: 哈尔滨工程大学, 2012.]

[16] Xu H Y. Object recognition and motion tracking based on viedo [D]. Shanghai: Shanghai Jiao Tong University, 2013. [许涵洋. 基于视频的物体识别及运动跟踪[D]. 上海: 上海交通大学, 2013.]