Journal of Universal Computer Science, vol. 15, no. 18 (2009), 3307-3324
submitted: 22/9/09, accepted: 23/12/09, appeared: 28/12/09

Layout Analysis for Camera-Based Whiteboard Notes

Szilárd Vajda
(Department of Computer Science
TU Dortmund
44221, Dortmund, Germany
szilard.vajda@udo.edu)

Thomas Plötz
(School of Computing Science
Newcastle University
NE7 1NP, Newcastle upon Tyne, United Kingdom
t.ploetz@ncl.ac.uk)

Gernot A. Fink
(Department of Computer Science
TU Dortmund
44221, Dortmund, Germany
gernot.fink@udo.edu)

Abstract: A domain where, even in the era of electronic document processing, handwriting is still widely used is note-taking on a whiteboard. Such documents are either captured by a pen-tracking device or – which is much more challenging – by a camera. In both cases the layout analysis of *realistic* whiteboard notes is an open research problem.

In this paper we propose a camera-based three-stage approach for the automatic layout analysis of whiteboard documents. Assuming a reasonable foreground-background separation of the handwriting it starts with a locally adaptive binarization followed by connected component extraction. The latter are then automatically classified as representing either simple graphical elements of a mindmap or elementary text patches. In the final stage the text patches are subject to a clustering procedure in order to generate hypotheses for those image regions where textual annotations of the mindmap can be found.

In order to demonstrate the effectiveness of the proposed approach we report results of a writer independent experimental evaluation on a data set of mindmap images created by several different writers without any constraints on writing or drawing style.
Key Words: whiteboard notes, handwriting recognition, mindmap recognition, camera based recognition, writer independent document layout analysis
Category: H.3.1, H.3.3, H.4.1

1 Introduction

In many areas writing down notes or texts manually using, for example, pens has been replaced by machine-based techniques. Very prominently, it is nowadays standard to write an email using a computer and a keyboard rather than actually writing a letter. Without any doubts, electronically supported creation of documents implies several advantages. Machine-printed texts are easy to read by virtually everybody. Furthermore, storage and retrieval are more convenient for electronic rather than for handwritten documents.

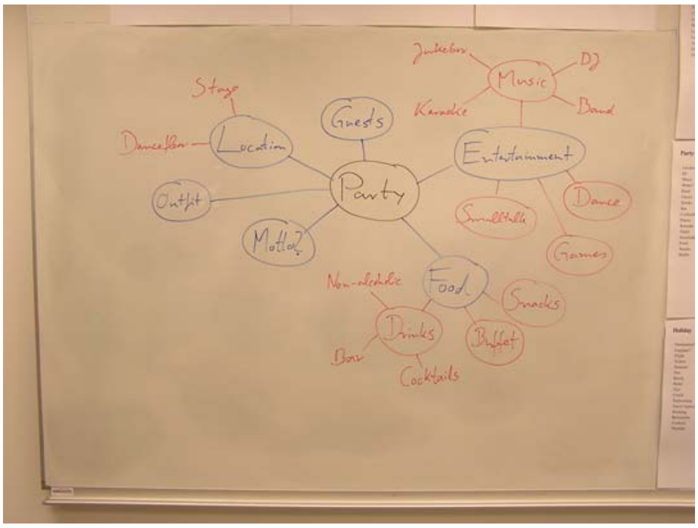
However, there are still certain application cases where the "traditional" way of handwriting is more favorable [Sellen and Harper, 2002]. Especially for creative processes like brainstorming any (electronic) equipment that might distract the attention of humans is likely to hinder the process of generating ideas. Basically, distraction kills creativity. Consequently, in such cases people often fall back on "low-tech" equipment for writing down their ideas, namely to pens and paper.

A standard means of writing down the results of a brainstorming session in a well structured way is *mindmapping* [Buzan, 2003]. A mindmap basically corresponds to a graph with nodes and edges. Nodes represent the ideas that are usually written down as short texts – mostly a single or just a small number of words each. Relations between certain ideas are visualized by (directed) edges between these nodes. Apart from that, there is no constraint in how to organize a mindmap, for example, w.r.t. writing style, writing direction etc. For group based brainstorming mindmaps are usually created on a whiteboard, which is nowadays standard equipment of a meeting room. In Fig.1 a mindmap created on a whiteboard and its digital counterpart is shown.

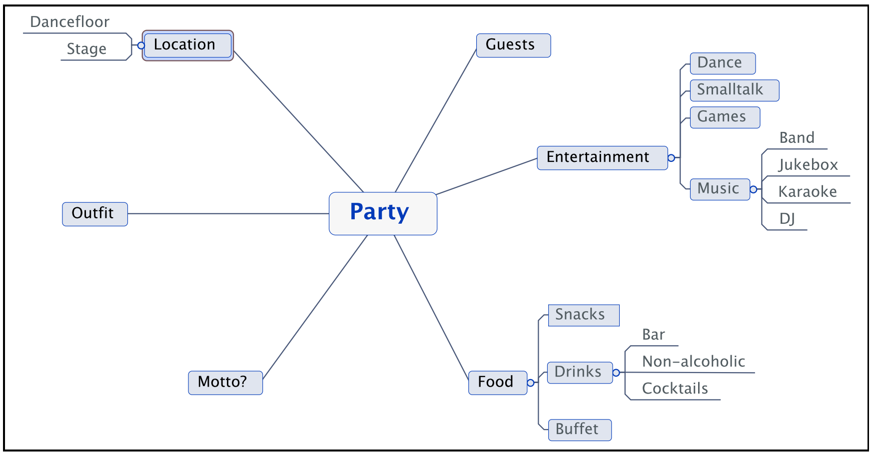
When restricting note taking to the use of pens and whiteboard, unfortunately, all advantages of electronically supported techniques vanish (Fig.1a). However, particularly for storage and retrieval, digital representations of whiteboard notes in general (Fig.1b) and especially mindmaps written on it are desirable. For their creation the paradigm of non-obtrusiveness remains, though.

In our work we develop a camera-based automatic whiteboard reading system [Plötz et al., 2008]. One goal is to monitor the dynamic process of creating mindmaps on a whiteboard using a video camera and to automatically extract a digital representation of the mindmap. The latter then can be used for the desired electronic storage and retrieval. By means of a projector recognized mindmaps can easily be reproduced directly at the whiteboard. This allows intuitive interaction (editing, erasing, browsing etc.) with the mindmap using natural means, i.e. pens, eraser and whiteboard.

One prerequisite for the successful recognition of a mindmap is its segmentation w.r.t. graphical elements (circles, lines, arrows) and text blocks. In this paper we present a writer independent approach for the automatic layout analysis of the structure of handwritten mindmap drawings. Still images of mindmaps written on whiteboards serve as input data. In a three-stage procedure we first extract relevant connected components, which are then fed into a classification system. At this second stage of the proposed procedure features calculated from



(a)



(b)

Figure 1: The same mindmap example: (a) created on a whiteboard (b) created by a computer software

the extracted connected components are automatically classified as either belonging to some graphical element or as being part of handwritten text. For a successful mindmap recognition we then agglomerate connected components of the same type to larger portions of structurally connected basic elements. Clusters of connected text components form single words that are the input for our handwriting recognition system. The output of the presented approach is a full segmentation of a mindmap image that includes region-based annotation at the level of graphical elements (circles, lines, arrows) and words. By means of a writer-independent experimental evaluation on a database of mindmaps we demonstrate the effectiveness of this new approach.

2 Related Work

A digital document may consist of a large variety of physical items such as text blocks, lines, words, figures, tables, etc. However, at a lower level, all these items are just connected components, which are a set of interconnected pixels containing no high-level description at all. The goal of document structure and layout analysis is to detect the different regions in the document and to identify the functional roles and relationships between them [Namboodiri and Jain, 2007].

While a human reader uses several clues like context and a-priori information about the script together with a complex reasoning mechanism, the machine can rely only on the extracted low-level information. This is the reason why automatic layout and structure analysis of an arbitrary document is a very challenging task. However, we should distinguish between printed documents and handwritten ones. While for printed documents we can presume a certain layout structure [Gatos et al., 2000] or regularity at letter level (font type, font size, boldness [Klink et al., 2000]), for handwritten documents there is usually a total lack of physical organization.

2.1 Whiteboard documents

While some impressive results have been achieved for the recognition of handwritten forms, postal documents [Fujisawa, 2008, Vajda et al., 2009], and mathematical formulas (cf., e.g., [Garain and Chaudhuri, 2003, Nomura et al., 2003, Tapia and Rojas, 2005]), the analysis and recognition of whiteboard notes is a relatively new issue in the scientific community and just few attempts can be found in this research field. The challenge to recognize such documents arises from the fact that their structure and content is completely unconstrained.

In [Liwicki and Bunke, 2005] the authors propose a system to recognize whiteboard notes by using an HMM based recognizer. In this system the image acquisition is performed on-line utilizing an infrared sensor mounted on the corner of the whiteboard and a normal pen covered by a special casing which sends the infrared signal. However, the work addresses just the problem of word recognition of well structured handwritten notes, preceded by a simple pre-processing to eliminate internal noise, without considering any extra information, which can occur in such a document.

A kind of e-Learning strategy using a whiteboard has been described in [Yoshida et al., 2006]. The authors use two cameras and a pen capture tool on the whiteboard to recognize Japanese characters based on some character matching. However, to detect the text regions from the whiteboard, they consider the software provided by the pen manufacturer.

In [Oliveira and Lins, 2007] the task of processing whiteboard images is addressed using portable digital cameras or cell-phones. However, the work is more related to image processing rather than to its analysis. The focus is on the detection of board boundaries and on image quality enhancement. The output of the described procedure can then be used for further analysis.

In our previous work [Plötz et al., 2008] we considered a similarly challenging task as our goal was to recognize whiteboard notes taken with a camera and without any available on-line information. The text detection strategy was based on the different pieces of low-level information extracted from the connected components. The drawback of this strategy is its rigidity as it considers global thresholds to distinguish between textual and non-textual items.

2.2 Other documents

Back in the 1980s research around textual documents has been extended to line drawings. The original raw data was scanned documents but the aim was not to recognize the structure/layout and content but to rebuild the high-level design from engineering drawings, recognize pipes, lines, roads, rivers in maps, etc.[Tombre and Lamiroy, 2008]. Considering the content of these documents, maybe they are more complex than printed materials but still operating with a limited and well defined set of graphical items.

Nowadays, the focus has been oriented toward text/graphic separation which is not obvious in some engineering drawing [Tombre et al., 2002] where text portions overlap other objects like lines, pictures, etc. Similar challenges can be encountered in postal documents [Vajda et al., 2009] where the address block should be separated from stamps, business cards [Mollag et al., 2009], or official documents [Roy et al., 2009] used in the administration. In [Vajda et al., 2009] the authors perform a run length smoothing in order to distinguish between text items and stamps but they presume a certain amount of text present in the document.

All these strategies to separate text from non-text have a common root relying on the method proposed by Fletcher and Kasturi [Fletcher and Kasturi, 1988]. They calculate on connected components (CC) the height, the width, the aspect ratio, pixel density, number of horizontal, vertical segments [Tombre et al., 2002], etc. Afterwards, based on some adaptive thresholding they design rule-based classifiers [Mollag et al., 2009] to separate the text layer from the rest of the document.

Such a rule-based strategy – even an adaptive one – can only work for printed documents where a considerable amount of printed text is available but fails for unconstrained documents like mindmaps where the size and the shape of characters is different and the calculated thresholds are not accurate.

3 Camera-Based Segmentation of Mindmaps

A mandatory pre-processing step for successful recognition of hand-drawn mindmaps is their segmentation. The goal of this process is to annotate regions of a camera image w.r.t. graphical elements and text. We developed a three-stage procedure that handles still images of mindmaps and produces a complete region-based annotation. The overall procedure is illustrated in Fig. 2. It starts with the extraction of relevant connected components (cf. Fig. 2c). In the second stage, the latter are automatically classified using a statistical modeling approach. Therefore, feature representations of all connected components are fed into a classifier that provides a labeling w.r.t. circles, lines, arrows and text (cf. Fig. 2d). In the last stage connected components of the same type (text snippets) are agglomerated by means of a hierarchical clustering procedure (cf. Fig. 2e). The output of our system is the agglomerated word snippets which provide a word-level description of the whole document which can be recognized further by a handwriting recognition system.

3.1 Connected Component Extraction

Connected component labeling is a well-established means for separating graphical elements in images. The structure of mindmap images suggests this procedure as there is apparently a rather clear distinction between handwriting in the foreground and a more or less homogeneous background (the surface of the whiteboard). Thus, connected components are very likely to be concentrated on the actual mindmap. Disregarding probable flaws in the image (e.g. inhomogeneous lighting, or non-opaque marker color) separating the mindmap by connected component analysis is reasonable.

For the purpose of connected component extraction the input image has to be binarized, which is accomplished by Niblack's algorithm [Niblack, 1986]. A variant of the basic approach is used that applies threshold optimization [Sauvola et al., 1997] and thresholding locally in a 51x51 pixels window. For an efficient computation integral images for plain proposed in [Shafait et al., 2008] by Shafait as well as for squared pixel intensities are analyzed. The actual extraction of connected components follows a straightforward approach of segmenting contiguous black pixel regions.

By means of heuristic post-processing connected components that obviously do not belong to the mindmap are suppressed by trivial filtering. Components with small size and extremly huge dimensions are discarded from further processing. The remaining set of connected components is not necessarily limited

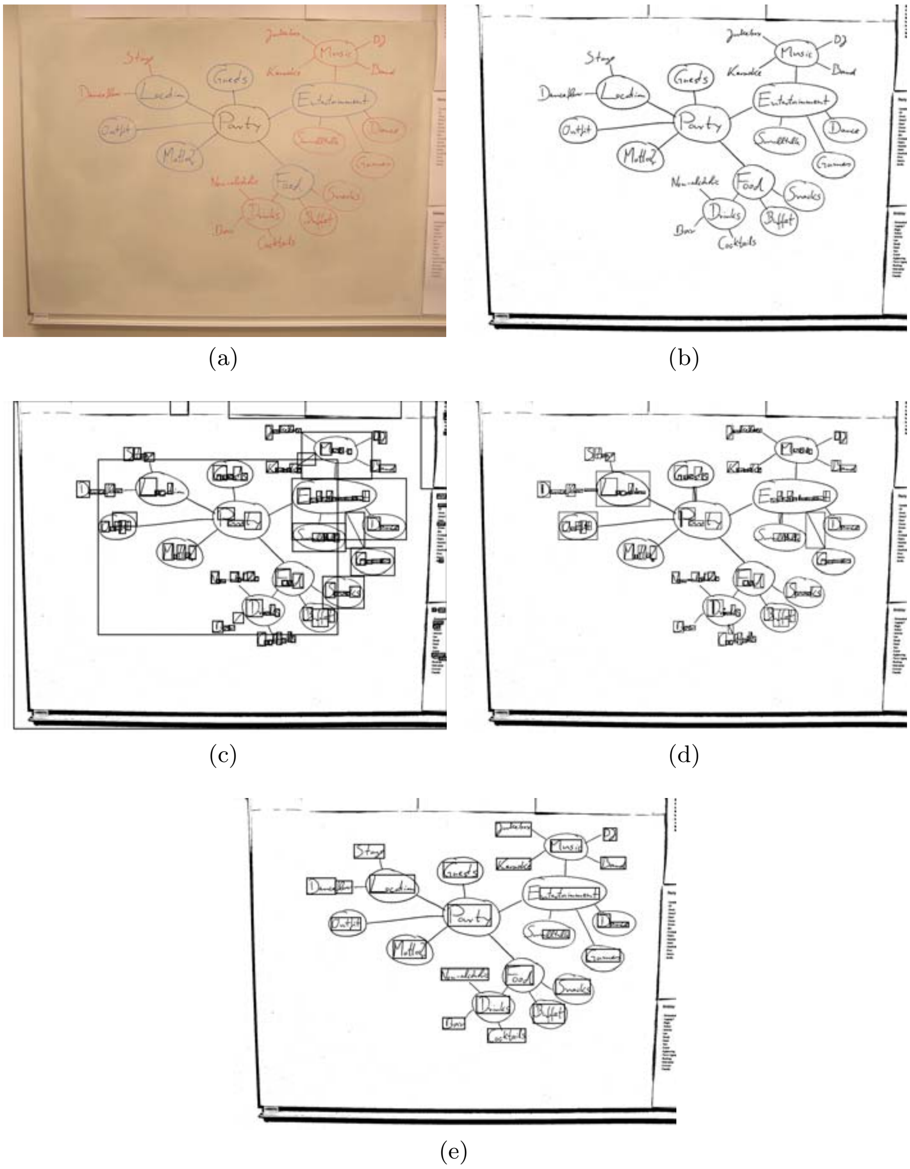


Figure 2: System overview: (a) original image (b) binary image (c) connected components detection (d) textual items detection (e) text items agglomeration

to well isolated, known graphical elements and text portions only. Instead unknown and touching elements together with additional clutter are very likely to occur (cf. Fig. 2c). Thus, for successful segmentation of mindmap images further analysis of the extracted connected components is required in future studies.

3.2 Classification of Connected Components

In the second stage of our segmentation approach the set of extracted connected components is classified w.r.t being either one of the known graphical elements, text, or unknown. In the latter case the particular connected component is discarded from further processing.

In a mindmap circles/ellipses, lines and arrows are used for the purpose of grouping, linking, and structuring. This fact is distinguishing the graphical elements from the textual ones, both on logical and physical level, respectively. Basically, all considered elements exhibit certain structural specialties. Textual components, for example, differ from others by their texture, (black) pixel density, size, etc. [van Beusekom et al., 2007]. Similar conclusions can be drawn for lines, circles and arrows. Consequently, reasonably discriminating features can be extracted from image data that will serve as input for a classification system.

Our main goal is to avoid the heuristic, threshold estimation based solutions proposed by different authors ([Fletcher and Kasturi, 1988, Mollag et al., 2009, Tombre et al., 2002, van Beusekom et al., 2007]). In contrary to setting up some rules to distinguish between the different types (text/non-text), we propose to train a neural network on this purpose. Such a learning mechanism in the decision can adapt more precisely to the data and its characteristic features than the different thresholds, based on some trial runs or heuristics.

We investigated two kinds of feature sets. On the one hand standard statistical features are calculated on image data. These measures are invariant in size and rotation. Roughly speaking they represent – to some extent – shape related properties of the analyzed connected components. Alternatively, intensities of gradient histograms (values ranging from 0 to 255, equally divided into 16 bins) of the connected components serve as features (gradient set).

The shape-set is based on the features proposed by Becker, the winner of the ICDAR 2005 Text Locating Competition [Lucas, 2005]. They have been used successfully for natural scene text detection. In order to also cope with the detection and discrimination of graphical elements we extended the original set by certain additional statistical measures. In the remainder of this paper the first set of features is referred to as shape feature set. The features description details can be found in Appendix A.

Using either the shape set or the gradient set two alternative feature representations for connected components are extracted. In the first case input data is represented by a 12 dimensional feature vector whereas in the latter the resulting feature space contains 16 elements.

The actual classification of the particular feature vectors is based on a Multi-Layer Perceptron (MLP)[Vajda et al., 2009]. By means of cross-validation the network topology has been adjusted. We use one hidden layer with 15 or 20 neurons and the sigmoid function as activation function. Model training is based on standard backpropagation. The input and output of the network is defined by the number of input features calculated for each component (12 or 16 based on their nature) and the number of classes to be identified (4, i.e. arrows, circles/ellipses, lines, and text).

In order to deal with input patterns that do not belong to one of the known classes the following rejection strategy is applied. Let us denote by s_1 and s_2 the two best outputs of the classifier, i.e. the scores computed for the best and second-best solution. A pattern is rejected if the difference $|s_1 - s_2|$ between the scores of the top two class hypotheses falls below a certain threshold ϵ. Therefore, the rejection rate can be controlled by adjusting the parameter ϵ.

3.3 Text Agglomeration

Once the classification of the different connected components is performed by the MLP, we can proceed to a higher level in the mindmap analysis. At this stage we step from a lower, connected component based level, to a more complex structural one, which projects a sort of vague layout analysis as we can already distinguish between text and non-text elements (lines, circles, arrows). However, the primary goal is not to detect the layout but to merge different identified textual connected components into so-called "word structures". This pseudo word level cannot really be equated with the physical word level as there is no information about what might be a word. This merging strategy is necessary for the further processing when a subsequently applied word recognition tool has to recognize the text.

Knowing that characters usually appear closer to each other than to other elements, by clustering they should group with their kind rather than with non-text elements. For that reason we discard all the items tagged by the classifier as being non-text and perform a hierarchical clustering trying to merge the remaining items into words. An example of the agglomeration is shown in Fig. 3.

We have considered different distances in order to measure the similarity between two clusters. As we selected an agglomerative clustering strategy, we explored the suitability of the Euclidean distance between the physical center of the two connected components. A similar measure is computed for the gravity centers. Furthermore, the minimal distance between the boxes bounding connected components is also considered. While these measures are easy to calculate their complexity is still high. Based on preliminary results we select the minimal distance of the bounding boxes to be considered in the further investigations.

A faster strategy is proposed by Yuan et al. [Yuan and Tan, 2005], where the distance is based on the size of components to be merged as well as on the Euclidean distance between the components. In [Huang and Tan, 2007] the same idea was used successfully in a greedy clustering approach to separate text, drawings, charts, etc. Therefore, as an alternative to the hierarchical clustering approach we explore the capabilities of greedy clustering using the following distance function proposed in [Huang and Tan, 2007]:

$$f(s_1, s_2) = \sqrt{\frac{k s_1 s_2}{s_1 + s_2}} \tag{1}$$

where s_1, s_2 represent the sizes of the two connected components in terms of number of black pixels and k is a parameter controlling the level of the grouping. In our case the value ok k is set to 10 based on some trial runs. Analyzing the function f it becomes clear that it is rotation invariant, symmetric, and it does not respect just the distances but also the sizes of the components.

In order to use this measure, we calculate the distance between the components c_1 and c_2 having the size s_1 and s_2. If this distance is smaller than the value given by the function f then component c_1 and c_2 are merged and form a new cluster. This operation is iterated while all the unique components are tested.

4 Evaluation

In order to evaluate the effectiveness of the proposed system we performed writer-independent experiments on real whiteboard images. In the following we first give a description of the data set. Then classification results for connected components analysis are presented that illustrate the capabilities of the system to discriminate between text and non-text elements (circles, lines, arrows). Finally, results achieved by clustering the connected components are discussed.

4.1 Data-set

Our whiteboard images based document set consists of 30 photos we took from mindmap-drawings on a whiteboard (e.g. Fig 4). Eleven different writers were asked to freely draw one mindmap for each of the topics "holiday", "party" and "study" (three writers sketched only two mindmaps). The writers were provided with a standard whiteboard marker set containing four different colors (black, blue, green, red) and a whiteboard eraser. Except for a basic set of words for each topic, which had to be used and an obligation to add at least three other words to the mindmap, there were no restrictions in creativity. In order to produce

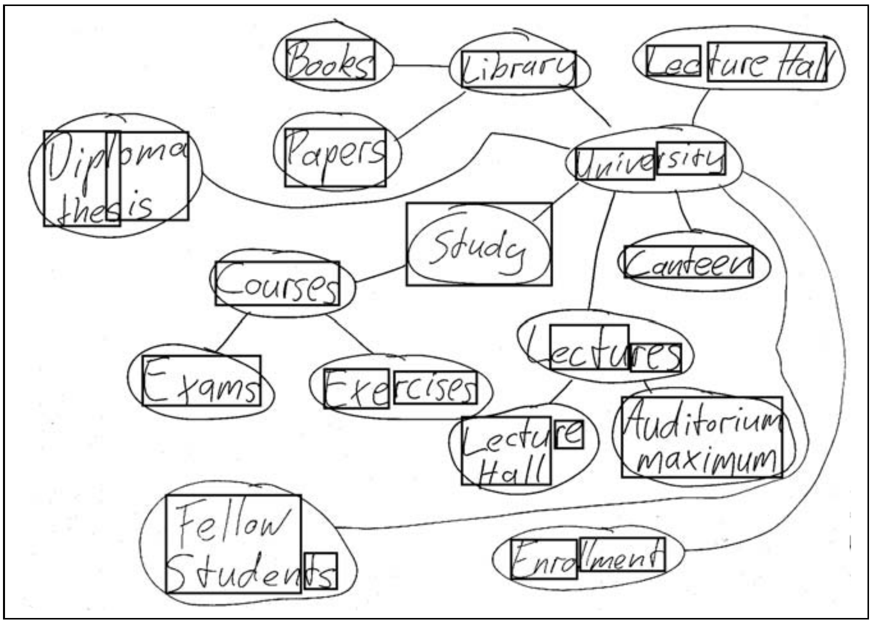


Figure 3: Results of text agglomeration (rectangles) on an exemplary mindmap image.

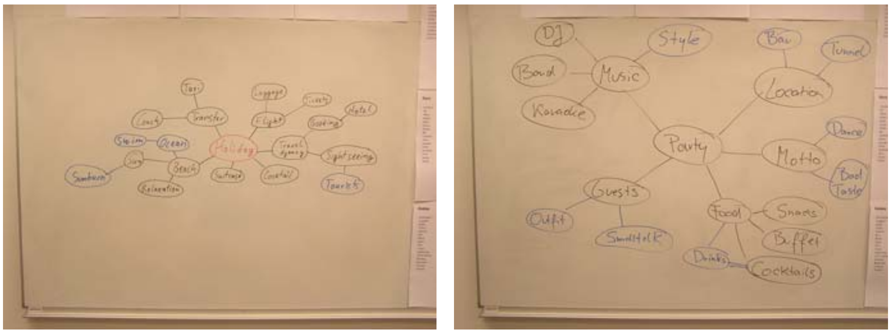
the ground-truth, a software has been developed where the user can manually label the different automatically detected connected components and label the connected components at word level.

After a writer had finished his mindmap, a photo of the whiteboard was taken with a digital camera set to a resolution of 2,048×1,536 pixels. Reasoned by the image acquisition process, we can encounter in the picture items like the wall, other printed documents linked to the whiteboard, and frame parts of the board, which are not part of the document. These items are considered being noise elements (cf. e.g. [Oliveira and Lins, 2007] for a comparable argumentation).

For the evaluation we split our data into a training and a test set. The training set consists of 19 mindmap images generated by 7 writers (no. 1, 3, 5, 7, 8, 9, 11). For testing we used 11 different images produced by the remaining 4 writers (no. 2, 4, 6, 10).

4.2 Results

In the first experiment we evaluated the classification capabilities of the second stage of our segmentation approach, namely the analysis of connected components w.r.t. the discrimination between texts, circles, lines and arrows. The connected component labeling procedure produced 1,488 (5,142) texts, 159 (646)



(a) (b)

Figure 4: Mindmap examples from the dataset describing: (a) a holiday concept (b) a party concept

Features	Accuracy[%]
shape set(12)	93.7
gradient set(16)	93.0

Table 1: Results of connected components classification using different feature sets.

lines, 146 (508) circles and 27 (62) arrows for the test and training set, respectively. We evaluated the two different feature sets described in Sec. 3.2. The overall classification accuracies (without rejection) for the shape set and the gradient set are given in Tab. 1. It can be seen that both types of features are suitable for the classification. However, using the shape features set produces slightly better results. For the sake of clarity we limit the further presentation of results to this feature representation. The particular classification results of each writer from the test set can be seen in Tab. 2.

While the text items are recognized with a high precision (98.5%) the arrows are often confused with lines. This confusion can be explained by the fact that just a few arrows are represented in our data set and there is not much differ-

	Writer			
	2	4	6	10
Accuracy [%]	92.1	96.2	91.5	94.2

Table 2: Writer specific classification results for connected components analysis using the shape feature set.

ϵ	Accuracy[%]	Misclassified[%]	Rejection[%] (% FP)
0.1	94.5	5.5	1.5 (0.7)
0.3	95.4	4.6	4.1 (2.3)
0.5	96.6	3.4	7.7 (4.6)
0.7	97.6	2.4	14.7 (10.5)
0.9	98.6	1.4	32.0 (26.8)

Table 3: Dependency of classification accuracy on the choice of the rejection threshold using the shape feature set

ence between them. Similar problems can be encountered for circles, which can erroneously be confused with text items as, e.g., "o", "D".

The rejection rates and the corresponding recognition accuracies for the different ϵ values (cf. Sec. 3.2) are given in Tab. 3. Setting a stronger rejection criterion by increasing the parameter ϵ provides a higher accuracy, a decreasing misclassification rate, and an increasing rejection percentage. However, the number of false positive rejections (given as absolute percentages in parentheses) also increases. Thus, care needs to be taken when adjusting the rejection threshold. For all further experiments reported in this paper we used small ϵ values, which practically implies no rejection.

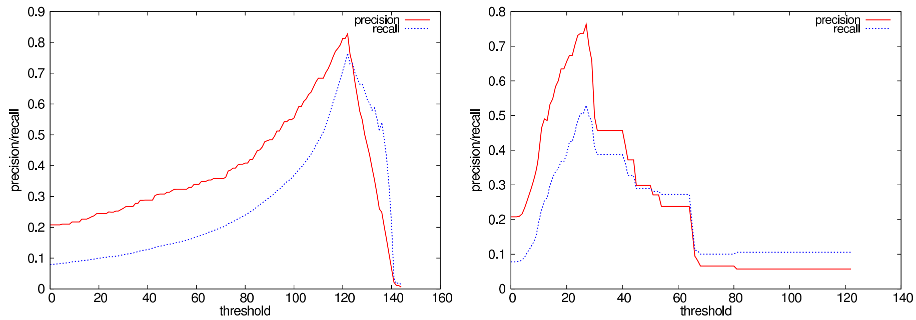
The bounding boxes of the annotated ground truth T and the agglomerated text components E are compared. The larger the overlap of the bounding boxes, the higher the level of match. A match m_p between two rectangles r, r' is defined as the quotient of their intersection area and their union area:

$$m_p = \frac{A(\bigcap(r, r'))}{A(\bigcup(r, r'))}.$$

The evaluation scheme is based on *precision* and *recall* known from the domain of Information Retrieval. However, a binary answer to whether there is a fitting ground-truth rectangle to an estimated one or not could not cope with partial matches. Therefore, the quality for a match m_p in this case lies in the range of $[0; 1]$. In order to calculate adapted versions of precision and recall the best match between a rectangle within the agglomerations and all rectangles within the set of annotations is taken into consideration – and vice versa. The best match $m(r, R)$ of a rectangle r within a set of rectangles R is defined as:

$$m(r, R) = \max \{m_p(r, r') | r' \in R\}.$$

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:



(a) Hierarchical agglomeration (b) Greedy clustering

Figure 5: Comparison of precision/recall values for one example document

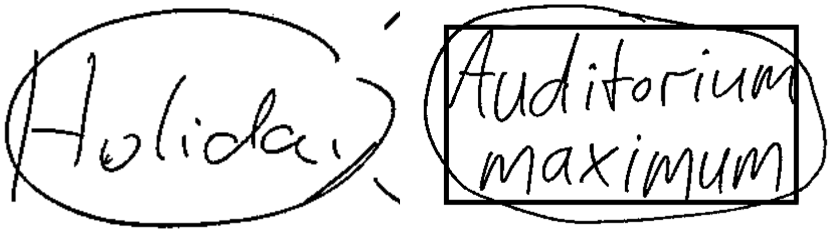
$$\text{recall} = \frac{1}{|T|} \sum_{r_t \in T} (r_t, E).$$

The *precision* relates to the quotient of the sum of the best matches of the agglomerated areas among the annotated regions and the number of all agglomerated areas:

$$\text{precision} = \frac{1}{|E|} \sum_{r_e \in E} m(r_e, T).$$

We evaluated the output of the agglomeration using both schemes described above (cf. Fig. 5). In Fig. 5a we display a typical result of the hierarchical clustering, stating in this case the maxima for precision and recall at 83% and 72%, respectively. One can see that the other clustering method (cf. Fig. 5b) reaches almost the same precision value (76%) while the maximum recall is significantly lower (53%). Despite the worse overall results, this algorithm might be preferable in some cases as it obviously reaches the optimum a lot faster. These diagrams also illustrate the agglomeration process – starting with the initial component set and finishing with one huge cluster. As more and more components get agglomerated, the granularity of the clustering approaches its optimum. Further grouping generates too large clusters and, consequently, lead to worse precision and recall values.

For the evaluation of the quality of the agglomeration of textual elements we use the method introduced in the context of the ICDAR 2005 Text Locating Competition [Lucas, 2005]. That way we produce comparable and comprehensible evaluation results.



(a) Touching items

(b) False agglomeration for multi-line texts

Figure 6: Segmentation challenges

5 Conclusion

In this paper we presented a segmentation approach for handwritten whiteboard notes that is based on a three-stage processing strategy. First we extract connected components, which are then classified w.r.t. belonging to known graphical elements or text. In order to obtain segmentation at word level in the final stage textual elements are merged by an automatic clustering procedure.

By means of a writer-independent experimental evaluation we demonstrated the effectiveness of the proposed approach. We successfully extracted graphical and textual elements of handwritten mindmaps of real-world whiteboard images. Clustering of connected components identified as being text produced reasonable word level hypotheses. The latter can now serve as input for an actual handwriting recognition system.

Analyzing the segmentation results provided by the proposed system certain challenges can be identified that still remain (cf. Fig. 6). As illustrated in Fig. 6a touching connected components need to be separated properly. Furthermore, line separation is required for multi-line text portions before feeding them to an actual recognition system (cf. Fig. 6b).

In our future work we will address the aforementioned issues. Furthermore, we will consider the recognition of the whole structure of the mindmap and the integration of the system with our handwriting recognizer [Plötz et al., 2008].

Acknowledgments

This work was supported by the German Research Foundation (DFG) within project **Fi799/3** or **Pl554/1**, respectively.

References

[Buzan, 2003] Buzan, T. (2003). *The Mind Map Book: Radiant Thinking - Major Evolution in Human Thought*. BBC Active.

[Fletcher and Kasturi, 1988] Fletcher, L. and Kasturi, R. (1988). A robust algorithm for text string separation from mixed text/graphics images. *PAMI*, 10(6):910–918.

[Fujisawa, 2008] Fujisawa, H. (2008). Forty years of research in character and document recognition - an industrial perspective. *Pattern Recognition*, 41(8):2435–2446.

[Garain and Chaudhuri, 2003] Garain, U. and Chaudhuri, B. B. (2003). On machine understanding of online handwritten mathematical expressions. In *International Conference on Document Analysis and Recognition*, pages 349–353, Edinburg, Scotland.

[Gatos et al., 2000] Gatos, B., Mantzaris, S. L., Perantonis, S. J., and Tsigris, A. (2000). Automatic page analysis for the creation of a digital library from newspaper archives. *International Journal on Digital Libraries*, 3(1):77–84.

[Huang and Tan, 2007] Huang, W. and Tan, C. L. (2007). A system for understanding imaged infographics and its applications. In *DocEng: ACM Symposium on Document Engineering*, pages 9–18, Winnipeg, Manitoba, Canada.

[Klink et al., 2000] Klink, S., Dengel, A., and Kieninger, T. (2000). Document structure analysis based on layout and textual features. In *International Workshop on Document Analysis Systems*, pages 99–111, Rio de Janeiro, Brazil.

[Liwicki and Bunke, 2005] Liwicki, M. and Bunke, H. (2005). Handwriting recognition of whiteboard notes. In *Conference of the International Graphonomics Society*, pages 118–122.

[Lucas, 2005] Lucas, S. M. (2005). Text locating competition results. In *International Conference on Document Analysis and Recognition*, pages 80–85, Seoul, Korea.

[Mollag et al., 2009] Mollag, A. F., Basu, S., Nasipuri, M., and Basu, D. K. (2009). Text/graphics separation for business card images for mobile devices. In *International Workshop on Graphics Recognition*, pages 263–270. Springer-Verlag.

[Namboodiri and Jain, 2007] Namboodiri, A. M. and Jain, A. K. (2007). *Document Structure and Layout Analysis in Digital Document Processing: Major Directions and Recent Advances, B. B. Chaudhuri (ed.)*. Springer-Verlag.

[Niblack, 1986] Niblack, W. (1986). *An introduction to digital image processing*. Prentice Hall.

[Nomura et al., 2003] Nomura, A., Michishita, K., Uchida, S., and Suzuki, M. (2003). Detection and segmentation of touching characters in mathematical expression. In *International Conference on Document Analysis and Recognition*, pages 126–130, Edinburgh, Scotland.

[Oliveira and Lins, 2007] Oliveira, D. M. and Lins, R. D. (2007). Tableau - processing teaching-board images acquired with portable digital cameras. In *International Workshop on Camera-Based Document Analysis and Recognition*, pages 79–86, Curitiba, Brazil.

[Plötz et al., 2008] Plötz, T., Thurau, C., and Fink, G. A. (2008). Camera-based whiteboard reading: New approaches to a challenging task. In *International Conference on Frontiers in Handwriting Recognition*, pages 385–390, Montréal, Quebéc, Canada.

[Roy et al., 2009] Roy, P. P., Pal, U., and Lladós, J. (2009). Seal detection and recognition : An approach for document indexing. In *International Conference on Document Analysis and Recognition*, pages 101–105, Barcelona, Spain.

[Sauvola et al., 1997] Sauvola, J., Seppanen, T., Haapakoski, S., and Pietikainen, M. (1997). Adaptive document binarization. In *International Conference on Document Analysis and Recognition*, volume 1, pages 147–152, Ulm, Germany.

[Sellen and Harper, 2002] Sellen, A. and Harper, R. (2002). *The Myth of the Paperless Office*. MIT Press.

[Shafait et al., 2008] Shafait, F., Keysers, D., and Breuel, T. M. (2008). Efficient implementation of local adaptive thresholding technique using integral images. In *Document Recognition and Retrieval*, pages 101–105, San Jose, California, USA.

[Tapia and Rojas, 2005] Tapia, E. and Rojas, R. (2005). Recognition of on-line handwritten mathematical expressions in the e-chalk system - an extension. In *International Conference on Document Analysis and Recognition*, pages 1206–1210, Seoul, Korea.

[Tombre and Lamiroy, 2008] Tombre, K. and Lamiroy, B. (2008). Pattern recognition methods for querying and browsing technical documentation. In *Iberoamerican Congress on Pattern Recognition*, pages 504–518, Havana, Cuba.

[Tombre et al., 2002] Tombre, K., Tabbone, S., Pélissier, L., Lamiroy, B., and Dosch, P. (2002). Text/graphics separation revisited. In *International Workshop on Document Analysis Systems*, pages 200–211, Princeton, NJ, USA. Springer-Verlag.

[Vajda et al., 2009] Vajda, S., Roy, K., Pal, U., Chaudhuri, B. B., and Belaïd, A. (2009). Automation of Indian postal documents written in Bangla and English. *International Journal of Pattern Recognition and Artificial Intelligence*, 23(8):1599–1632.

[van Beusekom et al., 2007] van Beusekom, J., Keysers, D., Shafait, F., and Breuel, T. M. (2007). Example-based logical labeling of document title page images. In *International Conference on Document Analysis and Recognition*, volume 2, pages 919–923, Curitiba, Brazil.

[Yoshida et al., 2006] Yoshida, D., Tsuruoka, S., Kawanaka, H., and Shinogi, T. (2006). Keywords recognition of handwritten character string on whiteboard using word dictionary for e-learning. In *International Conference on Hybrid Information Technology*, pages 140–145.

[Yuan and Tan, 2005] Yuan, B. and Tan, C. L. (2005). Multi-level component grouping algorithm and its applications. In *International Conference on Document Analysis and Recognition*, pages 1178–1181, Seoul, Korea.

A Appendix

A.1 Original Becker features [Lucas, 2005]

Contrast:

$$F_{\text{contrast}} = \min\left(1.0, \frac{|\mu_{fg} - \mu_{bg}|}{20}\right)$$

Edge density:

$$I_{\text{density}}(x,y) = \sqrt{\left(\frac{1}{121} \cdot \sum_{p \in N_{11}(x,y)} I_{\Delta}(p)^2\right)}$$

$$avg\_density = \frac{\sum_{y=0}^{height-1} \sum_{x=0}^{width-1} I_{\text{density}}(x,y)}{width \cdot height}$$

$$F_{\text{edge\_density}} = \min\left(1.0, \frac{avg\_density}{10}\right)$$

Homogeneity:

$$F_{\text{homogeneity}} = 1.0 - \frac{\min\left(1.0, \frac{\sigma_{fg}}{180}\right) + \min\left(1.0, \frac{\sigma_{bg}}{300}\right)}{2}$$

Histogram overlap:

$$F_{\text{hist\_overlap}} = 1.0 - \frac{num_{overlap}}{num_{bg}}$$

A.2 Shape features

Canny edge intensity: The average of the intensity of an edge of the canny edge image I_{canny} within the area of a connected component's bounding box.

$$F_{\text{canny\_intensity}} = \frac{\sum_{x_{cc}}^{width_{cc}-x_{cc}} \sum_{y_{cc}}^{height_{cc}-y_{cc}} I_{\text{canny}}(x,y)}{weight_{cc} \cdot height_{cc}}$$

Number of foreground gray levels: The number of gray values in the foreground (i.e. on the connected component) of the bounding box of a graphical element.

$$F_{\text{fg\_gray\_values}} = \frac{\sum_{i;hist_{fg}[i]>0} 1}{\sum_{i;hist_{fg}[i]>0} 1 + \sum_{i;hist_{bg}[i]>0} 1}$$

Foreground mean gray level:

$$F_{\text{fg\_mean}} = \frac{\sum_{i} hist_{fg}[i] \cdot i}{\sum_{i;hist_{fg}[i]>0} 1}$$

Relative amount of gradient orientations: Using the histogram $hist_{angles}$ of the angles of the gradient image, the number of the angles appearing at least once is calculated.

$$F_{\text{gradient\_orientations}} = \frac{\sum_{i;hist_{angles}[i]>0} 1}{360}$$

Relative amount of foreground pixels: The number of pixels of the connected component divided by its area.

$$F_{\text{fg\_pixels}} = \frac{\sum_{i} hist_{fg}[i]}{width_{cc} \cdot height_{cc}}$$

Standard deviations: As the standard deviation is a measure of dispersion of data, in our case these features select irregular/textured connected components.

– **Gray level intensity:**

$$F_{\text{gray\_level\_deviation}} = \sigma(I(cc))$$

– **Sobel gradient orientation:**

$$F_{\text{gradient\_orientation\_deviation}} = \sigma(I_{\text{Sobel\_directions}}(cc))$$

– **Sobel gradient magnitude:**

$$F_{\text{gradient\_intensity\_deviation}} = \sigma(I_{\text{Sobel\_magnitudes}}(cc))$$