Development and Evaluation of Text Localization Techniques Based on Structural Texture Features and Neural Classifiers

Christos Emmanouilidis[a], Costas Batsalas and Nikos Papamarkos[b]
[a]*Computational Systems Unit CETI/R.C.ATHENA, 58 Tsimiski St. Xanthi, 67100, Greece, Email: chrisem@ceti.gr*
[b]*Image Processing and Multimedia Laboratory, Dept. of Electrical & Computer Engineering, Democritus Univ. of Thrace, Xanthi, 67100, Greece, Email: papamark@ee.duth.gr*

Abstract

This paper presents a text localization approach for binarized printed document images. Emphasis is given to the feature extraction and feature selection stages. In the former, several document structure elements and spatial features, likely to convey useful information, are extracted. In the latter, evolutionary multi-objective feature selection is employed to identify combinations of features with simultaneous good performance in terms of text localization sensitivity and specificity. The selected features are applied to a range of classifiers. Performance results over document image sets from known databases are presented, employing the classifiers with or without feature selection. The results suggest that the hybrid techniques, which utilize the classifiers in combination with the customized pre-processing, feature extraction and feature selection stages, exhibit promising performance on a range of document images.

1. Introduction

A key problem in document image analysis is the segmentation of a mixed-type document into text and non-text regions, known as text localization. Text localization in printed document images is a key processing stage for page layout analysis (PLA) and in particular for text information extraction from document images [1]. Variations in text positioning, sizes, styles, as well as in image quality and logical structure, pose considerable challenges to text localization techniques. Most such approaches can be considered to fall under two broad categories, region-based and texture-based. The former usually employ a bottom-up approach, wherein connected components [2-6] or image edges [7, 8] are identified and connected. The latter are in most cases top-down techniques, wherein the image is split into major regions, based on texture properties, wherein the regions are sub-divided to sub-regions and so on [9-12]. Hybrid techniques blending the characteristics of the former two also exist [13-17]. Document region classification into text or non-text can be performed by employing a range of different techniques, including unsupervised [15] or supervised classifiers [18]. Key to the success of such classifiers is the extraction and selection of features which convey relevant discriminatory information. One option is to employ texture features, such as document structure elements (DSE) [19]. Other approaches include: global features for classifying rectangular regions from binary documents [20]; wavelet features [18]; and different types of line crossings [21].

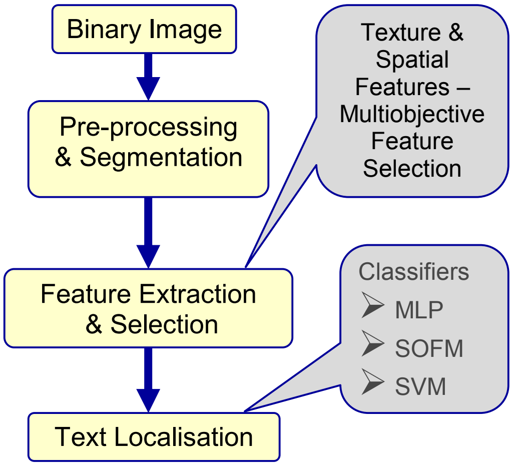

Binary Image
Pre-processing & Segmentation
Feature Extraction & Selection
Text Localisation
Texture & Spatial Features – Multiobjective Feature Selection
Classifiers
MLP
SOFM
SVM

Figure 1. Text Localization Processing Cycle

This paper implements a hybrid technique for text localization in binary printed document images. The work extends a previous approach [19] [15]. Specifically, it employs improved pre-processing; works with both texture and spatial features; applies both supervised and unsupervised classification

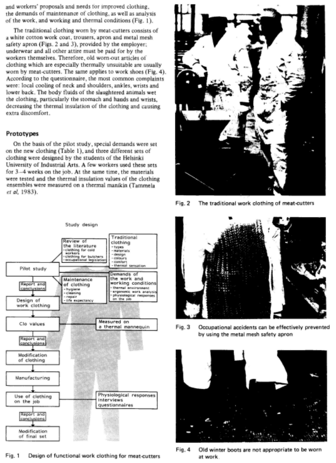
978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.254

strategies; and implements multi-objective feature selection, to simultaneously target classifier sensitivity and specificity [22, 23]. The obtained results suggest that text block extraction is fast, as the new approach can identify text regions while being tolerant of small skew errors, thus making it appropriate for camera-based printed document images, such as those taken from scientific journals. Evaluation is performed over document images taken from known repositories. Section 2 describes the pre-processing cycle; section 3 the feature extraction & selection; indicative results are discussed in section 4, followed by the conclusion.

2. Text Localization Processing Cycle

This paper implements a hybrid technique for text localization in binary printed camera document images, typically met in scientific journals. Binarized images convey less information compared to grayscale or color images and thus are likely to yield inferior results. Even in binarized images, results should vary depending on the binirization technique performance. Bearing in mind the above, we investigate to what extend a simplified hybrid localization approach technique with lower computational requirements can succeed in localizing text in such images, thus no emphasis is given on employing different binarization techniques. Our text localization processing includes four processing stages (**Error! Reference source not found.**): (a) Image binarization; (b) Pre-processing & segmentation; (c) Feature extraction / feature selection; and (d) Text localization.

In most cases, the document image is converted to a binary one using the simple binarization technique of Otsu [24]. However, in poorly illuminated documents, we can also apply more powerful binarization techniques [25]. Next a bottom-up approach is taken, connecting the components of the binary document form. That is, the documents' marks are extracted by enclosing each connected component in an orthogonal bounding box. After appropriate size-filtering, the boxes are extended, resulting in merging neighboring boxes in larger rectangular boxes. One more filtering stage is then applied, followed by the extraction of a series of texture and spatial features. The complete set of features can then be fed into the classifier. Both supervised (MLP, SVM) and unsupervised (SOFM) classifiers are employed. The classifiers may also be employed after a feature selection stage is completed. For that, we employ evolutionary multi-objective feature selection to identify minimal cardinality feature combinations which simultaneously exhibit adequate performance in terms of classifier sensitivity and specificity.The feature selection stage can be employed to achieve similar performance with a much reduced feature subset size. The main processing stages are presented in the following sections.



and workers' proposals and needs for improved clothing, the demands of maintenance of clothing, as well as analysis of the work, and working and thermal conditions (Fig. 1).

The traditional clothing worn by meat-cutters consists of a white cotton work coat, trousers, apron and metal mesh safety apron (Figs. 2 and 3), provided by the employer; underwear and all other attire must be paid for by the workers themselves. Therefore, old worn-out articles of clothing which are especially thermally unsuitable are usually worn by meat-cutters. The same applies to work shoes (Fig. 4). According to the questionnaire, the most common complaints were: local cooling of neck and shoulders, ankles, wrists and lower back. The body fluids of the slaughtered animals wet the clothing, particularly the stomach and hands and wrists, decreasing the thermal insulation of the clothing and causing extra discomfort.

Prototypes

On the basis of the pilot study, special demands were set on the new clothing (Table 1), and three different sets of clothing were designed by the students of the Helsinki University of Industrial Arts. A few workers used these sets for 3–4 weeks on the job. At the same time, the materials were tested and the thermal insulation values of the clothing ensembles were measured on a thermal manikin (Tammela *et al*, 1983).

Fig. 1 Design of functional work clothing for meat-cutters

Fig. 2 The traditional work clothing of meat-cutters

Fig. 3 Occupational accidents can be effectively prevented by using the metal mesh safety apron

Fig. 4 Old winter boots are not appropriate to be worn at work

Figure 2. Image from UW Image Database

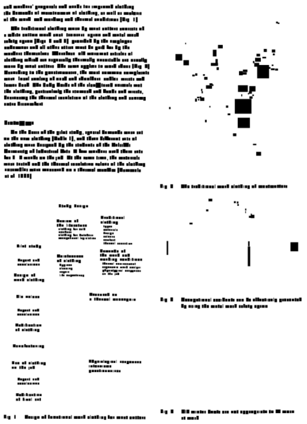


Figure 3. Retained connected components

3. Pre-processing

We employ a typical mark extraction approach by applying connected component analysis on the binary image. Each mark is included in a rectangular bounding box. The bounding boxes' height histogram is then estimated. The histogram peak corresponds to the most frequently appearing height of the document characters.

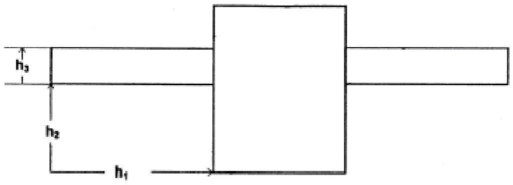


Figure 4. Extension the bounding box

In printed document images, we may assume that a character's bounding box height is likely to lie within the limits $H_{\max} / 2 \le h_i \le 10 H_{\max}$ (Figure 2- Figure 3). Taking all other bounding boxes out, we extend each remaining box towards the left and right by adding two rectangular boxes (**Error! Reference source not found.**), linking the dimensions of the added boxes to the common box height [19].

The length of the added box h_1 is set to $H_{\max}$, so bounding boxes in the same line and distance less than 2 x $H_{\max}$ can be merged. The value h_2 is set to $H_{\max}/2$ and can merge bounding boxes which include characters of high length, such as 'l' and 'h' or characters with tail, like 'p' and 'q', with their neighbor bounding boxes. The value h_3 is set to $H_{\max}/4$, thus merging neighboring bounding boxes that are in text lines with limited skew. The result is the extraction of the text lines of the document. The resulting boxes are then merged into new rectangular boxes. The new connected boxes are forming longer sequences if they correspond to non-isolated characters; otherwise they have a short length. This allows us to apply another filtering, retaining character groupings, requiring the ratio of block length to height is at least 3.5 x $H_{\max}$. The retained boxes can be seen in Figure 5. Most of them would correspond to text blocks but several boxes from images would still be present. The next step is to separate the boxes into text and non-text.

4. Feature Extraction

The classifiers performance strongly depends on the employed set of features. We seek to employ sets of features which are likely to convey strong discriminatory information. The first set consists of texture features, called document structure elements (DSE). Due to space limitations, the reader is referred to [19] for their full description, wherein they were shown to work well when fed to unsupervised classifiers. DSEs consist of binary 3x3 masks and there are 2^9 possible masks that can be used. A subset of 13 of them has been shown to be quite informative [26]. The second group of the texture features consists of 21 features, essentially pairs of neighboring DSEs, employed to resolve harder separation cases. Again due to space limitations the reader is referred to previous work for their description [19]. These two groups compose a set of 34 features, which we have employed in our experiments. Next we go one step beyond and incorporate the above texture features together with 10 spatial features, extracted from the bounding boxes [3]. These are the bounding box (1) Height, (2) Width, (3) Size (Height x Width), (4) eccentricity (Width/Height), (5) Saturation degree in the bounding box (The number of mark pixels to the box size). Additionally, we include four statistical spatial features suggested in [2]:

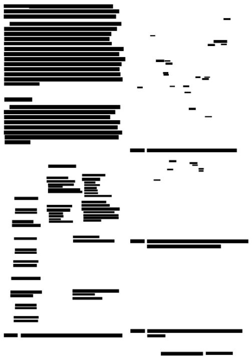


Figure 5. Retained boxes after second filtering

Additionally, we include four statistical spatial features suggested in [2]: (6) The transition variance in the block, ie the transition of background pixels to foreground pixels and vice versa in the row direction.

$$TransVar = \frac{\sum_i \left| t_i - t_{avg} \right|}{t_{avg}}$$

where t_i the transition in ith row and t_{avg} the average transitions of rows. (7) The density variation in the block; this feature computes the distribution of foreground pixels in the block. The block divided in ten equal box and compute

$$DV = \frac{\sum_i \left| d_i - d_{avg} \right|}{d_{avg}}$$

where d_i is the number of foreground pixels in the ith box and d_{avg} the average foreground pixels in the ten boxes. The next two features are related with the edges and calculated based the Sobel edge detection. (8) The edge-direction distribution in the block (9) The edge-direction variance in the block. The final feature taken into account (10) is the shortest distance of the bounding box from its neighboring bounding boxes. The idea here is that if a block includes a text line, then its distance from the other text lines will be smaller compared with the other bounding boxes, which, do not include text lines. Thus we derive a larger set of 44 features, including both texture and spatial features.

5. Multiobjective Feature Selection

Although evidence suggests that the extracted features are likely to convey discriminatory information, it would be useful to seek to identify subsets of features which would perform well both in terms of precision and recall, rather than achieving high classification rate. Our feature selection objective function is:

$$S^{j}(H(\mathbf{x})) = \underset{S^j \in S}{\arg\min}\{\mathbf{J}(\mathbf{x})\} = \underset{S^j \in S}{\arg\min}\left\{\left[\,\left|S^{j}\right|, sensitivity, specificity\right]\right\}$$

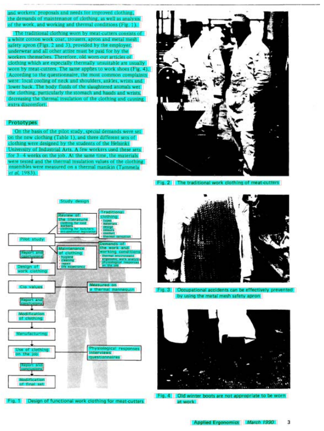
where the objective function $\mathbf{J}(\mathbf{x})$ is a vector consisting of three terms, the subset cardinality, $\left|S^{j}\right|$, and the classifier's specificity and sensitivity. The candidate solutions are non-dominated feature subsets, representing different trade-offs in the three dimensional subset cardinality/sensitivity/specificity objective space. We employ Evolutionary Multi-Objective Feature Selection (EMOFS) to identify such trade-offs in the form of the Elitist Niched Pareto Genetic Algorithm (ENPGA) [22, 23], employing the SSOCF crossover operator, which has been shown to outperform other operators in feature selection [27]. For more details on the approach the reader may refer to the above citations. Thus, several candidate solutions are derived and two subsets with 12 and 14 features with balanced precision and recall are selected.

6. Text Localization Results

We performed a series of experiments, employing supervised (MultiLayer Perceptrons – MLP and Support Vector Machines - SVM), and usupervised (Self-Organised Feature Maps - SOFM) classifiers. 25 images from the Oulu University Database [28] were used to extract the training dataset. This consists of 4706 blocks taken from the complete set of 44 features, ensuring that a variety of fonts and graphics are included. A separate validation set of data is extracted from 10 Oulu Database images and 35 images from the University of Washington English Document Image Database. For the MLP, we employed a standard sigmoid activation with 7, 8, 20 and 25 hidden units, when applying the selected 12, 14, 34 and 44 sets of features respectively, aiming at achieving satisfactory performance with reduced feature subsets, while avoiding overfit. For SVM classification, we employed Schwaighofer's wrapper [29] to Joachim's SVM light [30] implementation, known to be computationally efficient. The SOFM employed in [19] is also employed here. A summary of the results obtained is provided in **Error! Reference source not found.**, showing the employed classifier and the number of features against the achieved performance calculated over the validation set, in terms of classification rate, true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN), with 'positive' cases corresponding to 'text' [27]. The results are averaged over 10 runs for each case, while an example of text localization from a binary image is shown in Figure 6, with the text blocks marked in cyan.

Table 1. Comparative Performance

Classifier (features)	Class. Rate (%)	TP	TN	FP	FN
MLP (34)	94.75	5078 (98.58%)	1937 (85.97%)	73 (1.43%)	316 (14.02%)
MLP (44)	95.08	5065 (98.33)	1975 (87.66%)	86 (1.66%)	278 (12.33%)
MLP (12)	94.96	5011 (97.28%)	2020 (89.65)	140 (2.72%)	233 (10.34%)
MLP (14)	95.03	5036 (97.77%)	2000 (88.77%)	115 (2.23%)	253 (11.22%)
SVM (34)	92.09	5007 (97.2%)	1811 (80.38%)	144 (2.78%)	442 (19.61%)
SVM (44)	95.25	5082 (98.66%)	1970 (87.43%)	69 (1.34%)	283 (12.56%)
SVM (12)	94.06	4919 (95.49%)	2045 (90.76%)	232 (4.5%)	208 (9.23%)
SVM (14)	94.56	4977 (96.62%)	2024 (89.84%)	174 (3.37%)	229 (10.16%)
SOFM (34)	93.64	4952 (96.13%)	1981 (87.88%)	199 (3.86%)	272 (12.72%)
SOFM (44)	93.58	4823 (93.63%)	2106 (93.47%)	328 (6.36%)	147 (6.52%)



and workers' proposals and needs for improved clothing, the demands of maintenance of clothing, as well as analysis of the work, and working and thermal conditions (Fig. 1).

The traditional clothing worn by meat-cutters consists of a white cotton work coat, trousers, apron and metal mesh safety apron (Figs. 2 and 3), provided by the employer; underwear and all other attire must be paid for by the workers themselves. Therefore, old worn-out articles of clothing which are especially thermally unsuitable are usually worn by meat-cutters. The same applies to work shoes (Fig. 4). According to the questionnaire, the most common complaints were: local cooling of neck and shoulders, ankles, wrists and lower back. The body fluids of the slaughtered animals wet the clothing, particularly the stomach and hands and wrists, decreasing the thermal insulation of the clothing and causing extra discomfort.

Prototypes

On the basis of the pilot study, special demands were set on the new clothing (Table 1), and three different sets of clothing were designed by the students of the Helsinki University of Industrial Arts. A few workers used these sets for 3–4 weeks on the job. At the same time, the materials were tested and the thermal insulation values of the clothing ensembles were measured on a thermal manikin (Tammela et al, 1983).

Fig. 1 Design of functional work clothing for meat-cutters

Fig. 2 The traditional work clothing of meat-cutters

Fig. 3 Occupational accidents can be effectively prevented by using the metal mesh safety apron

Fig. 4 Old winter boots are not appropriate to be worn at work

Applied Ergonomics March 1990 3

Figure 6. Text localization in a binary image

The main findings are:

a. The test examples are unevenly distributed between 'text' and 'non-text' classes; therefore the overall misclassification rate can be misleading

b. Feature selection yields adequate performance with reduced feature cardinality and classifier complexity.

c. SVM is known to implicitly perform feature selection by itself; yet multi-objective feature selection reduces performance discrepancy between TP and TN classifications even in the SVM case.

d. SOFM underperforms compared to the supervised techniques; that was to be expected as SOFM training is not guided towards the direct optimization of the misclassification rate.

7. Conclusion

A hybrid technique for text localization in binary printed document images has been presented. The novelty of the approach is multi-fold: it employs improved pre-processing; it works with both texture and spatial features; it is demonstrated with both supervised and unsupervised classification; and it is implemented with multi-objective feature selection, to simultaneously improve classifier sensitivity and specificity. Future work will look at more advanced binarization techniques, whereas a more in-depth investigation of the coupling between multi-objective feature selection and text detection will be pursued.

6. References

[1] K. Jung, K. I. Kim, and A. K. Jain, "Text information extraction in images and video: a survey" *Pattern Recognition,* vol. 37, pp. 977-997, 2004.

[2] W.-Y. Chen and S.-Y. Chen, "Adaptive page segmentation for color technical journals' cover images" *Image and Vision Computing,* vol. 16, pp. 855-877, 1998.

[3] F. M. Wahl, K. Y. Wong, and R. G. Casey, "Block Segmentation and Text Extraction in Mixed Text/Image Documents" *Computer Graphics and Image Processing* pp. 375-390, 1982.

[4] B. Waked, Y. S. Ching, and S. Bergler, "Segmenting document images using diagonal white runs and vertical edges" *International Conference on Document Analysis and Recognition* pp. 194-199, 2001.

[5] H. L. Qing and L. T. Chew, "Newspaper Headlines Extraction from Microfilm Images" *in Proc. of ICPR'02,* vol. 3, pp. 208-211, 2002.

[6] L. O'Gorman, "Document Spectrum for Page Layout Analysis" *IEEETrans. On Pattern Anal. and Machine Intell.,* vol. 15, pp. 1162-1173, 1993.

[7] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR for digital news archive" in *Proceedings of the IEEE Workshop on Content based Access of Image and Video Databases*, Bombay, India, 1998, pp. 52-60.

[8] D. Chen, K. Shearer, and H. Boulard, "Text enhancement with asymetric filter for video OCT" in *Proceedings of the International Conference on Image Analysis and Processing*, Palermo, Italy, 2001, pp. 192-197.

[9] M. D. Matrakas and F. Bortolozzi, "Segmentation and Validation of Commercial Documents Logical Structure" *The International Conference on Information Technology: Coding and Computing* pp. 242-246, 2000.

[10] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video" *IEEE Trans Pattern Anal. Mach. Intell.,* vol. 22, pp. 385-392, 2000.

[11] V. Wu, R. Manmatha, and E. M. Riseman, "TextFinder: an automatic system to detect and recognise text in images" *IEEE Trans Pattern Anal. Mach. Intell.,* vol. 21, pp. 1224-1229, 1999.

[12] K. Jung, "Neural network-based text localization in color images" *Pattern Recognition Letters,* vol. 22, pp. 1503-1515, 2001.

[13] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images" *Pattern Recognition,* vol. 28, pp. 1523-1535, 1995.

[14] T. Ghandi, R. Kasuturi, and S. Antani, "Application of planar motion segmentation for scene text extraction" in *Proc. of ICPR*, 2000, pp. 445-449.

[15] C. Strouthopoulos, N. Papamarkos, and A. E. Atsalakis, "Text extraction in complex color documents" *Pattern Recognition,* vol. 35, pp. 1743-1758, 2002.

[16] E. Badekas, N. Nikolaou, and N. Papamarkos, "Text Localization and Binarization in Complex Color Documents" in *Proc. of MLDM*, 2007.

[17] K. Fan, C. Liu, and Y. Wang, "Segmentation and classification of mixed text/graphics/image documents" *Pattern Recognition Letters,* vol. 15, pp. 1201-1209, 1994.

[18] K. Etemad, D. Doermann, and R. Chellappa, "Multiscale segmentation of unstructured document pages using softdecision integration" *Pattern Analysis and Machine Intelligence, IEEE* vol. 19, pp. 92-96, 1997.

[19] C. Strouthopoulos and N. Papamarkos, "Text identification for document image analysis using a neural network." *Image and Vision Computing* vol. 16, pp. 879–896, 1998.

[20] D. X. Le, G. R. Thoma, and H. Wechsler, "Classification of binary document images into textual or nontextual data blocks using neural network models" *Machine Vision and Applications,* vol. 8 no.5, pp. 289-304, 1995.

[21] S. L. Taylor, R. Fritzson, and J. A. Pastor, "Extraction of data from preprinted forms" *Machine Vision and Applications,* vol. 5, pp. 211-222, 1992.

[22] C. Emmanouilidis, A. Hunter, and J. and MacIntyre, "A Multiobjective Evolutionary Setting for Feature Selection and a Commonality-Based Crossover Operator" *Proc. of CEC00,* vol. 1(2), pp. 309-316, 2000.

[23] C. Emmanouilidis, "Evolutionary Multiobjective Feature Selection and ROC Analysis with Application to Industrial Machinery Fault Diagnosis" *Evolutionary Methods for Design Optimisation and Control,* 2002.

[24] N. Otsu, "A Threshold Selection Method from Gray-Level Histogram" *IEEE Trans. System Man Cybernetics,* pp. 62-66, 1979.

[25] E. Badekas and N. Papamarkos, "Automatic Evaluation of Document Binarization Results" in *X CIARP*, Havana, Cuba, 2005, pp. 1005-1014.

[26] G. Nagy and S. Seth, "Hierarchical representation of optically scanned documents" *Proc. 7th Int. Conference on Pattern Recognition,* pp. 347-349, 1984.

[27] C. Emmanouilidis and A. Hunter, "A Comparison of Crossover Operators in Neural Network Feature Selection with Multiobjective Evolutionary Algorithms" in *Proc. of the GECCO Workshop on Evolutionary Computation in the Development of ANNs*, Las Vegas, USA, 2000, pp. 58-60.

[28] Oulu-University, "MediaTeam Oulu Document Database http://www.mediateam.oulu.fi/downloads/MTDB/

[29] http://ida.first.fraunhofer.de/~anton/software.html.

[30] T. Joachims, "Making large-scale SVM learning practical," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds.: MIT Press, 1999.