Improving text recognition by distinguishing scene and overlay text

Bernhard Quehl, Haojin Yang, Harald Sack
Hasso Plattner Institute, Potsdam, Germany
Email: {bernhard.quehl, haojin.yang, harald.sack}@hpi.de

ABSTRACT

Video texts are closely related to the content of a video. They provide a valuable source for indexing and interpretation of video data. Text detection and recognition task in images or videos typically distinguished between overlay and scene text. Overlay text is artificially superimposed on the image at the time of editing and scene text is text captured by the recording system. Typically, OCR systems are specialized on one kind of text type. However, in video images both types of text can be found. In this paper, we propose a method to automatically distinguish between overlay and scene text to dynamically control and optimize post processing steps following text detection. Based on a feature combination a Support Vector Machine (SVM) is trained to classify scene and overlay text. We show how this distinction in overlay and scene text improves the word recognition rate. Accuracy of the proposed methods has been evaluated by using publicly available test data sets.

Keywords: Video OCR, Multimedia retrieval, Machine learning.

1. INTRODUCTION

Text in images contains valuable information about the content and is exploited in many applications, e.g. image or video labeling, context disambiguation or video indexing. Text detection and recognition in videos is the most challenging task because of complex backgrounds, variations of fonts, size, color and orientation. Various methods have been proposed for the extraction of text fragments, from a printed document, captured images, handwritten documents, Web pages and license plate detection. Up to now no solution has been developed, which works equally well for all and delivers optimal recognition results. Standard used video OCR frameworks consist of two main steps: text detection and text recognition. The text detection process determines the position of text within the video image. Overlay text is created artificially and typically embedded in a heterogeneous background, while scene text is recorded by the camera and has low contrast and possible perspective distortions, which makes both difficult to be recognized via standard OCR software. Therefore, already located text needs to be separated from interfering background with the help of appropriate binarization techniques, before applying the OCR process to enable high quality results. The main contribution of this paper is the following: The type of the text in video images is determined via Support Vector Machine (SVM) to choose the appropriate combination of pre-processing procedures. For the known text type processing steps can be controlled dynamically, such as binarization the removal of complex backgrounds or contrast sharping algorithms. The rest of the paper is structured as follows: Section 2 presents related work, Section 3 explains the architecture of the proposed system, where our approach for text detection and recognition is depicted in detail. Section 4 provides the evaluation and experimental results. Finally the paper concluded with an outlook on future work.

2. RELATED WORK

Existing text extraction technologies for images or videos are based on the combination of sophisticated pre-processing procedures for text detection and the application of traditional OCR engines [5]. For instance, overlay text usually consists of similar color with high contrast and complex backgrounds. On the other hand, scene text is characterized by difficult lighting conditions, perspective distortions, low contrast or potential covering. Numerous methods which focus either on text localization for real world images [2, 8], or for overlay text [3, 4] have been published. Being dependent on the text type these approaches are typically only optimized for one single text type. Kim et al. [3] suggest an approach based on transient colors between inserted text and its adjacent background. Then, candidate regions are extracted by reshaping, using the projection profiling in the transition map. Hua et al. [4] presented a text localization method based on image corner points and used project profiling to detect text lines candidate. Finally, a verification step based on the feature derived from edge maps is taken to reduce false positives. For overlay text these approaches work very well, but scene text has rather different characteristics, such as low contrast between text and background or geometrical distortions, where project profiling or color features do not work anymore. Although the afore-mention methods evaluate video scenes, which possibly also contain scene text, but they are focused on overlay text. Neumann et al. [8] suggested a scene text extraction method. This method uses a Maximally Stable Extremal Regions (MSER) selector, which exploits

region topology. This method is sensitive to noise and complex characters or text. Artificial text with drop shadow or outline leads to misinterpretation when using MSER and subsequent refinement steps do not work anymore. Recently, a hybrid approach has been proposed. Googles PhotoOCR from Bissacco et al. [1] describes a system for text extraction in images, which achieves best results for both text categories. They developed an end to end system that includes detection as well as recognition. However, the competitive results which the authors claim are achieved by using a 10^4 times larger training set, which is not publicly available. The system proposed in this paper uses a standard OCR application (Tesseract [11]), which is not trained on scene or overlay text. Furthermore, to the best of the authors knowledge there are currently no approaches, which explicitly differentiate scene and overlay text to adapt post-processing steps.

3. SYSTEM ARCHITECTURE AND WORKFLOW

The text extraction system proposed in this paper is based on video OCR (VOCR) system of Yang et al. [12]. An overview of the system is given in the following Sections. The existing VOCR was mainly developed with a focus on overly text in video images. In our experiments with scene text a different set of parameters as well as varying post-processing methods have to be applied to recognize these texts successfully. Unfortunately, the chosen methods or parameter combination do not work equally well for all text types. In order to solve this problem in an automatic way different configurations for VOCR have been created, such as one optimized for scene and another one for overlay text. The drawback of this approach lies in the fact that it is not known beforehand which text type really occurs in the image. As a consequence the VOCR has to be executed for each configuration. Furthermore the results of each configuration have to be fused. In the end an automatic classification of the text type and dynamic control of the processing method or parameter is preferable.

To integrate this new approach into the existing VOCR processing chain the machine learning based verification module has been adapted to classify scene text, overlay text and non-text. Previously, the SVM-based verification was only used for removing difficult to detect false positive images, which have regular structures and can't be verified by the SWT-filter. The main contribution of this paper is the classification of already determined text bounding boxes of the video images into different categories such as scene, overlay, and non-text to improve overall text detection results, as well as recognition results. Once the correct text type is determined the appropriate binarization method and further post processing parameters are chosen automatically. The following Section describes briefly the detection and recognition module from Yang et al. [12], where details which are modified in our new system are depicted in detail.

3.1 Text Detection

Text detection refers to the localization of text in an image and the generation of bounding boxes around the text. A detailed description of the text detection system is provided in [12]. A brief overview of the text detection workflow is shown in figure 1.

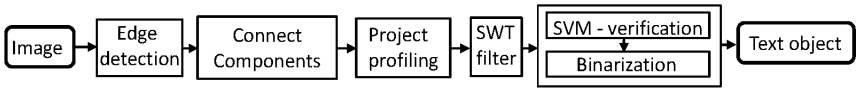

Figure 1 VOCR Workflow

A vertical edge map of the input image is computed using a Sobel-filter. Then, the morphological dilation operation is adopted to link the vertical character edges together. When generating the initial text line object, the Connect Components (CC) on the same horizontal text line with a neighbored distance smaller than their height are merged. The text line objects candidates are further processed in a subsequent text line refinement step. Finally, two verification steps are applied to remove non-text objects. In general, most false positives are sorted out by the Stroke Width Transformation (SWT)-filter. In some cases non-text regions are visually similar to edge directions and stroke width values of text and therefore might produce false positives e.g. windows, fences, railings or bricks. A machine learning based verification step is used to reduce false positives. In the approach presented here the machine learning verification and the subsequent binarization module is modified. In [12] a large number of different features have been evaluated and HOG, eLBP and image entropy features have achieved best results. We therefore also use these features as well.

3.2 Text Recognition

Detected text objects serve as input to the subsequent text recognition step. Recognition uses standard OCR software in order to extract text from bounding boxes, which sometimes fails, because technology from standard OCR is optimized for printed documents with high resolution. In contrast to print documents, video text comes with complex backgrounds and possibly low resolution or uneven illumination. In order to achieve better results the text images have to be further

refined by an optimized image binarization procedure. Local and global skeleton-based binarization methods are applied to extract video text from its background (see [12]).
Text recognition result can be further improved by using a multi-hypotheses approach, where several different thresholding methods are applied to generate binarization results. The corresponding OCR hypotheses are subsequently created by applying a standard OCR engine for each binarized image. A spell checking process is performed on every OCR string result. The final OCR result is achieved by applying heuristic constraints. In our approach, spell-checking[1] is performed as shown in figure 2.

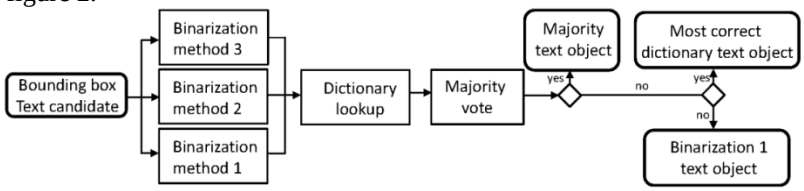

Figure 2 Multi-hypotheses workflow

If more than two hypotheses have been correctly recognized i.e. the result of the OCR string has been verified via dictionary lookup by the OCR engine, the final result is determined by the majority vote. If nothing can be verified via dictionary lookup, the hypothesis with the most valid characters is accepted, i.e. the dictionary term with less special characters inside or otherwise, the skeleton-based hypothesis as the final result is accepted.

4. SCENE AND OVERLAY CLASSIFICATIONON

The approach presented in this paper uses machine learning methods to determine the type of text found in a video frame (overlay or scene text). Depending on the text type our approach uses optimized binarization methods to remove complex background as well as type-specific pre-processing steps like contrast sharping algorithms, to improve the recognition results. In order to differentiate between scene text, overlay text and non-text a hierarchical approach is proposed, which uses two classifiers.

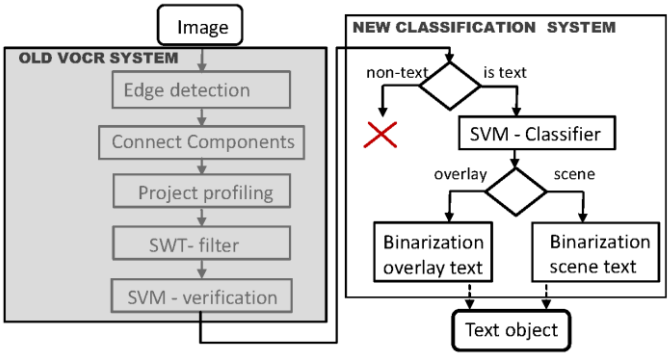

Figure 3 Right: The workflow of hierarchical approach Classification approach

Figure 3 presents the final System for video OCR as proposed in this paper. We follow a hierarchical approach that first uses the concept of the existing SVM-based verification stage the remove false positives and then applies a new SVM model to distinguish between scene and overlay text. Both classifier training strategies are evaluated in the subsequent sections.

4.1 Evaluation

Evaluation is performed on the icdar 2013 database "text localization" and "word recognition" task which provide a separate training set for scene and overlay text as well as evaluation set, respectively. Unfortunately, there is no text detection or recognition benchmark that contains mixed scene and overlay test set. Therefore, we use both sets – overlay and scene text – in a combined set in order to evaluate our initial assumption. In the following, we present and evaluate our approach that improves text detection and recognition by automatically distinguishing between both types.

[1] The open source spell checker Hunspell has been applied in the spell checking process. http://hunspell.sourceforge.net/ (last access: 06/09/2014)

4.2 Evaluation Text Detection

For evaluation of the text detection task the official metric, which is suggested in the icdar 2013 challenge [10] is used. As discussed in [12] a linear combination of HOG features with 9 bins, eLBP with 32 features and image entropy for each text line is used resulting in a 42-dimensional feature vector. For the first classifier a training set with 5076 positive (for text including scene and overlay text) and 5171 negative items (non-text) was created from icdar 2013 "text detection" training sets. Non-text examples are sampled from false positive results of the VOCR. For collecting an equal number of negative samples the false positives samples are extracted from the icdar 2013 training set images and other sources described in [12]. The Support Vector Machine (SVM) applies a RBF-kernel and all model hyperparameters have been optimized using grid search and cross validation.

Finally, we compare the detection result with and without using a SVM-verification and it turns out that F_1 measure for the text detection increased from $F_1=0.46$ to $F_1=0.55$ with using the SVM-verification.

4.3 Evaluation Text Recognition

The next step is to improve also the text recognition result of our video OCR system. We follow the hierarchical approach in Figure 3 that first uses the above mention SVM-verification stage to remove false positives and then applies a new trained SVM classifier model to distinguish between scene and overlay text. Once the correct text type is determined the appropriate binarization method is chosen automatically.

The second SVM classifier uses the same features and training model as previously mentioned, but a different training set containing only scene and overlay text samples. The overlay text is extracted from the icdar 2013 "text detection" training set using 1727 items to get equal number on scene text items the icdar 2013 "text detection" training set for scene text and the street view text (SVT) [13] is used containing 1750 items.

To estimate the recognition quality of our video OCR system the icdar 2013 "word recognition" has been executed, which provides a separate test set for scene and overlay text. In order to improve the recognition in the mixed test set appropriate binarization methods have to be applied for each text type. Therefore, we evaluated different binarization methods for scene and overlay text and optimized their respective parameters using the icdar 2013 "word recognition" benchmark (see Table 2, reported are correctly recognized word rate (WRR) as well as the edit distance [10] which is used as final ranking measurement).

		Overlay text			**Scene Text**		
	Methods	parameter	Edit Distance	WRR	parameter	Edit Distance	WRR
I.	*Skeleton local [12]*	*ß= 60*	*324*	*0.64148*	*ß =80*	*459*	*0.47838*
II.	*Skeleton global [12]*	*ß = -31*	*338*	*0.62196*	*ß=-20*	*436*	*0.47654*
III.	*Otsu [9]*	-	*381*	*0.60042*	-	*449*	*0.49601*
IV.	*Sauvola [10]*	*k=0.02*	*419*	*0.55316*	*k=-0.019*	*520*	*0.41950*
V.	*Tesseract [11]*	-	*375*	*0.60945*	-	*471*	*0.47612*
VI.	*Multi hypothesis[12]*	*Skeleton local, Otsu, Tesseract*	***244***	*0.72202*	*Skeleton global, Otsu, Tesseract*	***411***	*0.57313*

Table 1 Parameter optimization from different binarization methods

Table 1 shows the results of different binarization methods optimized for each text type. The methods *I-V* are based on single method and multi-hypothese (*VI*) based on this single methods, which is discussed in Section 3 (see 3.2-Text Recognition). The best recognition results have been achieved with the multi-hypotheses approach. The multi-hypotheses engine for the overlay text is based on local skeleton, otsu thresholding and tesseract-ocr engine whereas for the scene text it is based on global skeleton, otsu thresholding and the sharpened original image for tesseract-ocr engine. Furthermore to increase the edit distance several pre-processing steps are conducted that increase the resolution of small images (less than 30 pixels) by factor 2. Subsequently, the rand noises (artifacts from cutting see [10]) are removed for global skeleton and otsu thresholding by applying a horizontal-vertical projection profile refinement method (see [12]). The pre-processed image is classified in overlay or scene text and then sent to the corresponding multi-hypotheses engine.

Finally, scene and overlay texts of the icdar 2013 benchmark are merged to evaluate our system using edit distance and WWR.

Methods	Edit Distance	WRR
no Classifier	*766*	*0.612662*
with Classifier	*679*	*0.652811*
reference	*674*	*0.655582*

Table 2 Evaluation distinguishing between scene and overlay text

Table 2 shows the results obtained using the proposed classification scheme with classifier as well as using no differentiation of different text types (*no classifier*). It can be observed that the proposed approach improves in terms of edit distance as well as WRR. Compared with a manual classification of scene and overlay text (*reference*). Our proposed automatic classification approach comes quite close.

5. CONCLUSION AND OUTLOOK

In this paper we have presented a new approach for video OCR that automatically distinguishes between overlay and scene text by adapted text detection and recognition stages. Based on our experimental results we have also evaluated a sequential approach, which uses a multi-classifier to classify scene, overlay and non-text at once. However, the hierarchical approach outperforms the sequential approach. Furthermore, we have given evidence that the classification between scene and overlay text actually improves the quality of the recognition results in terms of edit distance and word recognition rate. Many applications are conceivable for automatic text type classification. For instance, apart from being used for choosing the best binarization method, it can also be used for selecting the best suited OCR system. Our future work will focus on integration of new binarization methods for the different kind of text. Furthermore, we intend to differentiate more types of text. For instance, text can also be classified into handwritten characters, text from presentation slides, or printed document text. Another possible extension could target language classification in order to set the correct language model for an OCR system. Obviously, an OCR with an English language model would not generate meaningful result on Chinese or Arabic letters.

REFERENCES

[1] Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: Photoocr: Reading text in uncontrolled conditions. Computer Vision, IEEE International Conference 785-792 (2013).

[2] Epshtein, B., Ofek, E., Wexler, Y.: Detecting text in natural scenes with stroke width transform. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pp. 2963-2970. IEEE (2010).

[3] Kim, W., Kim, C.: A new approach for overlay text detection and extraction from complex video scene. Trans. Img. Proc. 18, 401-411 (2009).

[4] Hua, X.S., Yin, P., Zhang, H.J.: Efficient video text recognition using multiple frame integration. In: 2002 International Conference on Image Processing (ICIP2002, pp. 22-25 (2002)

[5] Jung, K., Kim, K.I., k. Jain, A.: Text information extraction in images and video: a survey. Pattern Recognition 37(5), 977-997 (2004)

[6] Karatzas, D., Shafait, F., Uchida, S., Iwamura, M., i Bigorda, L.G., Mestre, S.R., Mas, J., Mota, D.F., Almazn, J., de las Heras, L.P.: Icdar 2013 robust reading competition. In: ICDAR, pp. 1484-1493. IEEE (2013).

[7] Lienhart, R., Wernicke, A.: Localizing and segmenting text in images and videos. IEEE Transactions on Circuits and Systems for Video Technology 12(4), 256-268 (2002)

[8] Neumann, L., Matas, J.: Text localization in real-world images using efficently pruned exhaustive search. In: Document Analysis and Recognition (ICDAR), 2011 International Conference on, pp. 687-691. IEEE Computer Society Conference Publishing Services, IEEE Computer Society Offices, 2001 L Street N.W., Suite 700 Washington, DC 20036-4928, United States (2011).

[9] Otsu, N.: A threshold selection method from gray level histogram. IEEE Transactions on System, Man, Cybernetics 19(1), 62-66 (1978).

[10] Sauvola, J., Pietikainen, M.: Adaptive document image binarization. Pattern Recognition 33(2), 225-236 (2000).

[11] Smith, R.: An Overview of the Tesseract OCR Engine. In: ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2, pp. 629-633. IEEE Computer Society, Washington, DC, USA (2007).

[12] Yang, H., Quehl, B., Sack, H.: A framework for improved video text detection and recognition. Multimedia Tools and Applications pp. 1-29 (2012).

[13] K. Wang, B. Babenko, and S. Belongie. End-to-end scene text recognition. In ICCV, 2011.