2009 10th International Conference on Document Analysis and Recognition

Document binarization based on connected operators

Benoît Naegel and Laurent Wendling

Nancy University, LORIA UMR 7503
Nancy, France
{benoit.naegel,laurent.wendling}@loria.fr

Abstract

An original binarization method based on connected operators is proposed in this paper. Connected operators enable to filter and/or segment an image by preserving its contours. The proposed binarization method enables to extract relevant document objects by means of the component-tree structure. This method was compared to other binarization methods and showed good behavior in various contexts.

1. Introduction

Document recognition systems usually require a binarization step that aims at separating text from background. Most of the document image binarization methods process an image at the pixel level: the classification often relies on discriminating thresholds, whether global or local. Therefore, a common issue of these methods is that they are not able to disconnect text components from other close components.

In order to prevent this drawback, an original binarization method based on connected operators is presented in this paper. Connected operators have been introduced in mathematical morphology and allow to process an image at the flat-zone level. As a consequence, they preserve image contours.

More specifically, we propose a method based on the processing of the connected components of the image threshold sets: these sets can be efficiently managed in a structure called *component-tree*. Image binarization can therefore be viewed as a classification process, in which only the relevant object components are kept.

2. Related work

In the early years there has been a continued and substantial interest in the field of binarization techniques [33, 35]. In bi-level thresholding techniques it is assumed that an image contains two classes: the objects and the background, which can be distinguished by comparing the grey level values with a preset threshold value. Most current techniques perform the binarization either globally [36] or locally [8]. In global approaches a single threshold is computed and applied to the whole image whereas local methods use different thresholds according to the region under consideration. Some hybrid methods have also been proposed [15]. Most of these algorithms rely either on statistical methods (for example Bayes classifier, maximum likelihood [7, 18, 19] and moment preservation [37]), or on signal processing (for example maximization of the entropy of the image [1, 17], minimization of the variance between the object and the background [27] and the Hadamard transform [3]). Other approaches are based on edge detection algorithms [5, 14, 15, 28, 39], on fuzzy classification [6] or on multi scale [34]. Some local approaches are based on the decomposition of an image by means of a quad-tree structure [11, 10]. However such a decomposition is quite arbitrary, and may broke image connected components. Using irregular pyramids in combination with a contrast criterion for document binarization has been investigated in [21].

Connected operators [32, 30, 12] have been introduced in the field of mathematical morphology and allow to transform an image by preserving its contours (i.e. they do not introduce contours that are not present in the image). Attribute-filters [2, 16, 38] are connected operators allowing to remove connected components according to some criteria. They can be efficiently implemented by using a tree structure called *component-tree* [26] (or called differently in [29, 4, 22]) that stores in each node a connected component of an image threshold set. Component-trees have been involved in many image processing tasks, such as image simplification [29], object detection [16, 24], image retrieval [23], caption text detection [20] or identification of ancient drop caps [25].

To our knowledge, component-tree based methods have never been considered until now for binarization. A bina-

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.42

IEEE computer society

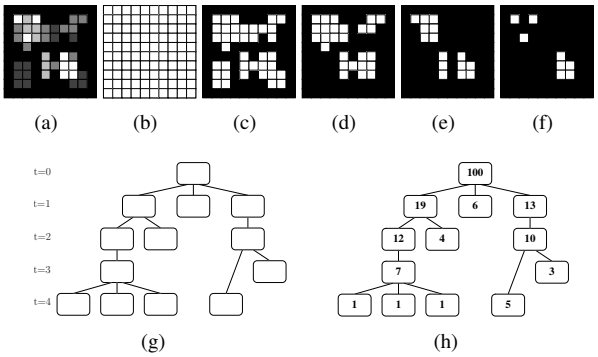

Figure 1. (a) A grey-level image F and its successive threshold sets $X_t(F)$ for t from 0 (b) to 4 (f). (g) The component-tree of F. (h) The same tree, enriched by an attribute (the size of the connected component of each node).

rization method based on this concept is presented in the sequel. As a consequence, this method belongs to the class of connected operators and therefore enable to extract relevant objects from a grey-level image while preserving its contours.

3. Background on component-tree

Let $F : E \rightarrow T$ be a discrete grey-level image with E the domain of definition of F ($E \subseteq \mathbb{Z}^2$) and $T = [t_{min}, \ldots, t_{max}] \subseteq \mathbb{Z}$ the set of values. Let $X \subseteq E$. We note $x \sim y$ the equivalence relation "x and y are connected in X". The connected components of X are the equivalence classes of X w.r.t. the equivalence relation $\sim$, *i.e.* the elements of the quotient set $X/\sim$ (noted $\mathcal{C}[X]$ in the sequel).

The threshold set of F at level t is defined by $X_t(F) = \{p \in E \mid t \leq F(p)\}$ for all $F \in T^E$. Let $t \in T$ and $X \subseteq E$. We define the cylinder function $C_{X,t}$ by $C_{X,t}(x) = t$ if $x \in X$ and t_{min} otherwise. Based on this definition, a discrete image $F \in T^E$ can be expressed as $F = \bigvee_{t \in T} C_{X_t(F),t} = \bigvee_{t \in T} \bigvee_{X \in \mathcal{C}[X_t(F)]} C_{X,t}$, where $\bigvee$ is the pointwise supremum of a set of functions, *i.e.* $F(x) = \sup_{t \in T}\{C_{X_t(F),t}(x)\}$.

Let $\mathcal{K} = \bigcup_{t \in T} \mathcal{C}[X_t(F)]$ be the set of connected components of all threshold sets. The relation $\subseteq$ is a partial order on $\mathcal{K}$. The transitive reflexive reduction of the relation $\subseteq$ on $\mathcal{K}$ induces a graph called the Hasse diagram of $(\mathcal{K}, \subseteq)$. This graph is a tree, the root of which is the supremum $R = \sup(\mathcal{K}, \subseteq) = E$. This rooted tree $(\mathcal{K}, L, R)$ is called the *component-tree of* F (see Fig. 1(g)). The elements $\mathcal{K}$, R and L are the set of the *nodes*, the *root* and the set of the *edges* of the tree, respectively.

Component-trees enable the storage of *attributes* at each node, related to the binary connected component associated to the node (for example the area of a component $X \in \mathcal{K}$: $|X|$ (see Figure 1(h))). Pruning a component-tree $(\mathcal{K}, L, R)$ of an image F according to criteria related to node attributes enables to perform filtering of F. The filtered image F_f is then defined as $F_f = \bigvee_{X \in \mathcal{K}_f} C_{X,m(X)}$ where $\mathcal{K}_f \subseteq \mathcal{K}$ is the subset of the remaining nodes after the pruning process and $m(X) = \min\{F(p) \mid p \in X\}$ is the *grey-value* associated to the component X. When performing segmentation, a binary result F_b can similarly be obtained as $F_b = \bigcup_{X \in \mathcal{K}_f} X$. Note that an operator that performs image filtering or segmentation based on component-tree pruning is a *connected operator*.

Let $\mathcal{M} \subseteq \mathcal{K}$ be the set of leafs defined by $\mathcal{M} = \{X \in \mathcal{K} \mid \forall Y \in \mathcal{K}, Y \not\subseteq X\}$. The components of $\mathcal{M}$ are called the *regional maxima* of F. The branch of the tree starting from the leaf $M \in \mathcal{M}$ is defined by the sequence of sets $\mathcal{B}_{\mathcal{K}}(M) = (X_k)_{k=1}^{n}$, such that $X_1 = M$, $X_n = R$ and $X_k \subseteq X_{k+1}, \forall X_k \in \mathcal{K}, k \in [1, \ldots, n-1]$.

4. Method

4.1. Overall description

The described method is based on the assumption that image is composed of white foreground and black background. The main strategy of our approach consists in selecting relevant connected components of the threshold sets of F, by means of its component-tree. The proposed approach is composed of three steps. Each step aims at removing irrelevant components according to a specific criterion: intensity, contrast, size.

The method is initialized by computing the component-tree $(\mathcal{K}, L, R)$ of F using Salembier's algorithm [29][1].

4.2. First step: rough binarization

In a first step the pixels are individually classified as object or background according to their grey-level. Indeed, we believe that the only way to discriminate an isolated grain of noise from a character part (for example the dot above the letter "i") is to consider its grey-value.

This classification step can be based on any two-class classifier: in this paper we have used a non-supervised K-Means classifier. The set $F_b \subseteq E$ of points classified as objects (considered as a binary image) is used as a mask in the sequel.

[1] According to [26] and as showed in our experiments, Salembier's algorithm is quadratic in the worst case; however it is generally twice as fast as Najman's one when $T = [0, \ldots, 255]$.

4.3. Second step: contrast maximisation

Let $X \in \mathcal{K}$ be a node of the component-tree. The neighborhood of X of size λ is defined as $N_\lambda(X) = \{p \in E \backslash X \mid \exists x \in X, d(x,p) \leq \lambda\}$. A measure of contrast of the node X derived from the Fisher discriminant is defined as $J_\lambda(X) = (m(X) - \mu_2)^2/(\sigma_1^2 + \sigma_2^2)$, where $m(X)$ is the *grey-value* of X, $\sigma_1^2 = \sum_{p \in X}(F(p) - \mu_1)^2/|X|$, $\sigma_2^2 = \sum_{p \in N_\lambda(X)}(F(p) - \mu_2)^2/|N_\lambda(X)|$, $\mu_1 = \sum_{p \in X} F(p)/|X|$, $\mu_2 = \sum_{p \in N_\lambda(X)} F(p)/|N_\lambda(X)|$.

The principle of this step is to keep, for each branch, the node for which this measure of contrast is maximum. More precisely, let $\mathcal{M}_b = \{X \in \mathcal{M} \mid X \cap F_b \neq \emptyset\}$ the set of leafs (regional extrema) intersecting the mask F_b. For each leaf $M \in \mathcal{M}_b$ is kept the node: $K_J(M) = \text{argmax}_{X \in \mathcal{B}_\mathcal{K}(M)}\{J_\lambda(X)\}$. The set $\mathcal{K}_b \subseteq \mathcal{K}$ of nodes remaining after this step is then defined by: $\mathcal{K}_b = \{K_J(M) \mid M \in \mathcal{M}_b\}$.

4.4. Third step: size criterion

A third step aims at extracting specific image objects, in order to remove nodes containing several objects that should appear disconnected in the final result.

To this aim, a criterion based on the bounding-box size of the component has been chosen: $B_{\alpha,\beta}(X) = \|(BB_x(X), BB_y(X)) - (\alpha, \beta)\|_2$, where $BB_x(X)$ (resp. $BB_y(X)$) denote the width (resp. the height) of the bounding box of X. For each branch of the *remaining leafs* $\mathcal{M}_c$ of $\mathcal{K}_b$ is retrieved the node *minimizing* this criterion. The set $\mathcal{K}_c \subseteq \mathcal{K}_b$ of nodes remaining after this step is then defined by: $\mathcal{K}_c = \{K_B(M) \mid M \in \mathcal{M}_c\}$, where $K_B(M) = \text{argmin}_{X \in \mathcal{B}_{\mathcal{K}_b}(M)}\{B_{\alpha,\beta}(X)\}$.

Finally, the final binarized image is defined by $F_f = \bigcup_{X \in \mathcal{K}_c} X$. The steps of the method are illustrated on Figure 2

5. Experiments

Data and evaluation Experiments have been performed on two kinds of grey-level documents: line drawings (see Figure 2) and images of ancient graphical drop caps (see Figure 3). Both types of documents are noisy. Considering line drawings, it is crucial to perform a binarization that allows to disconnect letters from the networks, in order to improve the document understanding. Binarization of drop caps should be accurate to ensure a fine pattern recognition step. Two criteria were used to assess the robustness of our operator in comparison with other binarization methods. First a basic OCR has been used to calculate the number of extracted characters into line drawings using each binarization method. Secondly a powerful statistical recognition method based on Generic Fourier Descriptors [40] has been applied on a database of binarized drop caps to show the robustness of the binarization methods in such a context.

Method	M1	M2	M3	M4
Extracted characters	**74**	37	53	52
Connected characters	**8**	57	38	3

Table 1. Percentages of extracted characters and characters still connected.

Let us call M1 our method based on the component-tree, M2 the binarization method of Trier and Taxt [36], M3 the binarization method based on fuzzy entropy of Cheng and Chen [6] and M4 the method of Sauvola and Pietikäinen [31].

Line drawings In order to test the ability of binarization to disconnect graphic characters from networks in line drawings we apply the well-known Fletcher and Kasturi [9] algorithm that allows to split documents in text and graphic layers (and others for unknown regions).

Around ten bad quality line drawings have been used to test the methods. Each document contains hundreds of characters and most of them are close to lines. Few make occlusions with the network. Table 1 gives the percentage of right extracted characters and still connected characters. Missing percentages are essentially due to noise and also for few missing extracted characters due to the implemented approach, as "I" and "l", which belong to the layer "other".

Table 1 shows the good behavior of our approach which supersedes others considering these line drawings tests. M2 provides clean binarization results. Nevertheless characters and graphic networks are rather thick due to the calculation of Laplacian on noise boundaries. That gives rise to numerous connected components which may be difficult to disconnect in further processing. That is also why the cumulative percentage is greater because basic characters as "I" and ''l" belong to the graphic layer. The results obtained using M3 are quite interesting but they are lower of those of M1 in all the experiments. Method M4 provides very noisy and hardly exploitable results. That is in concordance with the calculation which is sensitive to the size of the surrounding windows and the homogeneity criterion of the area. Such method is often coupled with a filtering as wiener filter in OCR based application [13].

Statistical recognition To assess the ability of the method to extract meaningful image parts in a classification context, a database of 100 ancient graphical drop caps from 10 classes (each class corresponding to a different letter) was binarized using the 4 methods (see Figure 3).

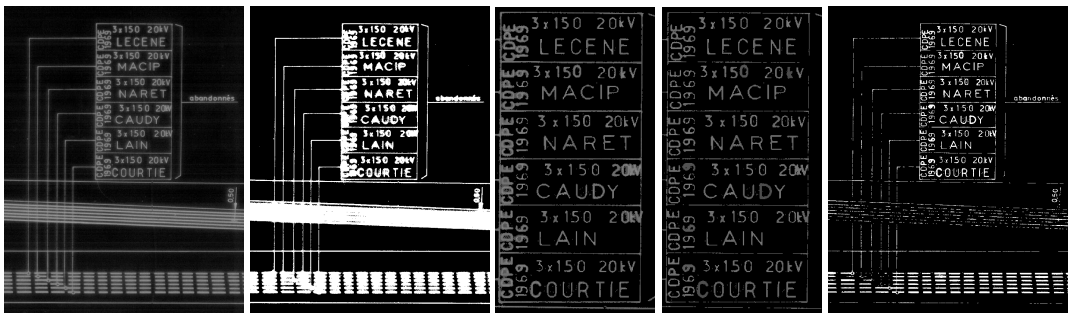

Figure 2. Illustration of the proposed binarization method on a line drawing. From left to right: Original image, first step, second step and third step (region of interest), final result.

Figure 3. From left to right: original drop cap, binarization using the proposed method (M1), method of Trier and Taxt (M2) and method based on fuzzy entropy (M3).

Method	M1	M2	M3	M4
Recognition rate	**30**	25	22	24

Table 2. Recognition rate (in percent) of ancient graphical drop caps.

For each set of binarized drop caps, the same head cluster was chosen for each class. A classification process based on the similarity between each binarized drop cap and each cluster head using the GFD shape descriptor [40] was done.

Table 2 presents the percentage of well classified drop caps. Considering this criterion, our approach performs better than the others. This result is explained by the fact that our method aims at extracting specific image objects. As a consequence irrelevant components (belonging to the texture part of the drop cap) are removed. It should however be noted that this recognition rate could be further increased by using a specific strategy for drop caps recognition (as in [25]).

6. Conclusion

In this paper an original binarization approach based on connected operators has been described. This method has been evaluated in various contexts and showed good performance in comparison with other binarization methods. While the approach is generic, it has been specifically designed to extract bright contrasted objects from dark background in the context of technical or graphical documents. As a consequence, the described approach is not suited to binarize, for example, natural images; however by defining criteria relevant for such images, the approach is still valid. One of the main interest of the approach lies in the processing of the connected components of the threshold sets as a whole, therefore preserving the contours of the original image. Moreover, this approach can be specialized for specific applications by defining further criteria related to the geometry, texture or shape of objects.

Future works will investigate the extension of this approach for specific tasks such as text and graphic separation.

References

[1] A. S. Abutaleb. Automatic Thresholding of Gray Level Pictures Using Two-Dimensional Entropy. *Computer Graphics and Image Processing*, 47:22–32, 1982.

[2] E. Breen and R. Jones. Attribute openings, thinnings, and granulometries. *Computer Vision and Image Understanding*, 64(3):377–389, 1996.

[3] F. Chang, K.-H. Liang, T.-M. Tan, and W.-L. Hwang. Binarization of document images using Hadamard multiresolution analysis. In *Proceedings of 5th International Conference on Document Analysis and Recognition (Bangalore, India)*, pages 157–160, Sept. 1999.

[4] L. Chen, M. Berry, and W. Hargrove. Using dendronal signatures for feature extraction and retrieval. *International Journal of Imaging Systems and Technology*, 11(4):243–253, 2000.

[5] Q. Chen, Q. Sun, P. Heng, and D. Xia. A double-threshold image binarization method based on edge detector. *Pattern Recognition*, 41:1254–1267, 2008.

[6] H.-D. Cheng and Y.-H. Chen. Fuzzy partition of two-dimensional histogram and its application to thresholding. *Pattern Recognition*, 32(5):825–843, 1999.

[7] S. Cho, R. Haralick, and S. Yi. Improvement of Kittler and Illingworth's Minimum Error Thresholding. *Pattern Recognition*, 22:609–617, 1989.

[8] L. Eikvil, T. Taxt, and K. Moen. A fast adaptative method for binarization of document images. In *Proceedings of 1st International Conference on Document Analysis and Recognition (Saint-Malo, France)*, volume 1, pages 435–443, 1991.

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

[10] E. Gabarra and S. Tabbone. Combining global and local threshold to binarize document images. In *Proceedings of 2nd Iberian Conference on Pattern Recognition and Image Analysis*, Estoril, Portugal, June 2005.

[11] T. Gadi and R. Benslimane. Fuzzy hierarchical segmentation. *Traitement du Signal*, 7(1):59–67, 2000.

[12] D. Gatica-Perez, C. Gu, M. Sun, and S. Ruiz-Correa. Extensive partition operators, gray-level connected operators, and region merging/classification segmentation algorithms: Theoretical links. *IEEE Trans. on Image Processing*, 10(9), 2001.

[13] B. Gatos, I. Praktikakis, and S. Perantonis. Adaptive degraded document image binarization. *Pattern Recognition*, 39:317–327, 2006.

[14] A. K. Jain and S. Bhattacharjee. Text Segmentation Using Gabor Filters for Automatic Document Processing. *Machine Vision and Applications*, 5(3):169–184, Summer 1992.

[15] J.-H. Jang and K.-S. Hong. Binarization of noisy gray-scale character images by thin line modeling. *Pattern Recognition*, 32(5):743–752, May 1999.

[16] R. Jones. Connected filtering and segmentation using component trees. *Computer Vision and Image Understanding*, 75(3):215–228, 1999.

[17] N. J. Kapur, P. K. Sahoo, and A. K. Wong. A New Method for Gray-Level Picture Thresholding Using the Entropy of the Histogram. *Computer Graphics and Image Processing*, 29:273–285, 1985.

[18] J. Kittler and J. Illingworth. Minimum Error Thresholding. *Pattern Recognition*, 19(1):41–47, 1986.

[19] T. Kurita, N. Otsu, and N. Abdelmalek. Maximum likelihood thresholding based on population mixture models. *Pattern Recognition*, 25(10):1231–1240, 1992.

[20] M. León, S. Mallo, and A. Gasull. A tree structured-based caption text detection approach. In *Proceedings of the Fifth IASTED Internation Conference*, pages 220–225, 2005.

[21] P. K. Loo and C. L. Tan. Using irregular pyramid for text segmentation and binarization of gray scale images. In *Proceedings of the 7th International Conference on Document Analysis and Recognition* , pages 594-598, Volume 1, 2003.

[22] J. Mattes and J. Demongeot. Efficient algorithms to implement the confinement tree. In G. Borgefors, I. Nyström, and G. S. di Baja, editors, *DGCI'00 - Discrete Geometry for Computer Imagery*, volume 1953 of *LNCS*, pages 392–405. Springer, 2000.

[23] V. Mosorov. A main stem concept for image matching. *Pattern Recognition Letters*, 26:1105–1117, 2005.

[24] B. Naegel, N. Passat, N. Boch, and M. Kocher. Segmentation using vector-attributes filters: methodology and application to dermatological imaging. In G. Bannon, J. Barrera, and U. Braga-Neto, editors, *ISMM 2007, 8th International Symposium on Mathematical Morphology. Rio de Janeiro, Brazil.*, volume 1, pages 239–250. INPE, 2007.

[25] B. Naegel and L. Wendling. Combining shape descriptors and component-tree for recognition of ancient graphical drop caps. In *Proc. of VISAPP09: 4th International Conference on Computer Vision Theory and Applications*, volume 2, pages 297–302. INSTICC, 2009.

[26] L. Najman and M. Couprie. Building the component tree in quasi-linear time. *IEEE Transactions on Image Processing*, 15(11):3531–3539, 2006.

[27] N. Otsu. A threshold selection method from grey-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-9(1):62–66, Jan. 1979.

[28] J. R. Parker, C. Jennings, and A. G. Salkauskas. Thresholding Using an Illumination Model. In *Proceedings of 2nd International Conference on Document Analysis and Recognition, Tsukuba (Japan)*, pages 270–273, 1993.

[29] P. Salembier, A. Oliveras, and L. Garrido. Anti-extensive connected operators for image and sequence processing. *IEEE Transactions on Image Processing*, 7(4):555–570, 1998.

[30] P. Salembier and J. Serra. Flat zones filtering, connected operators, and filters by reconstruction. *IEEE Transactions on Image Processing*, 3(8):1153–1160, Aug. 1995.

[31] J. Sauvola and M. Pietikäinen. Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236, 2000.

[32] J. Serra and P. Salembier. Connected operators and pyramids. In *Proc. of SPIE Image algebra and mathematical morphology*, volume 2030, pages 65–76, 1993.

[33] M. Sezgin and B. Sankur. Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging*, 13(1):146–168, 2004.

[34] S. Tabbone and L. Wendling. Multi-scale binarization of images. *Pattern Recognition Letters*, 24(1–3):403–411, 2003.

[35] Ø. Trier and T. Taxt. Evaluation of Binarization Methods for Document Images. *IEEE Transactions on PAMI*, 17(3):312–315, Mar. 1995.

[36] Ø. Trier and T. Taxt. Improvement of "integrated function algorithm" for binarization of document images. *Pattern Recognition Letters*, 16(3):277–283, Mar. 1995.

[37] W. Tsai. Moment-Preserving Thresholding: A New Approach. *Computer Vision, Graphics and Image Processing*, 29:377–393, 1985.

[38] E. Urbach, J. Roerdink and M. Wilkinson. Connected shape-size pattern spectra for rotation and scale-invariant classification of gray-scale images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(2):272–285, 2007.

[39] S. D. Yanowitz and A. M. Bruckstein. A New Method for Image Segmentation. *Computer Vision, Graphics and Image Processing*, 46(1):82–95, Apr. 1989.

[40] D. Zhang and G. Lu. Shape-based image retrieval using generic fourier descriptor. *Signal Processing: Image Communication*, 17:825–848, 2002.