Form Classification and Retrieval using Bag of Words with Shape Features of Line Structures

Florian Kleber, Markus Diem and Robert Sablatnig[a]

[a] Computer Vision Lab,
Vienna University of Technology,
1040 Vienna,
Austria

ABSTRACT

In this paper a document form classification and retrieval method using Bag of Words and newly introduced local shape features of form lines is proposed. In a preprocessing step the document is binarized and the form lines (solid and dotted) are detected. The shape features are based on the line information describing local line structures, e.g. line endings, crossings, boxes. The dominant line structures build a vocabulary for each form class. According to the vocabulary an occurence histogram of structures of form documents can be calculated for the classification and retrieval. The proposed method has been tested on a set of 489 documents and 9 different form classes.

Keywords: Layout Analysis, Form Classification, Bag of Words

1. INTRODUCTION

The classification of form documents allows automated extraction of filled-in data in form processing systems.[1,2] The retrieval of forms allows grouping and indexing of entire records due to the knowledge of the composition of records. A form class like e.g. *Table of Content* can be at the beginning of a record and infers about the rest of the record's content. According to Duygulu and Atalay[2] *"a form is a structured document which is composed of horizontal and vertical lines, preprinted data and user filled-in data"*. Due to the syntactical knowledge (defined structure) of a form type semantic information can be extracted (automated processing of the machine- or hand-printed user filled-in data[1]).

This paper deals with form classification and retrieval of Stasi documents. After the fall of the Berlin Wall[3] 600 million-odd snippets of Stasi documents have been discovered. The documents were fragmented in 1889 when Stasi officers tried to destroy secret files. The Fraunhofer Institute for Production Systems and Design Technology (IPK) Berlin is investigating methods for the reconstruction and has developed a system for the reassembling of torn Stasi-files.[3,4] After the reconstruction of torn, and preserved Stasi documents a grouping and indexing of single documents to entire records is done by archivists. Automated clustering methods based on the content (paper type, hand-written vs. machine printed, paper color, form type) support the work of the archivists to group entire Stasi records automatically. Form documents are used to get information of the content of a record (e.g. index, table of contents) and form classification allows a concrete search for a specified form. In contrast to current form processing systems the proposed method must be able to deal with degraded and incomplete form documents as well as with form structure variations within a single form class (template of certain Stasi forms can vary over the time). Problems of form identification are detailed by Arlandis et al.[1] and comprise forms with changes in layouts compared to former releases (similar layouts), introduced document skew and different print qualities, color and paper types. Hand-written filled in data can affect (global) form features and the occurence of unknown form types can cause additional errors in form processing systems.[1] Figure 1 shows exemplarily a reconstructed form document. It can be seen that due to gaps or missing parts form lines are broken and missaligned.

Further author information:
E-mail: kleber@caa.tuwien.ac.at, Telephone: +43 1 58801 18354

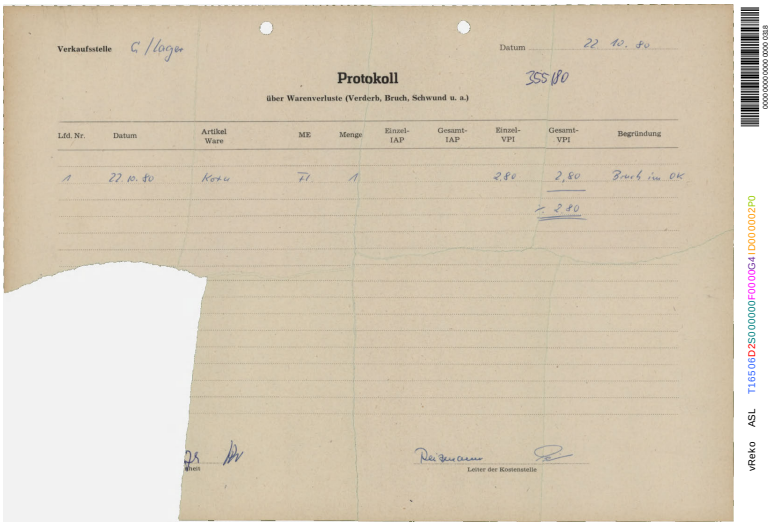

Figure 1. Real World example (exemplarily) of a reassembled form document of the Fraunhofer IPK reconstruction system.

The proposed method is based on shape features of the line information of different Stasi form types. For each form type the shape features are clustered to create a dictionary. Based on the dictionary an occurence histogram of shape features is calculated and compared to the form templates for classification (see Bag of Words (BoW)[5]). Different scales for the shape features are used to additionally represent the hierarchical structure of the form. The proposed method allows a form classification on the one hand, and a form retrieval of similiar forms based on a defined template, on the other.

This paper is organized as follows: Section 2 reviews the state-of-the-art of form classification methods. In Section 3 the proposed form classification and retrieval method is presented while Section 4 presents the results of the algorithm on a test set of Stasi form documents. Finally, a conclusion is given in Section 5.

2. RELATED WORK

In the literature current form classification methods use a hierarchical representation of form documents[2,6,7] based on structural features like frames, cells, lines or blocks.[1] Mandal et al.[6] propose a hierarchical method which uses global shape features (2nd and 4th order central moments of the average horizontal and vertical projection) to reduce the search space within a database. As a second step the form is classified by structural features (crossing types and relative positions of lines and crossings represented by a relationship matrix), which are compared to candidate forms stored in a database. The classification is based on the work of Fan et al.[8] Moments of horizontal and vertical projections as a global shape feature are not robust against distortions like missing and additional lines (binarization errors) in degraded documents. Thus, the pre-classification step can fail on noisy documents.

Duygulu and Atalay[2] represent form documents by lines and a XY-tree. The approach allows the identification or the retrieval of similar forms by calculating the distance between the XY-trees which is a hierarchical representation of the form document. It is stated that the proposed method can deal with form structure variations of a form type, since *"logically identical forms are expected to have the same or similiar hierarchical tree structure"*.[2] The method is restricted to form types that consist of boxes which are used to build the XY-tree. A different approach that uses the cell structure (size and position of cells/boxes) and a geometric hashing for form identification is presented in Shimotsuji and Asano.[9]

Arlandis et al.[10] determine a set of discriminant landmarks (areas) for a reference image with respect to all other form classes. The classification is based on a distance function and a discriminant threshold. It is stated that the proposed method can identify similarily filled-in forms and is tested with 7 different form classes

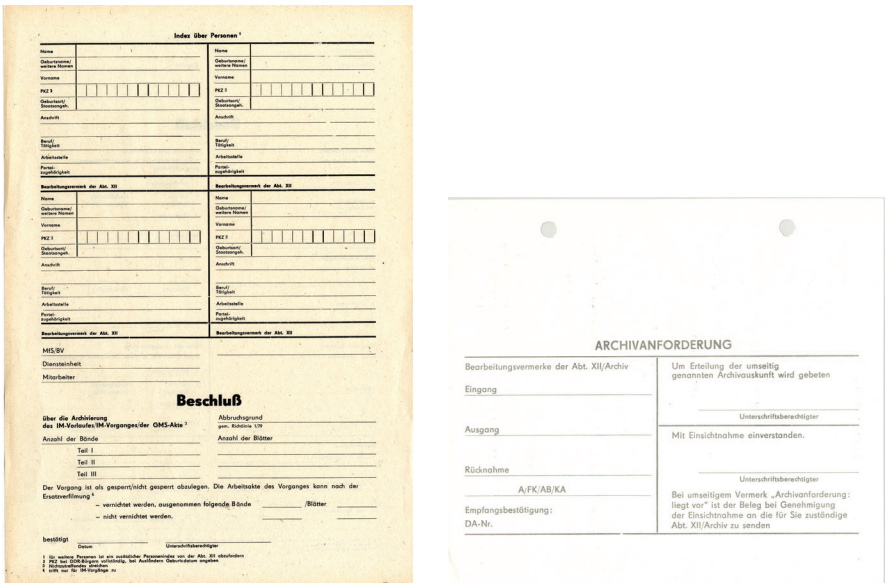

Figure 2. Form samples.

Ohtera and Horiuchi[11] are using the Histogram of the Hough-space for faxed form identification. The Hough space is exploited to determine the skew and to make a position adjustment by the center point of the Hough-space. The similarity of the Hough histograms for horizontal and vertical lines is used to classify the form document. The method has been tested on 10 different form types and it is stated that written characters must be preseparated from lines as a preprocessing step to avoid errors.

Byun et al.[12] determines distinctive form areas by partitioning the image into rectangular areas based on horizontal and vertical lines. A disparity score is calculated using Dynamic Programming (DP) to select the matching areas. The classification is based on the disparity values of the areas determined. The methodology has been tested on 246 form images with 6 training forms.

Saund[13] defines line crossings and endings (called junction/termination types) as described by Fan et al.[8] To be more discriminative links between these structural elements are established and represented as a data graph consisting of the junction/termination points as nodes and the links between them. For each form type a discriminative graph representation is choosen in the training. The Common-Minus-Difference (CMD) is used as a similarity measure and it is stated that the proposed method has a classification accuracy of 100% on all 11185 images of the NIST SpecialDatabase2 and SpecialDatabase6.[14] A graph lattice for representation of form documents and a BoW approach is used for classifcation.

3. METHODOLOY

The proposed method deals with the classification and retrieval of degraded form documents based on the line information (solid and dotted). Dominant line structures (line endings, crossings, T-junctions, ...) of a form type are determined and represent a dictionary for each form class. Based on the dictionary a feature histogram for a form can be calculated which allows a classification of the form type by comparing the histogram with the form-class histograms. Different scales allow for describing local as well as global structures of forms.

Due to the representation of forms as histograms of line structures (shapes) form template variations can be correctly classified. Thus, forms having the same line or similiar line structure and only changes within the text cannot be distinguished. Since broken or missing lines result only in minor changes in the feature histogram, degraded documents can be correctly classified. Figure 2 shows examplarily 2 forms occuring in the Stasi dataset. The size of the form ranges from approximately DIN A6 to DIN A4.

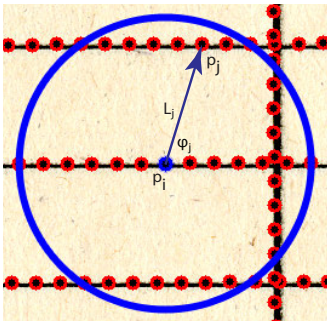

Figure 3. Sampled lines with a shape region with a scale of 120 pixels.

The following Section 3.1 describes the preprocessing, while Section 3.2 explains the structural line features. Section 3.3 outlines the classification using BoW.

3.1 Preprocessing

In the preprocessing step the skew is estimated using a combination of a gradient based approach and a Focused Nearest Neighbour Clustering (FNNC) of interest points.[15] The gradient based methodology is stable for forms with solid lines, while the FNNC is used if dotted lines or text determine the orientation of a form. The skew estimation is described in detail in Diem et al.[15] The estimated skew is a global skew. Small variations of single document fragments (miss-aligned) does not effect the structural features for the form classification. After the correction of the skew the image is binarized using Su et al.'s approach.[16] This method uses the image contrast defined by the local minimum and maximum within a local region and can be applied to degraded documents.

Based on the binary image lines are detected by analyzing horizontal and vertical run lengths.[17,18] Analyzing run lengths allow to remove hand-written and printed text and to close small gaps occuring due to ascenders and descenders of text. For the detection of dotted lines a template matching is done to determine dots. Based on the dots a nearest neighbour clustering is done for dotted line detection. Thus, the line detection and the differentiation between dotted lines and lines is performed automatically based on the binarized image.

3.2 Structural Features

The structural line features to describe lines and line crossings are modified Shape Context features proposed by Belongie et al.[19] The lines detected in the pre-processing step are the basis for the feature computation. Figure 3 shows a detail of a lined form with one vertical line. At each (sampling) point the line structure (shape) is calculated within a circular shape-region.

All solid lines l_s and all dotted lines l_d are sampled equally at a distance of ds pixel. The sampling distance ds defines the courseness of the line structure and is set to 10 pixel (spacing distance of dotted lines). Thus, all lines $l_{s,d}$ are represented by sample points p. For each point p_i an orientation histogram $H_i(\phi)$ is defined as line structure (shape feature):

$$NP = \{p_j \mid \|p_j - p_i\| < r\} \tag{1}$$

NP are defined as all Neighbour Points p_j within the radius r. The radius defines the scale of the shape and thus the geometric complexity. All neighbour points in NP are represented by their polar coordinates (L, ϕ) relative to p_i (center point of the current shape feature):

$$L_j = \sqrt{(x_j - x_i)^2 + (y_j - y_i)^2} \tag{2}$$

$$\phi_j = \arctan \frac{y_j - y_i}{x_j - x_i} \tag{3}$$

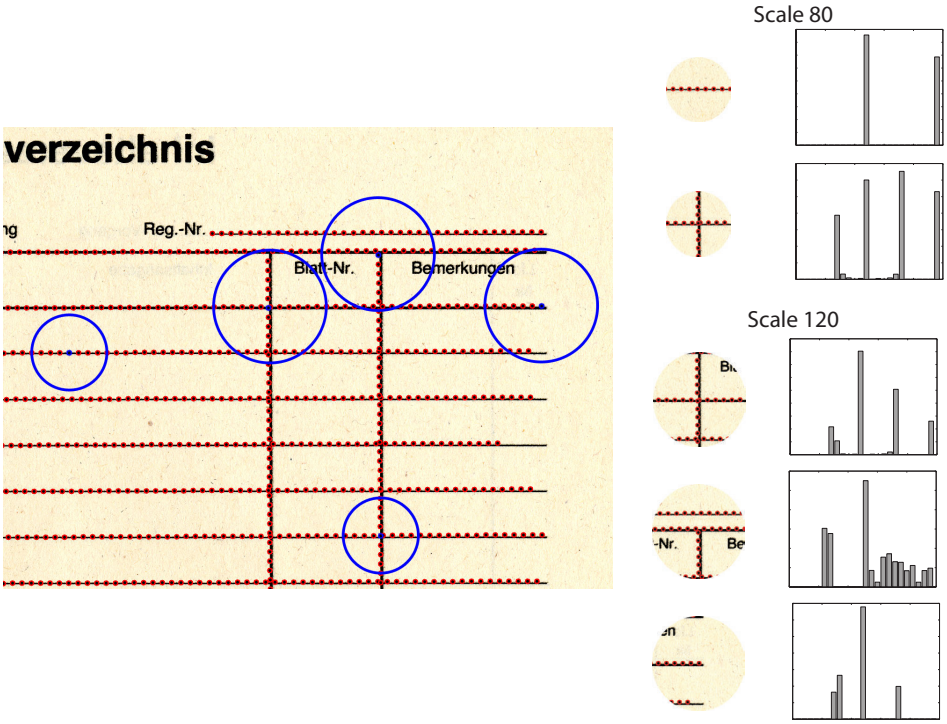

Figure 4. Structural shape features for sampled form lines.

where x and y are the image coordinates of $p_{i,j}$. Figure 3 illustrates the representation of a neighbour point p_j of p_i by its polar coordinates. All points of NP - defined by their angle and distance relative to the shape center - are accumulated into an orientation histogram $H_i(\phi)$ which is weighted by the distance L_j of every neighbour point. The weighted orientation histogram is defined as line structure.

Closer points to the center have less influence on the shape, thus weighting the orientation by the distance leads to more stable results. If the current point p_i belongs to a dotted line l_d, the orientation histogram is weighted by negative distances to distinguish between solid and dotted lines:

$$H_i(\phi) = -H_i(\phi) \; if \; p_i \in l_d \tag{4}$$

Different scales of shape regions are applied to represent local as well as global line structures. The distances L are normalized by the base scale (smallest scale). Figure 4 shows a detail of a form with 2 shape regions with a scale of 80 pixel (base scale) and 3 shape regions with a scale of 120 pixel, and its belonging structural features (weighted orientation histogram $H_i(\phi)$). The final line structure feature has a dimension of 24 angular bins (every 5 degrees) which locally describes the shape of binary solid, dotted lines and line junctions, robust against distortions like gaps and broken lines.

Compared to Mandal et al.[6] and Fan et al.[8] the line features are not restricted to a certain number of crossings (e.g. 9^6). Shapes with a scale smaller than the line spacing can be assumed as the shape primitves defined by Fan et al.[8] The lines are sampled every 10 pixels to reduce the computational effort. The proposed features are robust against broken lines, since missing points do not affect the shape. Figure 5 shows a junction, the current point p_i (red) and the corresponding structural features. All blue points represent the sample points p_j within the search window of p_i. It can be seen that the feature is robust against broken junctions or gaps with different sizes.

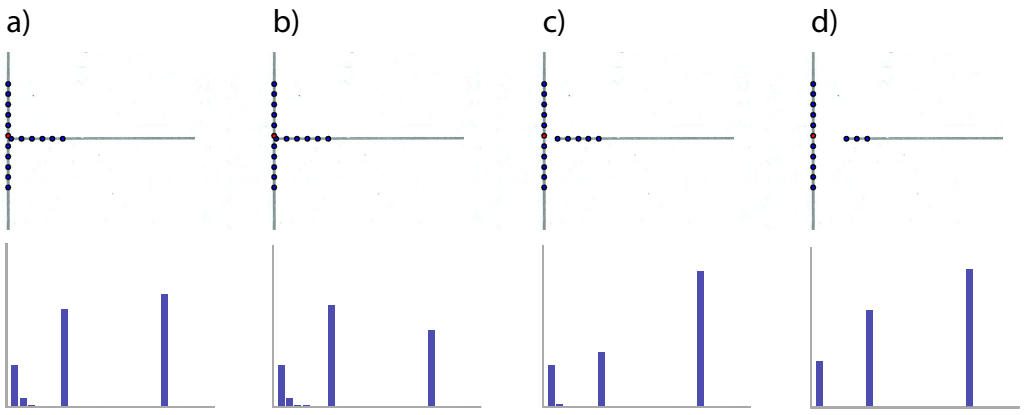

Figure 5. T-Junction as a detail of a form with the sampled line points and the structural features below: The red point shows the current point p_i; the blue points are all sample points p_j within in the search window of p_i. In a) the junction is not connected, whereas b) has a connected junction. c) and d) show a broken junction with gaps.

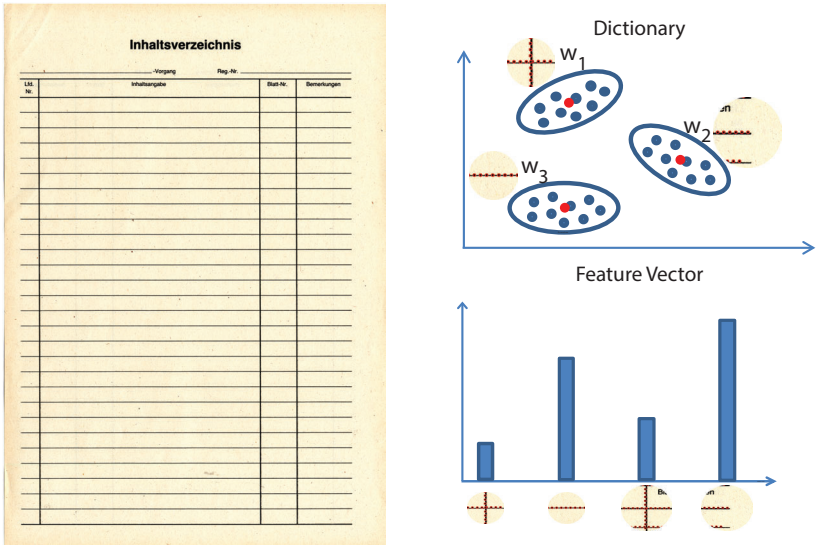

Figure 6. Dictionary (feature space with clustered *words*) and a feature vector of a form.

3.3 Classification using BoW

The classification and retrieval is based on the BoW approach, proposed by Csurka et al.[5] For each form type the structural features are calculated on a training dataset consisting of forms of the same class. A codebook (dictionary) for every form class is created by clustering the structural features using k-means and the cluster centers w_i form the words of the dictionary of size i. The words of all dictionaries represent frequent structures of all form types. Figure 6 (upper part) illustrates the codebook generation. The blue dots represent the form structures in the feature space, and the cluster centers (ellipse center point, red dots) build the final codebook.

Each form type is represented by a feature vector. The structural features of a form s_j are calculated and are assigned to the cluster center w_i (word) with the smallest (Euclidian) distance $\min_i \|s_j - w_i\|$; thus building a histogram of occurences of the cluster centers (words). Figure 6 (lower part) shows a typical form and the pre-determined codebook (frequent structures). Based on the codebook the histogram of occurences describes the final feature vector.

For the classification the occurence histogram (feature vector) of the unknown document is compared with every occurence histogram of the trained form classes. The class with the smallest distance defines the form type. For the form retrieval an arbitrary form is chosen and the feature vector is created. It is compared with

each feature vector of the documents in the dataset using the euclidean distance. The distances are sorted which defines a ranking for the similarity of the chosen form document. As parameter for the classification the dictionary size (number of clusters defined for the k-means algorithm) has to be choosen. Tests showed (see Section 4) that a dictionary size of 120 leads to the best results for the evaluated form types.

4. RESULTS

For the experiments a form dataset consisting of 9 different form types and a *non-form* class from the Stasi dataset has been created. Figure 2 shows exemplarily 2 form types. The size of the forms ranges appr. from DIN A6 to DIN A4 with a resolution of 300 dpi. The training dataset has at least 4 training forms for every class. The testset consists of 489 documents, comprising 287 forms and 202 non-form documents. The distribution of the number of documents within a form class represents the amount of the form types chosen within a single Stasi IM-record (unofficial collaborator file).

Table 1. The rows of the confusion matrix show the groundtruth labels (9 different form types and a class which contains no form documents), while the columns represent predicted labels (e.g. 2 forms of the type 0 ("Table of Contents") are falsely classified as form type 4). A scale size of 120, 580 and 840 (size of the shape regions) pixels, and a dictionary size of 120 words have been set. Overall accuracy: 87.11% (without non-form class) rsp. 80.98%.

	predicted										
	0	1	2	3	4	5	6	7	8	no form	#
0	47	3	1	2	2	·	12	·	·	·	67
1	·	39	3	·	·	·	·	·	·	·	42
2	·	·	70	·	·	1	·	·	·	·	71
3	·	·	·	26	·	·	·	·	·	·	26
4	·	·	·	1	9	2	·	·	·	·	12
5	·	·	·	6	·	13	·	·	·	·	19
6	·	·	·	·	·	·	31	·	·	·	31
7	·	·	·	3	·	1	·	1	·	·	5
8	·	·	·	·	·	·	·	·	14	·	14
no form	7	·	·	19	9	16	2	1	2	146	202

Table 2. Classification with a single scale size of 120 pixels (size of the shape regions), and a dictionary size of 120 words has been set. Overall accuracy: 80.35% (without non-form class) rsp. 77.91%.

	predicted										
	0	1	2	3	4	5	6	7	8	no form	#
0	42	6	3	·	·	·	13	1	·	2	67
1	·	37	5	·	·	·	·	·	·	·	42
2	·	·	60	·	5	6	·	·	·	·	71
3	·	·	·	26	·	·	·	·	·	·	26
4	·	·	·	1	8	3	·	·	·	·	12
5	·	·	·	6	·	13	·	·	·	·	19
6	·	·	·	·	·	·	30	1	·	·	31
7	·	·	·	2	1	1	·	1	·	·	5
8	·	·	·	·	2	·	·	·	12	·	14
no form	5	2	1	19	7	11	3	1	1	152	202

Table 3. Classification results regarding scales (dictionary size of 120).

scales [pixel]	accuracy (incl. non-forms) [%]	accuracy (w/o non-forms) [%]
120	77.91	80.35
120, 580	78.12	81.53
120, 580, 840	80.98	87.11

Table 4. Classification results regarding dictionary size (scales 120, 580, 840).

dictionary size	accuracy (incl. non-forms) [%]	accuracy (w/o non-forms) [%]
100	78.73	83.62
120	80.98	87.11
140	80.78	87.46

For evaluation, the scale (size of the shape region) and the number of scales as well as the dictionary size have been evaluated. A dictionary size of 120 words leads to the best results. Codebooks with less words combine different structural features within a single cluster, and a higher number of structural features causes sparse feature vectors.

Table 1 shows the result of the classification with a dictionary size of 120, and 3 different scales (size of the shape region) of the structural features comprising 120, 580 and 840 pixels. The classification regarding only forms has an overall accuracy of 87.11% and the accuracy of the classification including the non-form class is 80.98%. Missclassified documents of the non-form class are documents which have a similiar structure compared to forms (e.g. lined paper can be classified as form *"Table of Contents"*, if the lined structure of the paper is segmented).

Table 2 shows the confusion matrix with the same dictionary size of 120 words, if only a single scale (120 pixel) is applied. It can be seen that the overall accuracy drops from 87.11% to 80.35%, since global structures are not represented (leading to ambigous feature vectors). Table 3 summarizes the classification results regarding different scales, whereas Table 4 gives the classification results regarding different dictionary sizes at the same scales. A smaller dictionary size combines similar structures (clusters in the feature space) resulting in less descriptive features, and thus in an accuracy of 83.62% (without non-forms) compared to an accuracy of 87.11%.

5. CONCLUSION

In this paper a form classification and retrieval method robust against degraded documents and forms with slight variations in the layout has been proposed. A form is presented by a histogram of structural features of lines (solid and dotted) which have been trained offline for every form class. The method has been tested on Stasi documents with 9 different form types and achieved an overall accuracy of 87.11%. As future work features based on the layout of the (pre-) printed text will be combined with the proposed structural features. The combination will allow to classify also forms without or only sparse line information. Additionally forms differing only in the layout of the preprinted data can be correctly classified by combining layout features with the line information. Future tests will also comprise reconstructed documents (see Figure 1).

ACKNOWLEDGMENTS

The authors would like to thank the Fraunhofer-Institute for Production Systems and Design Technology (IPK), Berlin for supporting the work.

REFERENCES

[1] Arlandis, J., Perez-Cortes, J.-C., and Ungria, E., "Identification of Very Similar Filled-in Forms with a Reject Option," in [*Document Analysis and Recognition, 2009. ICDAR '09. 10th International Conference on*], 246 –250 (jul. 2009).

[2] Duygulu, P. and Atalay, V., "A hierarchical representation of form documents for identification and retrieval," *IJDAR* **5**(1), 17–27 (2002).

[3] Nickolay, B. and Schneider, J., [*Virtuelle Rekonstruktion "vorvernichteter" Stasi-Unterlagen. Technologische Machbarkeit und Finanzierbarkeit - Folgerungen für Wissenschaft, Kriminaltechnik und Publizistik*], vol. 21, 11–28 (Berlin, 2007).

[4] Schneider, J. and Nickolay, B., "The Stasi Puzzle," *Fraunhofer Magazine, Special Issue* **1**, 32–33 (2008).

[5] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C., "Visual categorization with bags of keypoints," *In Workshop on Statistical Learning in Computer Vision, ECCV* , 1–22 (2004).

[6] Mandal, S., Chowdhury, S., Das, A., and Chanda, B., "A hierarchical method for automated identification and segmentation of forms," in [*Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*], 705 – 709 Vol. 2 (aug.-1 sept. 2005).

[7] Heroux, P., Diana, S., Ribert, A., and Trupin, E., "Classification method study for automatic form class identification," in [*Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on*], **1**, 926 –928 vol.1 (aug 1998).

[8] Fan, K.-C., Wang, Y.-K., and Chang, M.-L., "Form document identification using line structure based features," in [*Document Analysis and Recognition, 2001. Proceedings. Sixth International Conference on*], 704 –708 (2001).

[9] Shimotsuji, S. and Asano, M., "Form identification based on cell structure," in [*Pattern Recognition, 1996., Proceedings of the 13th International Conference on*], **3**, 793 –797 vol.3 (aug 1996).

[10] Arlandis, J., Perez-Cortes, J.-C., and Ungria, E., "Identification of Very Similar Filled-in Forms with a Reject Option," in [*Proceedings of the 2009 10th International Conference on Document Analysis and Recognition*], *ICDAR '09*, 246–250, IEEE Computer Society, Washington, DC, USA (2009).

[11] Ohtera, R. and Horiuchi, T., "Faxed Form Identification using Histogram of the Hough-Space," in [*Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 2 - Volume 02*], *ICPR '04*, 566–569, IEEE Computer Society, Washington, DC, USA (2004).

[12] Byun, Y., Yoon, S., Choi, Y., Kim, G., and Lee, Y., "An Efficient Form Classification Method Using Partial Matching," in [*AI 2001: Advances in Artificial Intelligence*], Stumptner, M., Corbett, D., and Brooks, M., eds., *Lecture Notes in Computer Science* **2256**, 291–300, Springer Berlin / Heidelberg (2001).

[13] Saund, E., "A graph lattice approach to maintaining dense collections of subgraphs as image features," in [*2011 International Conference on Document Analysis and Recognition (ICDAR)*], 1069–1074 (2011).

[14] Dimmick, D., Garris, M., and Wilson, C., "Structured forms database," *Technical Report Special Database 2, SFRS, National Institute of Standards and Technology* (2001).

[15] Diem, M., Kleber, F., and Sablatnig, R., "Skew Estimation of Sparsely Inscribed Document Fragments," in [*Proc. of 10th IAPR International Workshop on Document Analysis Systems (DAS 2012)*], (2012).

[16] Su, B., Lu, S., and Tan, C. L., "Binarization of historical document images using the local maximum and minimum," in [*DAS*], 159–166 (2010).

[17] Zheng, Y., Liu, C., Ding, X., and Pan, S., "Form frame line detection with directional single-connected chain," in [*Proceedingsof the Sixth International Conference onDocument Analysis and Recognition (ICDAR)*], 699 –703 (2001).

[18] Diem, M., Kleber, F., and Sablatnig, R., "Document Analysis Applied to Fragments: Feature Set for the Reconstruction of Torn Documents," in [*Proceedings of the 9th International Workshop on Document Analysis Systems*], Doermann, D., Govindaraju, V., Lopresti, D., and Natarajan, P., eds., 393–400 (June 2010).

[19] Belongie, S., Malik, J., and Puzicha, J., "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(4), 509–522 (2002).