A Camera-Based Interactive Whiteboard Reading System

Szilárd Vajda, Leonard Rothacker, Gernot A. Fink
Department of Computer Science
TU Dortmund
Dortmund, Germany
{szilard.vajda,leonard.rothacker,gernot.fink}@udo.edu

***Abstract*—The recognition of mind maps written on a whiteboard is a challenging task due to the unconstrained handwritten text and the different graphical elements — i.e. lines, circles and arrows — available for drawing a mind map. In this paper we propose a prototype system to recognize and visualize such mind maps written on a whiteboard. After the image acquisition by a camera, a binarization process is performed, and the different connected components are extracted. Without presuming any prior knowledge about the document, its style, layout, etc., the analysis starts with connected components, labeling them as text, lines, circles or arrows based on a neural network classifier trained on some statistical features extracted from the components. Once the text patches are identified, word detection is performed, modeling the text patches by their gravity centers and grouping them into possible words by a density based clustering. Finally, the grouped connected components are recognized by a Hidden Markov Model based recognizer. The paper also presents a software tool integrating all these processing stages, allowing a digital transcription of the mind map and the interaction between the user, the mind map, and the whiteboard.**

***Index Terms*—whiteboard reading; unconstrained document layout analysis; growing neural gas modeling; handwriting recognition;**

Figure 1: Scene from a mindmap creation process around the idea of ”Whiteboard reading”.

I. Introduction

Nowadays, in the field of handwriting recognition the focus is shifted from classical topics like bank checks or postal documents recognition [1] to more challenging topics like historical documents recognition, personal memos or sketch interpretation [2] and lately to recognition of unconstrained whiteboard notes [3], [4]. The later is in the focus of the attention because it deals with unconstrained type of documents with no specific writing style, layout, etc.

Doing collaborative work (e.g. brainstormings, discussions, presentations) is quite common in corporate or academical environments. However, there is just a limited amount of work [3], [4], [5] to embed this whiteboard outcome in a smart environment scenario (e.g. a conference room). To provide not just a digital capture of the whiteboard, but also the recognition of that content in an interactive software framework, is one of the final goals of such a smart room.

Instead of tackling this issue by some specific (sometimes costly) hardware (e.g. special whiteboard, several cameras, pen, wireless microphone proposed by the e-Learning system [5]), we propose a system which uses only regular hardware (a simple whiteboard, markers, a low-resolution active camera and a projector) available in each common conference room scenario. Such commonly used hardware setup provides us a natural environment to actively support the collaborative mind mapping [6], allowing the users to keep their old habits writing down their ideas using just the whiteboard markers without bothering about some special equipment. The focus is rather on the content and not on the form. Such a mind map creation process is depicted in Fig. 1.

The current system focuses on two main aspects. First, we will present the system capable to recognize on-line the different text and non-text components and secondly, we will concentrate on the digital outcome of that recognition process: a digital, editable mind map framework and the interaction between the static whiteboard content, the user and the projected and already recognized mind map. Such an interaction is missing from the currently available systems.

The following sections of the paper are organized as follows. Related works concerning the whiteboard recognition will be discussed in the next section. Section III describes in detail the complete whiteboard reading system. Section IV is completely dedicated to the data and the experimental setup. Finally, Section V summarizes and highlights the strengths of the presented reading system.

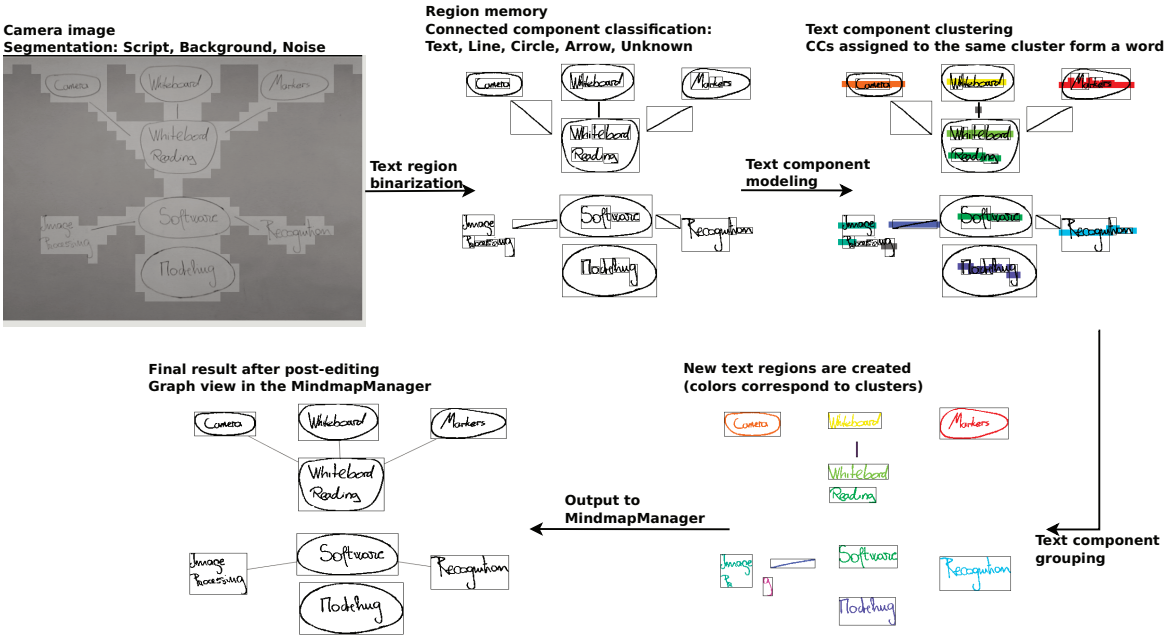


Figure 2: Overview of the proposed whiteboard reading system.

II. Related Work

Over the last two decades an impressive progress has been achieved in the field of handwriting recognition, even for large vocabularies [7] and multi-writer scenarios. However, the main constraint was always the same. To produce those sound results clean and well segmented data was necessary. In the whiteboard reading scenario addressed in this paper such presumptions can not hold. In particular, for collaborative works like mind mapping different layouts, writers, styles and text and non-text mixture should be handled.

The very first attempt to handle such a challenging task was addressed by Wienecke et al. [8], where complete handwritten sentences were recognized using a low-resolution camera. The proposed prototype system was quite promising, however it was able to recognize text only.

A similar type of research was conducted in [9], [10] where the data acquisition was performed on-line using an infrared sensor for tracking special pen. Even though the results are sound, it should be mentioned that only clear, well structured text lines were recognized.

To recognize Japanese characters on a whiteboard, the authors in [11] consider a complex hardware scenario including two cameras and a special pen to capture the writing. The text detection is not anymore performed by the system but rather by the software provided by the pen manufacturer. The system is used in an e-Learning scenario.

In a more recent work [4] of ours, we focused on a similar task, recognizing well-structured handwritten whiteboard paragraphs, considering only a camera and no on-line information. The results are really promising, however, the text detection on the whiteboard is based on some connected component (CC) estimation which is very rigid due to the usage of some thresholds.

Finally, in our recent work [3] we addressed the mind map recognition scenario (see Fig. 1), where beside the text components graphical elements like lines, circles, arrows were detected and a totally unconstrained document layout was analyzed. The goal of the current work is to improve that system by adapting the recognition to the different changing layouts, reconstruct the mind map and introduce a certain type of interaction between the user, the document and the whiteboard.

III. Whiteboard Reading System

In this section we concentrate on the whiteboard reading system, describing the complete process starting from the image acquisition, throughout the different processing steps and finally the recognition and the user interaction with the system. A system overview with its particular processing stages can be seen in Fig. 2.

A. Image acquisition and camera-projector calibration

For image acquisition a camera- and for user interaction a projector must be directed at the whiteboard. The camera is capturing the whole mind map creation process. Low-resolution gray level camera images are used for further processing (see Section III-B).

The camera-projector calibration is needed to project content to the whiteboard that is derived from the camera image. In order to be able to project information on the whiteboard for user interaction (see Section III-G), a mapping between the camera- and the projection image coordinate systems has to be obtained. The projection image contains the additional

user information and is shown on the whiteboard using the projector. This way the projection image can be seen as an overlay to the mind map drawn with a marker by the user.

For calibration a chessboard is projected on the whiteboard that is captured with the camera. Because the chessboard is rendered in the projection image, its chessboard corner coordinates are known in the projection image coordinate system. By finding chessboard corners in the camera image correspondences between both images are obtained. Finally, a homography can be estimated that maps each point from the camera coordinate system to the projection image coordinate system.

B. Image segmentation

The purpose of image segmentation is to separate elements written on the whiteboard with a marker from the whiteboard background and noisy image parts. Noisy parts are for example regions where the user stands (see Fig. 1). Afterwards the regions containing written content are further segmented by categorizing them into different mind map elements (text, line, circle, arrow).

1) Segmentation of the camera image: After image acquisition the objective is to only extract content written on the whiteboard. This relevant information is then added to a binary region memory (also refer to Fig. 1). The region memory represents the current state of written content on the whiteboard and is robust to irrelevant changes in the camera image, like illumination or in particular users standing in front of the whiteboard. Therefore the general assumption is that the camera image does not contain anything but the interior of the whiteboard and the camera or the whiteboard are not moved. In this scenario the system has to handle images that can consist of three different regions, namely:

- text (indicated by bright blocks in Fig. 1).
- background (indicated by dark blocks in Fig. 1).
- noise (indicated by blocks with grid pattern in Fig. 1).

As proposed by [8], segmentation is not done on pixel but on block level. The image is therefore divided into two layers of overlapping blocks. Each block is now segmented into one of the formerly mentioned categories on the basis of three features: gradients, gray level and changes between two consecutive images.

After categorizing all blocks the region memory can be updated.

Noise blocks are discarded because the whiteboard is potentially occluded at those locations. To be even more robust also blocks in a noise block's local neighborhood can be discarded. The occurrence of eventually appearing parts of the user's shape in the region memory can be minimized this way. The information contained in a falsely discarded block will simply be added to the region memory later.

Background blocks do not contain any written content, so the corresponding regions in the region memory can be erased.

Finally text blocks are binarized with the local Niblack method [12] and inserted into the region memory if their XOR errors with the region memory exceed a certain empirically selected threshold. This way the memory does not get updated for very small changes in the camera image but only if there is a modification to the written content. Those small changes are likely to be caused by illumination changes in the conference room. For further details please refer to [8].

The result as depicted in Fig. 2 consists of a binary representation of the whiteboard content and can be used for further processing.

2) Segmentation of the whiteboard image: A key issue to success is the accurate segmentation of the whiteboard content. We separate the whiteboard from the rest of the scene (see Section III-B1) but we do not have any prior information about the content itself. To recognize and reconstruct the mind map we need to separate text elements from non-text items, namely in such scenario, lines, circles and arrows. The detection process is based on connected components (CC) extracted from the binarized image. Working with CC is suitable in such scenarios as the components are easy to extract and no specific prior knowledge is necessary.

Instead of using heuristics rooting from the work of Fletcher and Kasturi [13] we propose a solution to classify CCs based on statistical learning. A descriptor of 12 components (i.e. contrast, edge density, homogeneity, etc.) is extracted from each CC and a multi-layer perceptron is meant to classify the pixel patches into text, line circle and arrow. For more details, please refer to [3]. This text component detector is suitable not only for Roman scripts but also for Chinese, Arabic or Bangla where even more complex characters shapes will occur.

C. Layout analysis

The layout analysis of a document consists of identifying the baseline elements composing the document and their spatial relationship among each other. While for printed documents a certain type of regularities like font type, font size, line structures, etc. can be expected, in a handwritten mind map document none of these is to be identified, hence the layout analysis is more challenging in such unconstrained handwritten document scenarios.

1) Layout modeling: As described above we first separate text items from non-text items. For further processing we will concentrate our effort to model only the different text patches. The lines, circles and arrows detected previously will serve to build the digital representation of the mind map into the so called "MindMap Manager" discussed later in Section III-F.

Our proposition is to adapt the model to the analyzed document considering the gravity centers of the text CCs (see Fig. 2a) and model the structure of the text patches (CCs) throughout these points. For each text component, the gravity center is calculated and at the left and right side of the bounding box a new center is calculated inheriting the height from the original gravity center. For larger components, exceeding the **average width**, estimated over all connected components from the document more points are derived. For each slice of **average width**/4 of the CC, a new gravity center is computer w.r.t. the pixels lying in that window.

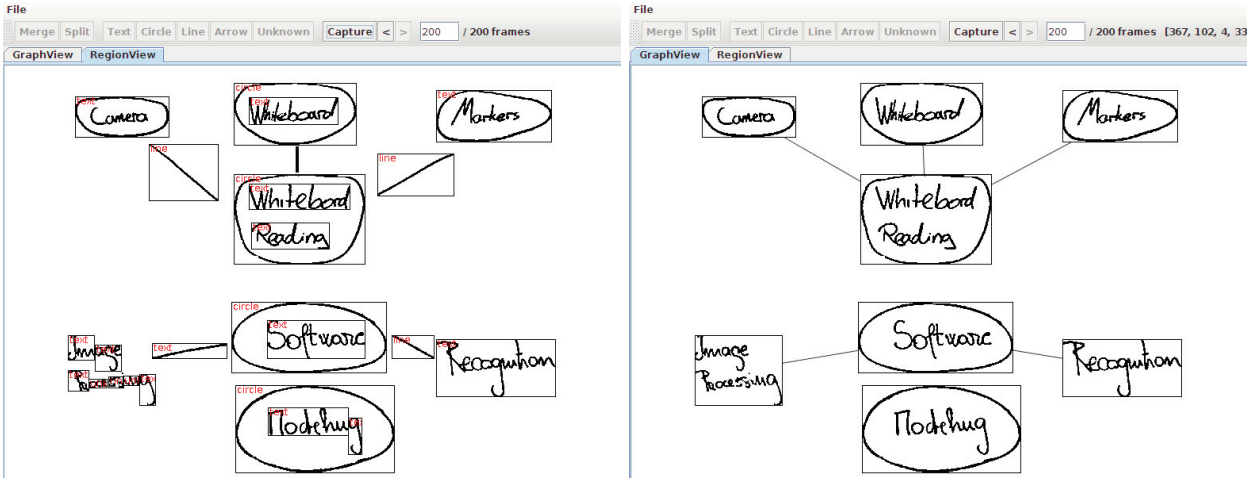


Figure 3: User interface of the *Mindmap Manager* showing the region view on the left and the graph view on the right. In the region view segmentation and recognition results can be corrected. The graph view shows the final graph representation of the mind map.

D. Word detection

Once the modeling part is done, the different gravity centers will form "dense" regions (see Fig.2) corresponding to possible words. These agglomerations into different clusters need to be identified in order to separate the different words from each other. For this purpose the DBSCAN algorithm [14] has been considered. While other clustering methods rely mainly on some distance metric in this case the distance is combined with the density.

The gravity centers will be clustered not only by the distances (between the different text patches), but also by the density which is definitely higher around the different text components (see Fig. 2).

Let $D_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ be the coordinates of the gravity centers, where $x_i, y_i \in R^+$ and n denotes the number of gravity centers in the set.

Let us denote by β-neighborhood of a point $p_k \in D_n$ the $N_\beta(p_k) = \{(p_k) \in D_n | dist(p_k, p_l) < \beta, k \neq l\}$, where $dist()$ is the Euclidean distance.

Considering the β-neighborhood of a point, we define the notion of p_k is density reachable from p_l if $p_k \in N_\beta(p_l)$ and $|N_\beta(p_l)| \geq P_{min}$, where P_{min} is the minimal number of points (gravity centers) which should be around point p_k.

The proposed clustering process is based on the β-neighborhood of a given point. We select as belonging to one cluster all the points which are density reachable considering a given number of k (number of neighbors). The expansion of each cluster is based on this idea allowing to get rid of the noisy points which density reachability indices are lower than for the others. For more details about the clusters expansion, please refer to work [14]. Finally, the original CCs' gravity centers are mapped to the different clusters established by DBSCAN (see Fig. 2).

E. Word recognition

For the recognition, we use the same recognizer that in our previous work was successfully applied to the task of reading text paragraphs in high-resolution whiteboard images [4]. Handwritten words are modeled by semi-continuous character HMMs. A sliding window is applied on the normalized text snippets considering a 8 pixel overlapping stripe for the sequence. For each frame a set of nine geometric features and the approximation of their first derivatives are computed [15]. In total, 75 models considering upper and lower case letters, numerals and punctuation marks have been trained on the IAM database [4]. The HMM models are decoded with a time-synchronous Viterbi beam search algorithm [16].

F. The Mindmap Manager

The *Mindmap Manager* is the front end of the presented system. Segmentation, grouping and recognition results are consecutively — for each change in the region memory — saved and made accessible to the user for post-editing and exporting. Fig. 3 shows its user interface. The region view (left side of Fig. 3) contains editing functionality for incorrectly segmented or classified components. Please note that those components refer to classified CCs that in case of text elements might have undergone further grouping. After selecting a component its category can be changed or the entire component can be split horizontally or vertically. If the categories of two different components are compatible, they can be merged.

The graph view (right side of Fig. 3) contains a graph representation of the mind map. Text and circle elements are treated as nodes and lines and arrows are treated as edges (compare region and graph view in Fig. 3). This way the user is able to rearrange the layout of the nodes by simply dragging and dropping them. Connected edges will follow. Besides the possibility to rearrange the mind map layout the graph view

has no further editing capabilities and is rather intended to contain the final outcome of the system. A compatibility with existing digital mind map formats would allow to use other, already existing tools to further manipulate the documents.

The graph representation has to be created from the previously mentioned components. Because there is no prior knowledge of which nodes are connected by a line, there is needed some estimation. By transferring components classified as lines to Hough space [17], a parametric representation of the line can be estimated. Finally, two nodes being connected by the line component are determined trough intersection of the parametric line with neighboring components.

G. Interaction with the whiteboard

The purpose of user interaction is to give a feedback of segmentation and recognition results while the mind map creation is still in progress. This way the user can react accordingly (e.g. writing things clearer) to improve the overall performance. The feedback is given by highlighting newly recognized elements on the whiteboard using the projector. The highlighting color indicates the classification result (text, line, circle, arrow). To use the camera and the projector in conjunction a camera-projector calibration has to be performed initially (also see Section III-A).

Segmentation and recognition results can be retrieved whenever the region memory changes (see Section III-B1). Ideally those updates to the region memory occur only if there is a change to the written content on the whiteboard. In such cases changed CCs are determined by computing a difference (XOR) image in the region memory. From their bounding boxes highlighting events are now generated that additionally contain the category (text, line, circle, arrow) for highlighting in a specific color. The bounding boxes are given in camera image coordinates and have to be mapped to projection image coordinates for rendering. This mapping can be computed through the formerly estimated homography (see Section III-A). After rendering, the user will see a colored rectangle around all recently changed parts of the mind map for a few seconds. In order to be more robust to false updates of the region memory (e.g. due to illumination changes), highlighting events will only be generated if the change in the region memory exceeds a certain threshold.

Finally we have to deal with the effect that projections to the whiteboard are also captured by the camera. This way projections can result in CCs being extracted that do not correspond to written content on the whiteboard. The idea is to filter the area in the camera image, where there will be a projection, from the region memory.

IV. Experiments

In this section a brief description of the data and the results achieved by the described method will be presented.

A. Data description

The dataset consist of 31 mind maps written by 11 different writers around the topics “study”, “party” and “holiday”. 2 writers sketched only 2 mind maps. The data is annotated with respect to the nature of connected components (line, circle, text, arrow) and words [3].

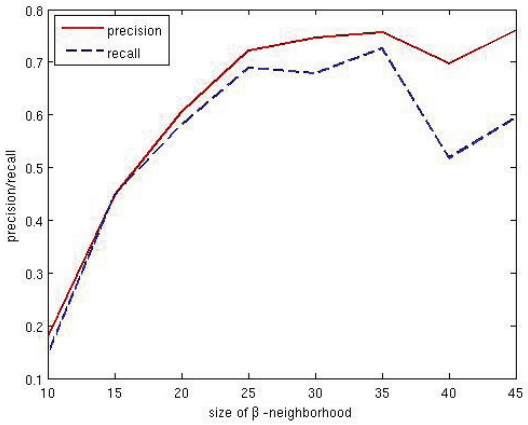


Figure 4: Layout analysis results (Precision and recall) for the document shown in Fig. 6.

B. Results

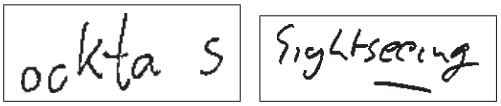
To evaluate thoroughly the method we need to evaluate the text detection solution, the subsequent modeling strategy and the text recognition. As the text detection method was proposed in [3] we just use those results and we focus our evaluation on the layout analysis and the recognition. For more details on the results concerning the text detection, please refer to [3].

Text Detection: The neural network provides an average recognition score of 95.7% for the different CCs as being text, line, circle or arrow. However, while for text components the recognition scores are high (99.4%), for lines and arrows there are elevated confusion rates.

Layout Analysis: For the evaluation of the proposed method we use the method introduced in the context of the ICDAR 2005 Text Locating Competition [18]. The evaluation scheme is based on *precision* and *recall*. The *recall* then is the quotient of the sum of the best matches of the ground truth among the agglomerated areas and the number of all annotated bounding boxes within the ground truth.

We evaluated the output of the agglomeration (modeling) using both schemes described above. In Fig. 4 we display a typical result of the hierarchical clustering, stating in this case the maxima for precision and recall at 75% and 72%, respectively. The average recall value for the test documents is 58.09%.

While in some cases, the agglomeration is successful, in some other cases it fails because of some CC recognized as non-text (e.g. M in “Motto” or D in “Dance” in Fig. 6) or due to some distances which lead to agglomeration or separation (see “Guests” in Fig. 6) of different text items. Overall, in 211 cases the agglomeration produced non-text word hypothesis, in 194 cases some parts of the word (mainly characters)



(a) missing letters (b) agglomeration with noise

Figure 5: Typical error cases occurred in the agglomeration

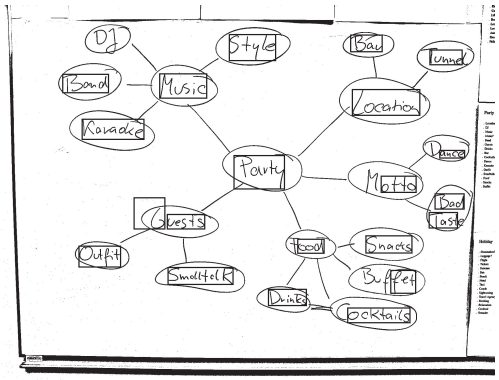


Figure 6: Layout analysis results on an exemplary mind map.

were missing. Finally, in 353 cases complete words, or words preceded or followed by noises (mainly lines) were detected. Some typical agglomeration errors are depicted in Fig. 5.

Word Recognition: The recognition of the word snippets by the HMM are reported only on those 353 words considered as successful (complete) for the grouping. The lexicon (164 entries) was generated from the transcripts including all tentatively written words irrespective of segmentation errors. To reject test fragments with erroneous ink elements a rejection module was used, defined as an arbitrary sequence of character models. The overall word recognition score in the different snippets is 40.5%. 83.3% of the snippets were recognized correctly (i.e one word snippets). The low scores can be explained by the fact that the recognizer is trained on completely different data, while the recognition is performed on low-resolution image snippets, with huge writing style variations and containing also additional noise components inherited from the grouping process.

V. Summary

In this paper we proposed a basic prototype reading system to automatically recognize mind maps written in an unconstrained manner on a whiteboard. Instead of considering expensive equipment, only common tools like e.g. whiteboard, markers, camera, and projector were considered in the recognition scenario, which usually are available in whatever conference room.

Instead of establishing some rules, the method adapts to the layout of each analyzed document. The modeling of the text components by their gravity centers followed by Density Based Spatial Clustering will provide the solution to merge the detected text patches (connected components) into words which serve as input for a handwriting recognizer. For this preliminary work, the recognition results, even though the recognizer was trained on completely different data, are not satisfying yet, but with some post-processing of the grouping more complete word agglomerations can be submitted for recognition. The software tool and the interactivity with the whiteboard provides a straightforward solution for a human-computer interaction in this challenging automatic whiteboard reading scenario.

Acknowledgment

This work has been supported by the German Research Foundation (DFG) within project **Fi799/3**.

References

[1] S. Vajda, K. Roy, U. Pal, B. B. Chaudhuri, and A. Belaid, "Automation of Indian postal documents written in Bangla and English,," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 8, pp. 1599–1632, Dec. 2009.

[2] M. Weber, M. Eichenberger-Liwicki, and A. Dengel, "a.scatch - a sketch-based retrieval for architectural floor plans," in *International Conference on Frontiers of Handwriting Recognition*, 11 2010, pp. 289–294.

[3] S. Vajda, T. Plötz, and G. A. Fink, "Layout analysis for camera-based whiteboard notes," *Journal of Universal Computer Science*, vol. 15, no. 18, pp. 3307–3324, 2009.

[4] T. Plötz, C. Thurau, and G. A. Fink, "Camera-based whiteboard reading: New approaches to a challenging task," in *International Conference on Frontiers in Handwriting Recognition*, 2008, pp. 385–390.

[5] D. Yoshida, S. Tsuruoka, H. Kawanaka, and T. Shinogi, "Keywords recognition of handwritten character string on whiteboard using word dictionary for e-learning," in *Proceedings of the 2006 International Conference on Hybrid Information Technology - Volume 01*, ser. ICHIT '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 140–145.

[6] P. Farrand, F. Hussain, and E. Henessy, "The efficiency of the "mind map" study technique," *Journal of Medical Education*, vol. 36, no. 5, pp. 426–431, 2003.

[7] T. Plötz and G. A. Fink, "Markov models for offline handwriting recognition: a survey," *IJDAR*, vol. 12, no. 4, pp. 269–298, 2009.

[8] M. Wienecke, G. A. Fink, and G. Sagerer, "Towards automatic video-based whiteboard reading," in *International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 2003, pp. 87–.

[9] M. Liwicki and H. Bunke, "Handwriting recognition of whiteboard notes," in *Conference of the International Graphonomics Society*, 2005, pp. 118–122.

[10] ——, "Handwriting recognition of whiteboard notes – studying the influence of training set size and type," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 1, pp. 83–98, 2007.

[11] D. Yoshida, S. Tsuruoka, H. Kawanaka, and T. Shinogi, "Keywords recognition of handwritten character string on whiteboard using word dictionary for e-learning," in *International Conference on Hybrid Information Technology*, 2006, pp. 140–145.

[12] W. Niblack, *An introduction to digital image processing*. Birkeroed, Denmark: Strandberg Publishing Company, 1985.

[13] L. Fletcher and R. Kasturi, "A robust algorithm for text string separation from mixed text/graphics images," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 10, no. 6, pp. 910–918, November 1988.

[14] H.-P. Ester, M.and Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[15] M. Wienecke, G. A. Fink, and G. Sagerer, "Toward automatic video-based whiteboard reading," *IJDAR*, vol. 7, no. 2-3, pp. 188–200, 2005.

[16] G. A. Fink, *Markov Models for Pattern Recognition, From Theory to Applications*. Heidelberg: Springer, 2008.

[17] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 2001.

[18] S. M. Lucas, "Text locating competition results," in *International Conference on Document Analysis and Recognition*, 2005, pp. 80–85.