Towards Automatic Video-based Whiteboard Reading

Markus Wienecke Gernot A. Fink Gerhard Sagerer
Bielefeld University, Faculty of Technology
33594 Bielefeld, Germany
E-mail: mwieneck@techfak.uni-bielefeld.de

Abstract

As whiteboards have become a popular tool in meeting rooms, there has been a growing interest in making use of the whiteboard as a user interface for human computer interaction. Therefore, systems based on electronic whiteboards have been developed in order to serve as meeting assistants for e.g. collaborative working. However, as special pens and erasers are required, the natural interaction is restricted. In order to render this communication method more natural it was proposed to retain ordinary whiteboard and pens and to visually observe the writing process using a video camera [11, 9]. In this paper a prototype system for automatic video-based whiteboard reading is presented. The system is designed for recognizing unconstrained handwritten text and is further characterized by an incremental processing strategy in order to facilitate recognizing portions of text as soon as they have been written on the board. We will present the methods employed for extracting text regions, pre-processing, feature extraction, and statistical modeling and recognition. Evaluation results on a writer independent unconstrained handwriting recognition task demonstrate the feasibility of the proposed approach.

1 Introduction

Whiteboards have become a very popular tool not only for presentations and educational purposes but also in meeting rooms for the exchange of ideas during group discussions, for project planning, system design, etc. The increasing popularity of whiteboards as natural means of expression is mainly due to their ease of use. They provide a large writing space and any information that is no longer needed can be easily erased.

In recent years systems have been developed that make use of the whiteboard as user interface for human computer interaction. Typically, these systems employ electronic whiteboards which are able to detect the position of the pen in order to allow the construction of an electronic version of the image using a computer (cf. e.g. [4]). Additionally, the pen trajectory can be interpreted by an on-line recognition module to automatically recognize what was written on the board.

However, electronic whiteboards exhibit some disadvantages. As special pens and erasers are necessary, the natural interaction is restricted. For example, the electronic whiteboard does not notice erasing parts of what is written, if the writer uses a finger or paper towel instead of the eraser provided. Therefore, a promising alternative might be to retain ordinary whiteboard and pens and to observe the writing process using a video camera (cf. e.g. [11, 9]).

In this paper a prototype system for automatic video-based whiteboard reading is presented. In contrast to the approaches proposed in [11, 1] which only permit the recognition of a limited set of symbols, our system is designed for recognizing handwritten text. As the pen is rarely visible in the image and thus online recognition based on the pen trajectory is not feasible the proposed system is characterized by an incremental offline recognition approach. Thus, the writing process is continuously observed and recognition starts automatically as soon as a region of handwritten text is visible in the image. Besides saving processing time as only small portions of text are recognized at each time, this approach also allows obtaining a rough estimation of the time-structure of the handwriting process. This is an important prerequisite for systems for meeting assistance and e-learning as the handwriting can be related to the gestures or the speech of the user.

In the following section we will give a short review of relevant related work. The architecture of the proposed system is presented in section 3. Afterwards we describe in subsequent sections the methods for text extraction, preprocessing, feature extraction, and statistical modelling and recognition. Evaluation results will be presented in section 8 in order to demonstrate the effectiveness of the proposed approach.

[1]This work was in part supported by the German Research Foundation (DFG) within project **Fi799/1**.

In Proc. Int. Conf. on Document Analysis and Recognition, pages 87-91, Edinburgh, Scotland, 2003. IEEE.

2 Related Work

In the field of human computer interaction natural input modalities like speech, handwriting, and gestures have been extensively studied in recent years. Thus, as whiteboards are popular tools in meetig rooms and the computing machinery has become increasingly powerful, there has been a growing interest in making use of the whiteboard as a user interface for human computer interaction. Therefore, systems have been developed in order to serve as meeting assistants for e.g. collaborative working. Usually, such systems are based on electronic whiteboards and special pens and erasers, which are capable of detecting the current pen positon during writing in order to construct an electronic version of the image in the computer for further processing (cf. e.g. [4]).

In contrast to electronic whiteboards where the pen trajectory is directly recorded, systems based on visual input first have to detect pen movements or relevant image regions in the image sequence. One approach is to use a special marker that has a distinctive color that makes it easy to track [1]. Thus, a temporal trajectory is obtained that can be recognized using on-line methods. A contrary approach is to extract and analyze the relevant image regions after the writing process has finished [11, 9].

As the system presented in this paper is not restricted to a small set of commands but is designed for recognizing unconstrained handwritten text the approach is not only closely related to the task of locating text in image sequences but also to off-line handwriting recognition. Whereas the problem of text detection in image sequences was so far mostly studied for machine printed text [2][6] there has been a lot of work in the field of off-line handwriting recognition. See e.g. [12] for an extensive survey. In order to avoid errors introduced by segmenting the text into words at an early stage, segmentation-free methods based on Hidden Markov Models (HMMs) are successfully applied for handwriting recognition. For example, an advanced system for writer-independent unconstrained text recognition which is also tested on a large database [7] produced by several hundreds of writers can be found in [8].

3 System Architecture

The system presented in this paper continuously observes the writing process and automatically starts recognizing as soon as a line of handwritten text is visible. Therefore, the incremental processing strategy depicted in figure 1 is proposed. Thus, after grabbing the image, all text regions currently visible are extracted. In order to avoid recognizing the same text region multiple times in the image sequence we employ a region memory consisting of all the different text regions extracted so far. If a new, not yet memorized text region is found, several pre-processing steps are applied to compensate for the highly varying background intensity and to normalize the handwriting. After that, features are extracted using a sliding window approach which are finally fed into a statistical recognition module based on Hidden Markov Models (HMMs).

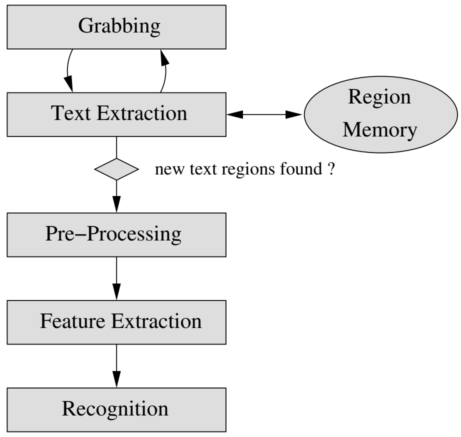

Figure 1. System architecture

4 Extracting Text Regions

The image sequences required for whiteboard reading are captured using a standard video camera working in interlaced PAL mode thus obtaining images with a resolution of 756×576 pixels. The camera is positioned approximately 3 to 4 meters in front of the whiteboard, and the observed writing space covers an area of approximately 70×50 cm. Thus, the effective resolution is less than 30 dpi, which is about ten times smaller compared to scanned documents mostly used for handwriting recognition.

As the extraction of the handwritten text regions in the image sequence is an essential module for further processing some requirements have to be met. First of all, the extraction of the text regions has to be robust with regard to noisy images, i.e. non-text regions that belong to the writer or caused by non-uniform lighting should be suppressed. Additionally, the text extraction has to avoid splitting up words in order to facilitate lexicon-based recognition. Another important requirement is a short processing time that keeps the delays between writing and recognition as short as possible. In order to satisfy these conditions a two-step approach for robust and fast extraction of handwritten text regions is used (figure 2).

Firstly, the gray-scale image is divided into blocks of equal size (40×40 pixels) on which a feature vector is calculated for discriminating text from noise. It is assumed that text blocks can be identified using the following characteristics: They contain contour pixels caused by the pen strokes,

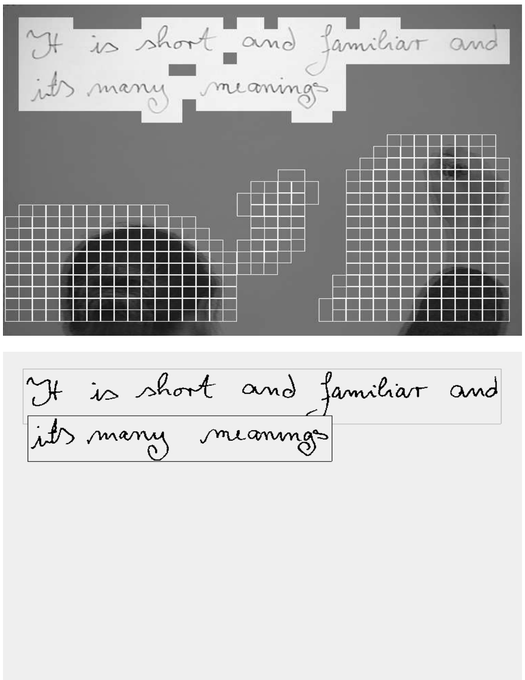

Figure 2. Text extraction: The blocks corresponding to text (light gray), background (dark gray), and noise (framed) are shown above, the extracted text regions below.

show an average pixel intensity similar to the empty whiteboard, and, compared to noise regions, remain relatively stable over time. Therefore, the feature vector consists of (1) the average pixel intensity of the block, (2) the average difference of pixel intensity between two consecutive images, and (3) the number of edge pixels – determined using a Canny edge detector. Thresholds are then applied in order to decide wether the image block is part of a text region, an 'empty' whiteboard region, or noise.

After that, the image blocks which are assumed to contain text and are sufficiently far away from noise regions as well are binarized and the connected components corresponding to ink strokes are calculated. The connected components of all text blocks are then clustered based on the distances between their bounding boxes in order to obtain regions containing the handwritten text.

In order to avoid recognizing the same text region repeatedly in consecutive images a region memory is employed that contains all regions recognized so far. Thus, in each time-step the extracted regions are compared with the regions stored in the memory. If the similarity (based on the connected components) exceeds a threshold it is assumed that the newly extracted region has already been recognized previously and, therefore, will not be recognized again. The extracted text regions that cannot be found in the memory are then used for further pre-processing.

Besides increasing the processing speed, the region memory also permits the handling of corrections – an important feature for human-computer interaction. The handling of corrections is accomplished by permanently comparing the region memory with the current image. If, for example, the writer wipes out something, the region previously stored in the memory disappeares in the image and therefore will be also deleted from the memory.

5 Preprocessing

It is a well-known fact that the normalization of handwriting has a great impact on recognition accuracy. This particularly applies to the task of video based whiteboard reading which is much harder than reading scanned documents. One reason is that the illumination conditions often cause severe difficulties. As no specialized lighting is employed the background intensity of the image is highly varying so that no global threshold exists for discriminating between the foreground and the background. Therefore, a modified version of Niblack's binarization method presented in [14] is used for determing thresholds locally. These local thresholds are afterwards applied for normalizing the intensity information.

Another difficulty of whiteboard reading is the distortion of the handwriting. The lack of any reference lines together with the circumstance that subjects are often not very familiar with writing on boards results in distorted patterns of handwriting that often show baseline drifts and huge variations of the corpus height within one line. Motivated by this observation, the vertical position, skew, and slant of each text region are corrected locally (figure 3b). Thus, after separating individual lines of handwriting within the extracted region using projection histograms, each line is split by searching for white-spaces between the segments of handwriting. A threshold is used to avoid segments that are too short for calculating reliable normalization factors.

Whereas the vertical position and skew is corrected by applying a baseline estimation method using linear regression similar to the approach described in [10], the method for slant correction is based on edge orientation. The Canny edge detector is applied in order to obtain the edge orientation data which is accumulated in an angle histogram. The mean of the histogram is then used as slant angle.

In order to normalize the size of the handwriting, we count the number of local extrema of each handwritten line and put this number in relation to the width of the line. The scaling factor is based linearly on this relation, because the larger this relation the narrower the writing-style.

The last pre-processing step concerns a more precise estimation of the lower and upper baseline (figure 3c). We observed that the straight lines calculated for skew correction by linear regression are often only very rough approxima-

a)
b)
c)

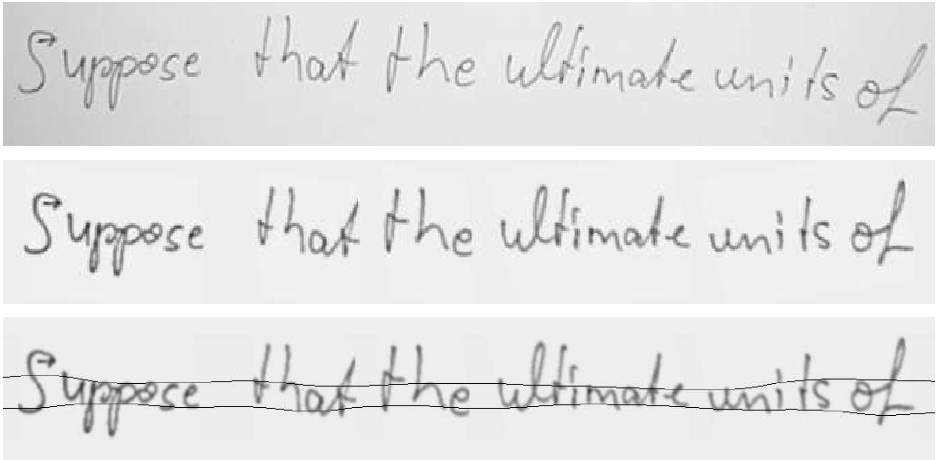

Figure 3. Pre-processing: from top to bottom: a) raw image region, b) intensity, slope, and slant correction, c) estimated baselines

tions. However, a couple of features used for recognition depend on the position of the baselines. Therefore, a cubic spline is calculated for approximating the lower baseline using the local minima of the handwriting. In order to avoid considering local minima belonging to descenders a threshold is applied depending on the distance of the minimum to the straight baseline estimated by linear regression. Similarly, the upper baseline (corpus height) is approximated by a cubic spline using the local maxima of the handwriting.

6 Feature Extraction

The pre-processed images are used as input data for the feature extraction step. A sliding window technique is applied similar to the approach described in [8]. In our case, a window of the image's height and four columns width is moved with an overlap of two columns from left-to-right over the image and several geometrical features are extracted.

For each column of the sliding window the following seven features are calculated: (1) Number of black-to-white transitions (after a binarization of the windowed text image), (2) position of the mean value of the intensity distribution with respect to the baseline (3) distance from the uppermost text pixel to the baseline, (4) distance from the lowermost text pixel to the baseline, (5) distance of uppermost and lowermost text pixel, (6) average intensity between uppermost and lowermost text pixel, (7) average intensity of the column. The features (2)-(5) are normalized by the core-size, i.e. the distance of lower and upper baseline, in order to increase the robustness against variations in writing-size. Afterwards, all features are averaged over the four columns of the window.

In order to consider the direction of the lower and upper contour as well as the gradient of the mean value of the intensity distribution, we additionally calculate three directional features. Therefore, we estimate lines through the four lower contour points, upper contour points, and mean values within the sliding window, and use the line orientations as features (8), (9), and (10), respectively.

For considering a wider temporal context, we additionally compute an approximate horizontal derivative for each component of the feature vector, so that a 20 dimensional feature vector is obtained (10 features per window + 10 derivatives). In order to decorrelate the feature vectors and to improve the class separability we integrate linear discriminant analysis (LDA) in the training and recognition phase (cf. [3]). The original feature representation is optimized by applying a linear transformation and simultaneously reducing the dimensionality of the feature space from 20 to 12 dimensions.

7 Statistical Modeling & Recognition

For the design of statistical recognition systems the availability of a sufficiently large database of training samples is an important prerequisite. Ideally, for a video-based system it would be desirable to obtain a large amount of image data recorded while observing a subject writing on the whiteboard. However, recording and labeling of such video data requires a substantial manual effort. Therefore, we decided to apply our recognition system presented in [13], that was trained and tested on the IAM-database of scanned documents [7].

Configuration, training, and decoding of Hidden-Markov Models (HMMs) for the handwritten text recognition task is carried out in the framework of the methods and tools provided by the ESMERALDA development environment [5]. As general setup we use semi-continuous HMMs with a shared codebook containing 2048 Gaussian mixtures with diagonal covariance matrices. Models for 52 letters, ten numbers, twelve punctuation marks and brackets, and one white space character are trained using standard Baum-Welch reestimation. For decoding of the recognition model standard Viterbi beam-search is used.

8 Results & Discussion

In order to evaluate the proposed system we asked six subjects to write portions of text taken from the test set used in [13] on the whiteboard. No constraints with respect to the writing style were given. In contrast to the training patterns resulting from scanned forms, where rulers on a second sheet put below were used to align the baseline horizontally, the video-based data often show baseline drifts and variations of the corpus height.

The test set thus obtained contains a total of 527 words corresponding to 67 lines of text. Using this test set we at first investigated whether the sequence of extracted text regions corresponds to the chronological order in which the text was written. We observed that only 9% of the extracted text regions were in the wrong chronological order.

For evaluating the recognition accuracy we conducted two experiments: lexicon-free recognition using a tri-gram language model on the character level and word-based recognition using a lexicon containing 400 words. For lexicon-free recognition we obtained a character error rate of 29.5%, where the tri-gram language model has a perplexity of 7.9 on the test set. By applying a lexicon in the second experiment a word error rate of 37.6% was achieved.

In our opinion these results are quite encouraging, because the task is challenging in several aspects: it is writer independent, handwriting is unconstrained, and there exists a mismatch between training and testing conditions caused by the reduced resolution, varying illumination, and, particularly, the larger variations in writing style. In order to improve the recognition accuracy we will investigate adaptation techniques dealing with mismatched training and testing conditions in the near future.

9 Conclusion

We presented a system for automatic whiteboard reading based on visual input. It is characterized by an incremental processing strategy, i.e. the text lines are extracted as soon as they are visible in the image. The pre-processing and feature extraction methods applied generate a data representation which is to a certain extent robust against variations concerning the writing style and the reduced quality of the video-based data. Evaluation results on a writer independent unconstrained handwriting recognition task were presented that demonstrate the feasibility of the proposed approach.

References

[1] M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions. In H. Burkhardt and B. Neumann, editors, *European Conf. on Computer Vision*, volume 1406 of *LNCS-Series*, pages 909–924, Freiburg, Germany, 1998. Springer-Verlag.

[2] P. Clark and M. Mirmehdi. Recognising text in real scenes. *Int. Journal on Document Analysis and Recognition*, 4:243–257, 2002.

[3] J. G. A. Dolfing and R. Haeb-Umbach. Signal representations for Hidden Markov Model based on-line handwriting recognition. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume IV, pages 3385–3388, München, 1997.

[4] S. Elrod, R. Bruce, R. Gold, D. Goldberg, F. Halasz, W. Janssen, D. Lee, K. McCall, E. Pedersen, K. Pier, J. Tang, and B. Welch. Liveboard: A large interactive display supporting group meetings, presentations and remote collaboration. In *Proceedings of ACM CHI'92 Conference*, pages 599–607, May 1992.

[5] G. A. Fink. Developing HMM-based recognizers with ESMERALDA. In V. Matoušek, P. Mautner, J. Ocelíková, and P. Sojka, editors, *Lecture Notes in Artificial Intelligence*, volume 1692, pages 229–234, Berlin Heidelberg, 1999. Springer.

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

[7] U.-V. Marti and H. Bunke. A full english sentence database for off-line handwriting recognition. In *Proc. Int. Conf. on Document Analysis and Recognition*, pages 705–708, Bangalore, 1999.

[8] U.-V. Marti and H. Bunke. Handwritten sentence recognition. In *Proc. Int. Conf. on Pattern Recognition*, volume 3, pages 467–470, Barcelona, 2000.

[9] E. Saund. Bringing the marks on a whiteboard to electronic life. In *Proceedings of CoBuild'99. Second International Workshop on Cooperative Buildings*, pages 69–78, Pittsburgh, 1999. Springer.

[10] A. W. Senior and A. J. Robinson. An off-line cursive handwriting recognition system. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(3):309–321, 1998.

[11] Q. Stafford-Fraser and P. Robinson. Brightboard: A video-augmented environment. In *CHI*, pages 134–141, 1996.

[12] T. Steinherz, E. Rivlin, and N. Intrator. Offline cursive script word recognition – A survey. *Int. Journal on Document Analysis and Recognition*, 2(2):90–110, 1999.

[13] M. Wienecke, G. A. Fink, and G. Sagerer. Experiments in unconstrained offline handwritten text recognition. In *Proc. 8th Int. Workshop on Frontiers in Handwriting Recognition*, Ontario, Canada, August 2002. IEEE.

[14] Z. Zhang and C. Tan. Restoration of images scanned from thick bound documents. In *In Proceedings of International Conference on Image Processing*, Thessaloniki, Greece, October 2001.