Miss.Dharmini Esther Thangam.P, Mrs.Akila Agnes.S / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622
www.ijera.com Vol. 3, Issue 1, January -February 2013, pp.1023-1028

The Stroke Filter Based Caption Extraction System

Miss.Dharmini Esther Thangam.P[#1](PG scholar),Mrs.Akila Agnes.S[#2](Asst.Prof)
Department of Computer Science and Engineering, Karunya University
Coimbatore, India

Abstract-

Captions in videos provide rich semantic information the contents of the video can be clearly understand with help of the captions. Without even seeing the video the person can understand about the video by knowing the captions in the video. By considering these reasons captions extraction in videos become an important prerequisite. In this paper we use a stroke filter, the stroke filter identifies the strokes in the video and usually caption regions have strokes such that the strokes identified belongs to captions by which the captions are detected. Then the locations of the captions are localized. In the next Step the caption pixels are separated from the background pixels. At last a post processing step is performed to check whether the correct caption pixels are extracted. In this paper a step by step sequential procedure to extract captions from videos is proposed.

***Keywords*-stroke filter, caption detection, caption extraction,caption pixels**

I. INTRODUCTION

Video processing is a particular case of signal processing, which often employs video filters and where the input and output signals are video files or streams. The use of digital video is becoming widespread all over the place, extending throughout the world. Captions in videos provide rich semantic information which is useful in video content analysis, indexing and retrieval, traffic monitoring and computerized aid for visually impaired. Caption extraction from videos is a difficult problem due to the unconstrained nature of general-purpose video. Text can have arbitrary color, size, and orientation. Backgrounds may be complex and changing. The previous caption extraction method considers each video frame as an independent image and it extracts caption from each video frame. However, captions occurring in video usually persist for several seconds. Consider if a movie played by 30 frames per second contains large number of video frames and if each video frame is considered as an independent image it takes more time and the same captions may be extracted again and again [5].To overcome this disadvantage in our method we consider the temporal feature in videos such that the time interval in which same caption is present and then caption extraction is carried out. By considering temporal feature approach the computational load is reduced and more accurate captions are extracted. The proposed caption extraction method consists of three important steps. They are caption detection, caption localization and caption segmentation. For caption detection we use stroke based detection method rather than edge, corner or texture based detection because sometimes background has dense edges so background edges may have been mistaken as that of caption edges in this kind of situation edge based detection doesn't works well, If the background has densely distributed corners corner based detection doesn't suits and in texture based detection we must previously know about the texture of the captions [4]. Therefore stroke filter is used because captions only Have stroke like edges. For caption localization spatial-temporal localization is used whereas the previous methods used DCT algorithm [7] .The drawback of DCT algorithm is that it doesn't considers the time interval of the captions so the same captions are found again and again. For caption segmentation we use spatial-temporal segmentation, the previous methods used only temporal segmentation [11]. The disadvantage of the previous method is that there are many types of close up and medium shots make it difficult to accurately distinguish them, it involves many complex steps and does not handle false positives due to camera operations, whereas the proposed system aims to guarantee each clip contains same caption, so that the efficiency and accuracy of caption extraction is greatly improved.

II. PROPOSED APPROACH

The proposed approach is a step by step sequential procedure to extract captions from videos. The proposed system functions as follows

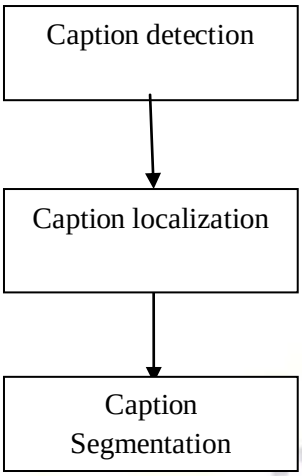


Fig .1Steps in proposed system

A. Caption Detection

A stroke filter is used to extract captions from videos. The stroke filter identifies the stroke like edges in the captions. Strokes of a pen or brush are the movements or marks that we make with it when writing or painting. The stroke filter which can detect the strokes in video images is designed as follows. First, we define a local image region as a stroke-like structure, if and only if, in terms of its intensities,(1) it is different from its lateral regions(2) its lateral regions are similar to each other (3) it is nearly homogenous. By using stroke filter the stroke like edge pixels are denoted by 1 and the remaining pixels are denoted by 0[17].The stroke like edge pixels are taken as contours. Sometimes the edge pixels may also come from complex background so to differentiate the caption pixels from edge pixels we use some techniques they are (i)the strokes always have a range of height and width which depend on the size of the characters and length and width of characters in captions may not be too long. Let $w(c_i)$ and $h(c_i)$ denote the width and height of minimum enclosing rectangle(MER) of contour c_i, if

$$W(c_i)<=\text{threshold1},h(c_i)<=\text{threshold2} \quad (1)$$

Where the thresholds are chosen based on the size of characters to be detected. If the above condition (i) is not satisfied the contour c_i is removed as non stroke edges in complex background.(ii)for each contour the gradient direction $\tilde{O}_k$ has to found out. The gradient $\tilde{O}_k$ is quantized into four values based on horizontal, vice-diagonal, vertical, main diagonal directions. Then each contour is broken down into several sub contours based on their connectivity and gradient direction [1]. Then each where the thresholds are chosen based on the size of characters to be detected. If the above condition (i) is not satisfied the contour c_i is removed as non stroke edges in complex background.(ii)for each contour the gradient direction $\tilde{O}_k$ has to found out. The gradient $\tilde{O}_k$ is quantized into four values based on horizontal, vice-diagonal, vertical, main diagonal directions. Then each contour is broken down into several sub contours based on their connectivity and gradient direction. Then each sub contour must contain a parallel edge with the same gradient direction in a specified interval depending on the width of character strokes.

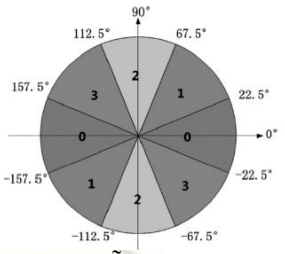


Fig.2 quantize $\tilde{O}_k$ into four values

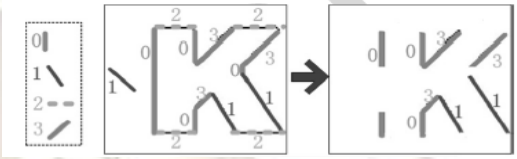


Fig. 3 subcontours in the character K

If a contour does not satisfy the above mentioned two conditions it is removed as non stroke edge and the edge doesn't belongs to captions [12]. If contour satisfies the above mentioned conditions it is considered as stroke like edge pixels and thus the captions are detected.

B. *Caption Localization*

There are two types of caption localization is performed. They are spatial localization and temporal localization Spatial localization involves finding the location of the caption in the video frame. Temporal localization involves finding the time interval in which the captions are found in consecutive video frames.

1)Spatial localization-We use SVM classifier to identify the location of the caption in the video frame. Compared with other classifiers such as neural network and decision tree, support vector machine (SVM) is easier to train and has better generalization ability. Therefore, we use the SVM classifier and choose radial basis function (RBF) as kernel function of this SVM classifier. The basic idea of SVM is to implicitly project the input space onto a higher dimensional space where the two classes are more linearly separable. Given the caption candidate produced by the caption detection module, we first normalize it to a height of 15 pixels height and corresponding resized weight. Then, we use a sliding window with 15X15 pixels to generate several samples[14]. We use the trained SVM classifiers to classify each of these samples and fuse the results. The SVM output scores of theses samples are averaged. If the average is larger that a predefined threshold, then the whole text line is regarded as a text line, otherwise it is regarded as a false alarm. We

set the threshold value as 0.3 based on the trade-off between the false acceptance rate (FAR) and false rejection rate (FRR).

2)Temporallocalization-Temporal localization involves identifying the time interval of the captions. It identifies the consecutive frames that containing the same caption and forms a clip with caption or a clip with no caption $S_{t1,t2}(x,y)=\sum(E_{t1}(x,y) \wedge E_{t2}(x,y))$ (2)

The logical AND operation is performed to find out whether the consecutive frames containing the same captions. If the result is 1 the both edge pixels are same otherwise the edge pixels are different. All the edge pixels of consecutive frames are performed logical AND operation to check whether the frames contains captions. If the frames containing same captions are identified then they are combined to form a clip[16]. The more important thing related to temporal localization is that whether the clip generated contains captions ,since only the clips containing captions are meaning full for the subsequent computation[8]. If the consecutive frames does not contain captions they form a clip with no caption. Consider consecutive video frames $I_{t1}(x,y)$, $I_{t2}(x,y)$, $I_{t3}(x,y)$…..... $I_{t(k-1)}(x,y)$ does not contain captions but frame $I_{tk}(x,y)$ contains caption then frames $I_{t1}(x,y)$ to $I_{t(k-1)}(x,y)$ will form a clip with no caption. So no further computation has to be carried out in clip containing no caption.

C. *Caption Segmentation*

The caption segmentation involves separating the caption pixels from pixels in caption region located. A color based approach is used to extract the caption pixels. It is based on the idea that usually caption pixels are of same color and the color of the caption pixels is different from that of background pixels. Mostly the color of the captions will be same over different video frames. After getting the color models of the caption pixels, the probability of each pixel in the image to caption pixel can be calculated[21]. According to the probability, most of the caption pixels can be segmented from the background. Another method is that partition the RGB color space into 64 cubes. We exploit the temporal homogeneity in color of caption pixels to filter out from background pixels. Let $B_{t1}(x,y),B_{t2}(x,y).....B_{tn}(x,y)$ be the pixels of the video frame of the clip containing captions. By performing logical AND operation of the pixels we can identify the pixels of the captions and the background because the caption pixels have more population than the background pixels.

$$B_{t1}(x,y)\wedge B_{t2}(x,y)\wedge.....B_{tn}(x,y) \quad (3)$$

Thus the captions are extracted from the videos.

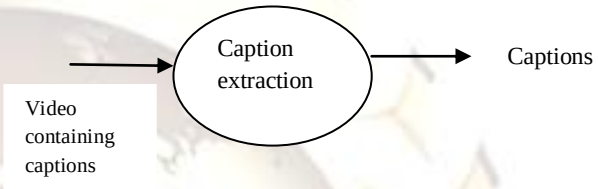
D. Refining

The captions which have extracted may have some of the background pixels or some of the caption pixels may be in the background itself. So refining of captions extracted must be carried out to get the accurate captions. The sliding window protocol is used to refine the segmentation results . we choose an 8 by 8 window as sliding window size[18]. A number of caption pixels in sliding window is checked against a threshold value. If the number of the caption pixels exceeds a threshold value, then it implies that some of background pixels are with the caption pixels so again segmentation is to carried out to get the accurate captions[20].

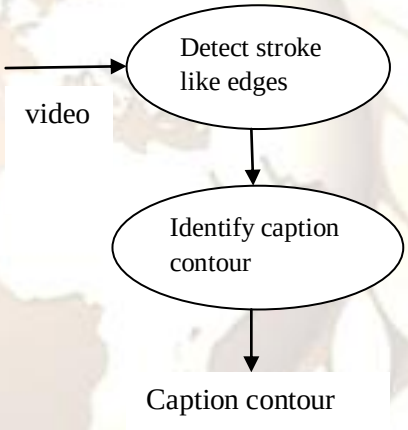
These sequential steps are performed to extract captions from video.

III. DATA FLOW DIAGRAMS

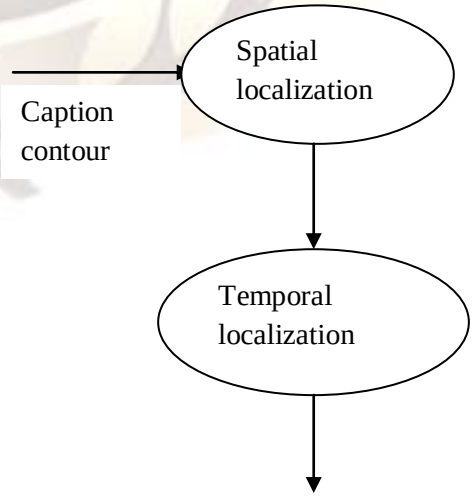
A. Context Level Dfd



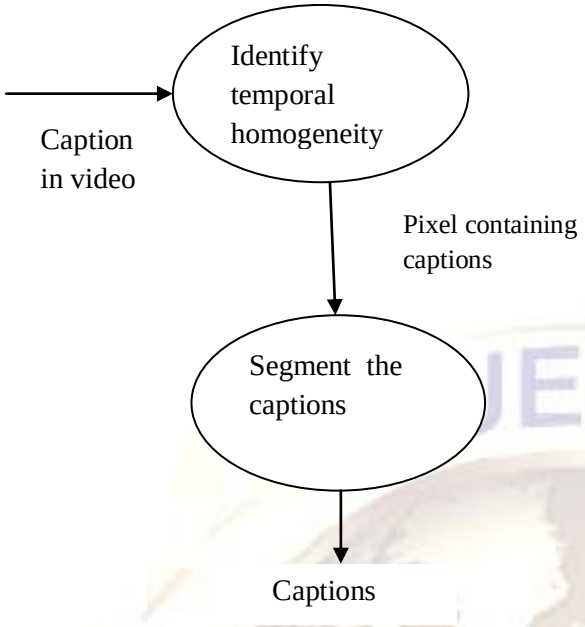
B. Level 1 Dfd Of Caption Detection



C. Level 1 DFD Of Caption Localization



D.*Level 1 DFD Of Caption Segmentation*



These are the data flow diagram of the proposed approach.

IV. EXPERIMENTAL EVALUATION

The datasets that may be chosen as movies, news, sports, talk shows and entertainment shows. In our experiment we use dataset as movies and sports video. We evaluate the performance of the proposed system based on three factors. They are based on temporal localization, spatial localization and segmentation of captions. For experiments we take 50 sample videos. We extract captions from this videos and we identify the precision and recall.

A.Temporal Localization

Evaluating the performance of temporal localization is an important factor because computations are carried out after the temporal localization .Two metrics are used for the evaluation of temporal localization. They are(1)Precision(2)Recall.

Precision→$P_{tem}=N_c/N_d$

Recall→$R_{tem=}N_c/N_g$

Where N_c denotes the number of boundaries correctly detected , N_d denotes the number of boundaries output by the system, Ng denotes the number of ground truth boundaries .The ground truth boundaries involves three cases .They are (i) from a caption to a different caption

(ii) from no caption to a caption (iii) from a caption to no caption. The comparison of two temporal localization methods in different dataset is given as

Thus the stroke filter method has more excellent performance than the other two methods with the increase of 8.6%,8.5% in P_{tem} and 0.53%,7% in $R_{tem.}$

Method	N_g	N_d	N_c	P_{tem}	R_{tem}
Texture based	1660	1730	1572	0.909	0.947
Corner based	1660	1695	1543	0.910	0.930
Proposed Method	1660	1669	1660	0.995	1.000

B.Spatial Localization

The performance of spatial localization is measured on the definition that a correct localization rectangle is counted if and only if the intersection of a located text rectangle(LTR) and a ground-truth text rectangle(GTR) covers at least 90% of their union.Recall $R_{loc}=N_{rc}/N_{tc}$ and precision $P_{loc}=N_{rc}/N_{gc}$ where

N_{rc} → number of text rectangles located correctly

N_{gc} →number of ground truth rectangle

N_{tc} →total number of rectangles located by the proposed system.The performance of the proposed system is given by calculation is

	R	P
Location	$R_{loc=40/50=0.8}$	$P_{loc=48/50=0.96}$

C. Segmentation

The failure of segmentation results occur due to two important reasons. They are (i)due to the incorrect localization that is some part of the caption is outside the bounding box or the entire part of the caption is not inside the bounding box (ii) if the background is of same color of that of pixels segmentation may result to some amount of failure. Two metrics are used for evaluation of segmentation performance .They are R_{seg} and P_{seg}. Recall $R_{seg}=N_{rp}/N_{gp}$ and precision $P_{seg}=N_{rp}/N_{tp}$ where

N_{tp} → number of text pixels segmented by the system

N_{gp} →number of text pixels ground truth

	R	P
Segmentation	$R_{seg=42/50=0.84}$	$P_{seg=46/50=0.92}$

The experimental results of our proposed method implemented in sports video is

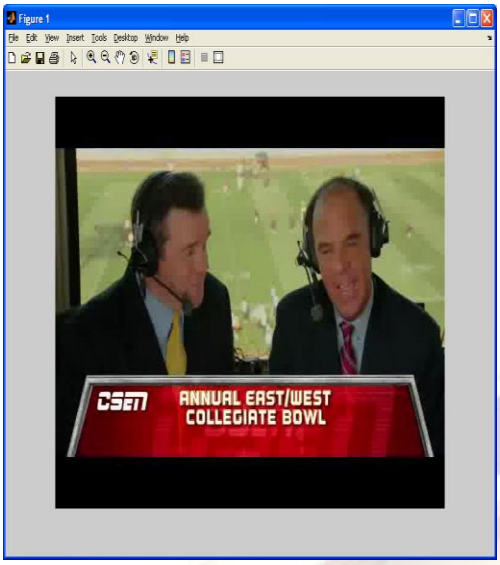


Fig. 4 input video

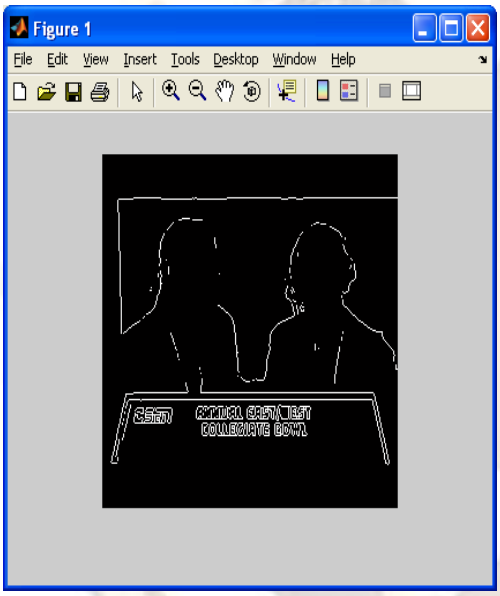


Fig .5 stroke edge detection

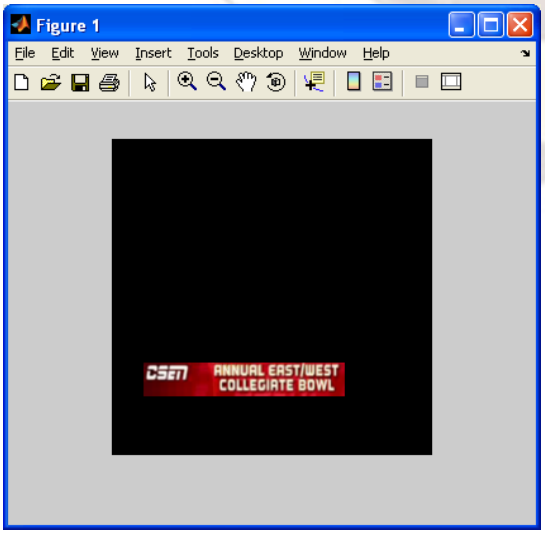


Fig .6 localization

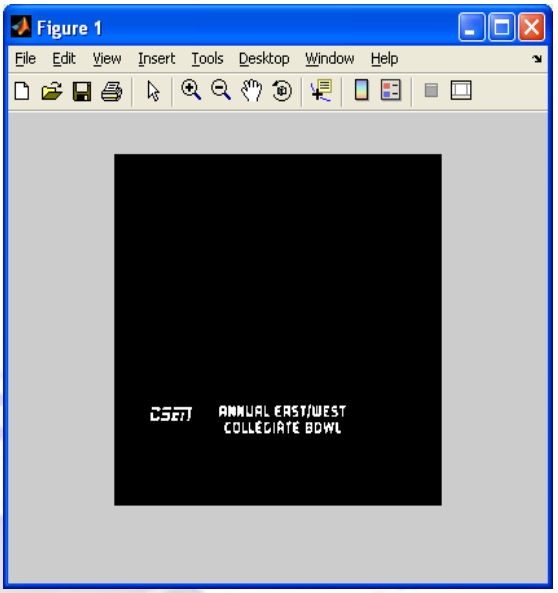


Fig .7 Extraced Captions

IV. CONCLUSION

Our proposed system performs step by step sequential procedure to extract captions from videos. This method is simple, efficient and flexible. A stroke like edge detector based on contours is used, which can effectively remove non-caption edges introduced from complex background and greatly benefit the follow-up spatial and temporal localization processes. This paper can be further developed by using optical character recognition module to identify each character in the captions. This method is used in the traffic monitoring in such a way that a video camera is located at each checkpoint which records the video of all the vehicles crossing that checkpoint .So when vehicles violate traffic rules the number of that vehicle can be identified extracting the captions(number plate) from the video recorded in the checkpoint.Thus the method is used in the traffic monitoring.

ACKNOWLEDGMENTS

The authors would like to thank ALMIGHTY GOD whose blessings have bestowed in me the will power and confidence to carry out my work

REFERENCES

[1] Xiaoqian Liu and Weiqiang Wang (April 2012) "Robustly Extracting Captions in Videos Based on Stroke-Like Edges and Spatio-Temporal Analysis ," in Proc. IEEE Transactions on multimedia, vol 14,No 2.

[2] Bertini.M, C. Colombo, and A. D. Bimbo,(Aug.2001) "Automatic captionlocalization in videos using salient points", in Proc. Int.Conf.Multimedia and Expo, pp. 68–71.

[3] Cho.J, S. Jeong, and B. Choi,(Aug. 2004) "News video retrieval using automatic indexing of korean closed-caption," Lecture Notes in Computer Science, vol. 2945, pp. 694–703.

[4] Fan.J,D.K.Y. Yau, A.K. Elmagarmid, and W.G. Aref,(2001)"Automatic Image Segmentation by Integrating Color-Edge Extraction and Seeded Region Growing", IEEE Trans. on ImageProcessing, 10(10): 1454-1466.

[5] [5] Justin miller, X. R. Chen, W. Y. Liu, and H. J. Zhong, (Sep 2001) "Automatic location of text in video frames", in Proc. ACM Workshop Multimedia :Information Retrieval, Ottawa, ON, Canada.

[6] Jagath Samarabandu and Xiaoqing Liu(2007) "An Edge-based TextRegion Extraction Algorithm for Indoor Mobile Robot Navigation", IEEE Transactions of Knowledge and DataEngineeringpp.273-280.

[7] K. C. K. Kim, H. R. Byun, Y. J. Song, Y. W. Choi, S. Y. Chi, K. K. Kim, Y. K. Chung, (Aug. 2004) "Scene text extraction in natural scene images using hierarchical feature combining and verification," in Proc. Int.Conf. Pattern Recognition, vol.2, pp. 679–682.

[8] M. R. Lyu, J. Song, and M. Cai,(Feb. 2005) "A comprehensive method for multilingual video text detection, localization, and extraction," IEEE Trans.Circuit and Systems for Video Technology, vol. 15, no. 2, pp. 243–255.

[9] S.Liao, Max W. K. Law, and Albert C. S. Chung,(May 2009)" Dominant Local Binary Patterns for Texture Classification", IEEE Transactions on Image Processing, Vol. 18, No. 5.

[10] Qixiang Ye, Wen Gao, Weiqiang Wang, Wei Zeng, (May 2009) "A Robust Text Detection Algorithm in Images and Video Frames ", IEEE Trans. Circuits Syst. Video Technol., vol. 19(5), pp. 753–759.

[11] K. I. Kim, K. Jung, and J. H. Kim, "Texture-based approach for textdetection in images using support vector machines and continuously adaptive mean shift algorithm," Pattern Anal. Mach. Intell., vol. 25,no. 12, pp. 1631–1639, 2003.

[12] C. Garcia and X. Apostolidis, "Text detection and segmentation incomplex color images," in Proc. IEEE Int. Conf. Acoustics, Speech,and Signal Processing, Istanbul, Turkey, 2000, pp. 2326–2329.

[13] A.K.Jain and S. Bhattacharjee, "Text segmentation using Gabor filtersfor automatic document processing,"Mach, Vis. Appl,, vol. 5, no. 3, pp.169–184, 1992.

[14] Lyu, M., Song, J., Cai, M., 2005. A comprehensive method for multilingual video textdetection, localization, and extraction. IEEE Trans. CSVT 15 (2), 243–255.

[15] Tupin, F., Maitre, H., Mangin, J.F., Nicolas, J.M., Pechersky, E., 1998. Detection of linear features in SAR images: Application to road network extraction. IEEE Trans. GeoScience Remote Sensing 36, 434–453.

[16] Vapnick, V., 1995. The Nature of Statistical Learning Theory. Springer.Wang, Y., Liu, Y., Huang, J.C., 2000. Multimedia content analysis using both audioand visual clues. IEEE Signal Process. Mag. 17 (6), 12–36.

[17] Wu, V., Manmatha, R., Riseman, E.M., 1997. TextFinder: An automatic system todetect and recognize text in images. IEEE Trans. PAMI 21 (11), 1224–1229.

[18] Ye, Q., Huang, Q., Gao, W., Zhao, D., 2005. Fast and robust text detection in imagesand video frames. Image Vision Comput. 23, 565–576.

[19] Zheng, Y., Li, H., Doermann, D., 2004. Machine printed text and handwritingidentification in noisy document images. IEEE Trans. PAMI 26 (3), 337–353.

[20] Canny, J.F., 1986. A computational approach to edge detection. IEEE Trans. Pattern Anal. Machine Intell. 8, 679–698.