A document binarization method based on connected operators

Benoît Naegel\textsuperscript{a,}\textsuperscript{*}, Laurent Wendling\textsuperscript{b}

\textsuperscript{a} Nancy University, LORIA UMR 7503, 54506 Vandœuvre-lès-Nancy, France
\textsuperscript{b} Paris Descartes University, LIPADE, 75270 Paris, France

\textbf{A R T I C L E   I N F O}

Article history:
Received 10 July 2009
Received in revised form 5 March 2010
Available online 10 April 2010
Communicated by C.L. Tan

Keywords:
Document binarization
Connected operators
Component-tree
Mathematical morphology

\textbf{A B S T R A C T}

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 strategy was compared to other binarization methods and showed good behavior in various contexts.

\textsuperscript{*} Corresponding author. Tel.: +33 0 3 29 53 60 41; fax: +33 03 29 53 60 00.
E-mail addresses: benoit.naegel@loria.fr (B. Naegel), lwendlin@math-info.univ-paris5.fr (L. Wendling).

\section{1. Introduction}

Connected operators have been introduced in the context of mathematical morphology \cite{Serra1993,Salembier1995,Gatica-Perez2001} and allow to process an image while preserving its contours. Connected operators can be defined from various ways \cite{Garrido1998,Meyer1998,Soille2003}. A common way to design anti-extensive connected operators is to consider the set of all the connected components of all the threshold sets of an image. These components can be organized into a tree called \textit{component-tree} \cite{Salembier1998,Najman2006}. The component-tree structure has been used in many contexts over the years. An advantage of this structure relies on its low computational construction cost. Powerful image processing methods based on arbitrary attributes \cite{Breen1996} can be designed in combination with this structure. Moreover, component-tree based attribute-filters can be computed in a real-time fashion, allowing to design interactive filtering methods \cite{Westenberg2007}.

In the field of document image processing, recognition systems usually require a binarization step that aims at separating both text and objects from background. Most of the document image binarization methods process an image at the pixel level: the individual pixel classification often relies on discriminating thresholds, whether global or local.

We propose in this paper an original document binarization technique based on the component-tree paradigm. This method is based on an optimization procedure applied to each branch of the tree. The method allows to binarize an image while avoiding some well-known drawbacks of binarization methods, such as connections between distinct foreground objects or removal of small objects. Moreover, the proposed method is almost free of parameters, since it is based on a local optimization strategy. The method has been applied on line drawing documents, ancient graphical drop caps and color maps. The proposed methodology is generic and can be transposed to other application fields. A short version of this work has been presented in \cite{Naegel2009b}.

The outline of the paper is as follows. Previous works related to connected operators and binarization methods are described in Section 2. In Section 3 some background on connected operators and component-tree are introduced. Our binarization method is presented in Section 4. Section 5 describes the experimental setup used to assess the performance of the proposed work, and the results of these experiments. Finally, in Section 6, conclusions based on these results are drawn.

\section{2. Related work}

Attribute-filters \cite{Breen1996,Jones1999,Urbach2007} are connected operators allowing to remove threshold components according to given criteria. They can be efficiently implemented by means of the \textit{component-tree} structure \cite{Salembier1998,Chen2000,Mattes2000,Najman2006} 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.
(Sal embier et al., 1998), object detection (Jones, 1999; Naegel et al., 2007), image retrieval (Mos rov, 2005), caption text detection (León et al., 2005) or identification of ancient drop caps (Naegel and Wendling, 2009a). Component-trees are also linked to the level lines or level sets of an image, which are contrast invariant structures carrying information strongly related to human visual perception (Monasse and Guichard, 2000; Desolneux et al., 2001).

The notion of contrast is commonly used as a relevant criterion in binarization methods. Contrast information has been introduced in the field of connected operators towards various ways. First, the notion of dynamics was introduced by (Grimaud, 1992) from which was derived the -operators (Soille, 2003). These connected operators enable to remove regional extrema having a contrast lower than a given parameter. The notion of volumic filtering was later introduced in (Vachier, 1998) in order to be closer from human visual perception. This notion can be seen as an extension of dynamics. In these works, however, the notion of contrast is always considered with respect to the extrema image components: we will introduce in Section 4 a notion of contrast that can be applied to any threshold components.

A large amount of work has been devoted to image binarization techniques over the years (Trier and Taxt, 1995a; Sezgin and Sankur, 2004). 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. Binarization methods are commonly divided in two categories: global and local approaches. 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 (Jang and Hong, 1999).

In a recent survey (Sezgin and Sankur, 2004), forty thresholding methods from various categories are compared. Most of these algorithms relies either on statistical methods (for example Bayes classifier, maximum likelihood (Mardia and Hainsworth, 1988; Cho et al., 1989; Kittler and Illingworth, 1986; Taxt et al., 1989; Kurita et al., 1993)), on fuzzy classification (Cheng and Yanowitz and Bruckstein, 1989), on extensive connected operators consists in analyzing separately the connected components of the image threshold sets (Breen and Jones, 1996; Sal embier et al., 1998). This can be done efficiently by means of the component-tree structure.

3.3. Component-tree

Let \( P(E) = \{X|X \subseteq E\} \). Let \( X_t \subseteq T^E \rightarrow P(E) \) be the threshold operator defined by \( X_t(F) = \{p \in E|F(p) > t\} \) for all \( F \in T^E \).

Let \( t \in T \) and \( X \subseteq E \). We define the cylinder function \( C_X(\cdot) \) by \( C_X(t) = t \) if \( x \in X \) and \( t_{\text{max}} \) otherwise. Based on this definition, a discrete image \( F \in T^E \) can be expressed as \( F = \bigcup_{\text{leafs}} C_X(\cdot) = \bigcup_{x \in X} C_X(x) \), where \( Y \) is the pointwise supremum of a set of functions, i.e. \( F(\cdot) = \sup_{x \in X} \{C_X(x)\} \). As \( T \) is finite.

Let \( K = \bigcup_{x \in X} C_X(x) \) be the set of connected-components of all threshold sets. The relation \( \subseteq \) is a partial order on \( K \). The transitive reflexive reduction of the relation \( \subseteq \) on \( K \) induces a graph called the Hasse diagram of \( K \). This graph is a tree, the root of which is the supremum \( R = \text{sup}(K, \subseteq) = E \). This rooted tree \( (K, R) \) is called the component-tree of \( F \) (see Fig. 1(g)). The elements \( K, R \) and \( I \) are the set of the nodes, the root and the set of the edges of the tree, respectively.

3.4. Tree branch

The concept of tree branch has been introduced in (Jones, 1999). Let \( M \subseteq K \) be the set of leaves defined by \( M = \{X \in K|\forall Y \subseteq X, Y \not\subseteq X\} \). The components of \( M \) are called the regional maxima of \( F \). The branch of the tree starting from the leaf \( M \subseteq M \) is defined by the unique sequence of nodes \( B_k(M) = \{X_k\}_{k=0}^{\infty} \), such that \( X_0 = M, X_k = R, \forall k \in [1, \ldots, n - 1], X_k \subseteq X_{k+1} \land \forall Y \subseteq K, X_k \not\subseteq Y \Rightarrow Y = X_k \).

3.5. Attributes

Component-trees allows storing attributes at each node, related to the binary connected component associated to the node (for example the area of a component \( X \subseteq K : |X| \) (see Fig. 1(h))). This allows to store for each connected component some attributes useful for image processing as bounding-box size, area, compactness (Breen and Jones, 1996). Some attributes can be computed very efficiently by using the inclusion relation between connected-components belonging to the same branch.

Pruning a component-tree \((K, R, L)\) of an image \( F \) according to criteria related to node attributes enables to perform filtering of \( F \). The filtered image \( F_l \) is then defined as \( F_l = \bigcup_{X \in L} C_X(\cdot) \) where \( L \subseteq K \) is the subset of the remaining filtered nodes after the pruning process and \( m(\cdot) = \min\{F(p)|p \in X\} \) is the grey-value associated to the component \( X \). When performing segmentation, a binary result \( F_b \) can similarly be obtained as \( F_b = \bigcup_{X \in X} C_X(\cdot) \).

Let \( X \subseteq K \) be a node of the component-tree. \( X \) may contain distinct flat-zones having the same value. Since an unique value is assigned for all the points of \( X \), an operator that performs either
image filtering or segmentation based on component-tree pruning is a connected operator.

4. Method

The main idea of our method relies on the concept of component-tree branch. Let $F$ be an image containing bright objects on dark background. Our method is based on the following assumption: each foreground object (characters, graphics, symbols) is represented by an unique node of the tree and two different foreground objects belong to two distinct branches of the component-tree. Obviously the reasoning is the same when considering dark objects with bright background, using the dual component-tree (i.e. the component-tree of the negative image).

Therefore, the principle is to retrieve for each branch of the tree an unique node maximizing a given criterion. Eventually, a same node can be retrieved for several distinct branches. The binarized result is obtained from the union of all retrieved nodes. This strategy can be seen as a local procedure aiming at finding the best possible threshold for each branch of the component-tree. It can also be viewed as a region-growing procedure, in which regional maxima are considered as seeds and stopping criterion is based on components attributes. Fig. 2 illustrates the principle of our approach.

4.1. Parameters of the method

The proposed approach is composed of three distinct steps. Each step aims at removing irrelevant components according to a specific criterion: intensity, contrast, size. The maximization procedure used to retrieve the target node in each branch is automatic.

However, in order to be adapted to a large range of applications, the method is based on a set of parameters:

1. $\lambda$: the size of the neighborhood used with the $J_1$ criterion;
2. $(a, b)$: the approximate bounding-box size of the target objects used with the $B_{a, b}$ criterion.

4.2. Contrast attribute

Contrast information is useful in many image processing tasks to assess perceptual significative parts and plays usually a key role in binarization methods. Since our method acts at the threshold component level, we need a measure of contrast enabling to assess the contrast of such a component. To this aim, we assess the local contrast of the component with respect to its neighborhood.

More formally, let $X \in \mathcal{X}$ be a node of the component-tree. The neighborhood of $X$ of size $\lambda$ is defined as $N(X) = \{p \in \mathcal{F} | d(x, p) \leq \lambda\}$, where $d(\cdot, \cdot)$ denotes the Euclidean distance. Let:

\[
\begin{align*}
\mu_1 &= \sum_{p \in X} F(p)/|X|, \\
\mu_2 &= \sum_{p \in N(X)} F(p)/|N(X)|, \\
\sigma_1^2 &= \sum_{p \in X} (F(p) - \mu_1)^2/|X|, \\
\sigma_2^2 &= \sum_{p \in N(X)} (F(p) - \mu_2)^2/|N(X)|,
\end{align*}
\]

respectively, be the grey-level mean of component $X$, the grey-level mean of the neighborhood of $X$, the grey-level variance of component $X$ and the grey-level variance of the neighborhood of $X$. 

---

Fig. 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).

Fig. 2. Illustration of the maximization process for one branch of the component-tree. (a) Grey-level image. (b) Maximization procedure starting from the upper left image regional maximum. (c–e) Maximization procedure.
Given an estimate of the inter-class variance \( (\mu_1 - \mu_2)^2 \) and of the intra-class variance \( \sigma_1^2 + \sigma_2^2 \), a separation between the class can be computed in order to maximize the inter-class variance while minimizing the intra-class variance (Fukunaga, 1972; Otsu, 1979; Duda et al., 2000). This criterion has been commonly used in popular thresholding methods (Otsu, 1979). A measure of contrast of the node \( X \) based on the Fisher’s discriminant can then be defined as:

\[
J_r(X) = (m(X) - \mu_r)^2 / (\sigma_1^2 + \sigma_2^2).
\]

Here, \( m(X) \) is used instead of \( \mu_r(X) \) because we want to assess the contrast of the component alone with respect to its neighborhood in order to keep or reject this component. In fact \( m(X) \) represents the mean of the grey-level of the (flat) component \( C_{X,m(X)} \).

### 4.3. First step: rough binarization

This step is based on the assumption that object pixels can be distinguished from background pixels based on their grey-level. This assumption is commonly used in binarization methods, however it is usually not sufficient to obtain a relevant segmentation of image. In order to remove most of the image background parts, pixels are classified individually according to their grey-levels. The threshold must be conservative enough to keep all the object components.

This step is necessary since the only way to discriminate an isolated, contrasted, grain of noise from a character part (for example the dot above the letter “t”) is to consider its grey-value. In our method this step is based on a k-Means clustering (with \( k = 2 \)). The set \( F_b \subseteq E \) of points classified as objects (considered as a binary image) is used as a mask in the sequel to localize potential image objects.

### 4.4. Second step: contrast maximization

In second step, contrasted nodes are kept following a maximization procedure. Given the set of regional maxima intersecting the mask \( F_b \), the principle is to keep, for each branch, the node for which the contrast measure (see Section 4.2) is maximum.

Let \( M_b = \{ X \in M \mid X \cap F_b \neq \emptyset \} \) the set of leaves (regional extrema) intersecting the mask \( F_b \). For each leaf \( M \in M_b \) is retrieved the node: \( \hat{K}_r(M) = \text{argmax}_{X \in M_b} J_r(X) \). The set \( F_r \subseteq K \) of nodes remaining after this step is then defined by: \( \hat{K}_r = \{ \hat{K}_r(M) \mid M \in M_b \} \).

This step is illustrated on Figs. 3 and 4 on two regional maxima. Starting from a regional maximum, the contrast measure \( J_r \) is computed for all the nodes belonging to the branch associated to this maximum. Fig. 4 shows the evolution of the contrast measure with respect to the level of the branch node starting from the branch leaf (right) towards the root (left). In the case of the regional maxima depicted in Fig. 4 the maximum value is automatically obtained at the threshold \( t = 115 \) in Fig. 4 (a) and the threshold \( t = 132 \) in Fig. 4 (b).

### 4.5. Third step: size criterion

In order to impose further constraints on the extracted objects, a third step may be optionally used. In a document application, the aim may be to keep and highlight characters. To this aim, a criterion based on the bounding-box size of the component can be used: \( B_{X,a} = ||(BB_{X}(X)) - (a,b)||_2 \), where \( BB_{X}(X) \) denote the width and the height of the bounding-box of \( X \) and \( (a,b) \) denote the width and the height of the desired component size. For each branch of the remaining leaves \( M_r \) of \( \hat{K}_r \) is retrieved the node minimizing this criterion. In this way, for each branch is kept the node whose bounding-box size match at best \( (a,b) \). The set \( K_a \subseteq K_r \) of nodes remaining after this step is then defined by: \( K_a = \{ K_a(M) \mid M \in M_r \} \), where \( K_a(M) = \text{argmin}_{X \subseteq C_{X,a}} B_{X,a}(X) \).

Following the described methodology, other attributes commonly used in character extraction methods could be used as compactness, density, elongation.

The final result is defined by \( F_f = \bigcup_{K \subseteq K_r} X \). The steps of the method are illustrated on Fig. 5.

### 5. Experiments

The method has been implemented in C++ and the computation of the component-tree \((K, L, R)\) of \( F \) was based on Salembier’s algorithm (Salembier et al., 1998).

According to (Najman and Couprie, 2006) 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 \( T = [0, \ldots, 255] \).
Experiments have been performed on noisy line drawing grey-level documents (see Fig. 5), ancient graphical drop caps and color maps. It is crucial to perform a binarization that allows to disconnect letters from the background, in order to improve the document understanding.

5.1. Parameters

Since the node selection procedure is based on an optimization procedure aiming at finding the most relevant node of each branch, this method is almost parameter free. The parameter $\lambda$ controls the

Fig. 4. Illustration of the maximization process for two regional maxima (in white on top left subfigures) included in the regions depicted on Fig. 3. The graphics show the evolution of the contrast measure (y-axis) versus the level of branch node (x-axis).

Fig. 5. Illustration of the proposed binarization method on a line drawing. (a–f) Original image. (b–g) Rough binarization. (c–h) Maximization procedure to retrieve the most contrasted node for each branch. (d–i) Constraint on the bounding-box size of the nodes. (e–j) Final result.
size of the neighborhood used in the computation of the contrast criterion. Therefore, this parameter is related to the sharpness of object edges. In all our experiments we have taken $k = 1$. This value ensures to keep fine image details. The parameters $(\alpha, \beta)$ control the size of image objects that will be kept. These parameters are obviously application dependant.

To assess the robustness of our operator in comparison with other binarization methods, a basic OCR was 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 (Zhang and Lu, 2002) has been applied on a database of binarized drop caps to show the robustness of the binarization methods in such a context.

Let us call M1 our method based on the component-tree, M2 the binarization method of Trier and Taxt (1995b), M3 the binarization method based on fuzzy entropy of Cheng and Chen (1999) and M4 the method of Sauvola and Pietikäinen (2000).

5.2. 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 (1988) algorithm that allows to split documents in text and graphic layers (and others for unknown regions). The Fig. 6 shows the application of such an approach on two binarized images (using M3).

Ten bad quality images representing line drawings have been used to evaluate the methods. Each document contains hundreds

<table>
<thead>
<tr>
<th>Method</th>
<th>M1</th>
<th>M2</th>
<th>M3</th>
<th>M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Extracted characters</td>
<td>74</td>
<td>37</td>
<td>53</td>
<td>52</td>
</tr>
<tr>
<td>Connected characters</td>
<td>8</td>
<td>57</td>
<td>38</td>
<td>3</td>
</tr>
</tbody>
</table>

Fig. 6. Text/Graphic decomposition based on Fletcher and Kasturi method (Fletcher and Kasturi, 1988). (a) Image. (b) Binarization. (c) Graphic layer. (d) Text layer.

Fig. 7. Comparison of binarisation methods on a line drawing document. From left to right column: M1 (our method), M2 (Trier and Taxt, 1995b), M3 (Cheng and Chen, 1999) and M4 (Sauvola and Pietikäinen, 2000).
of characters and most of them are close to lines. Few make occlusions with the network. A comparison of binarization methods is illustrated on Fig. 7.

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 “l” and “I”, 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 “P” 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. We have modified the defuzzification criterion to set the threshold to the kernel of form fuzzy measure to better disconnect it. Indeed, basic algorithm is not suitable because the threshold averages the noise between background and network and a degraded result is obtained. Despite such improvement, numerous components remain connected in concordance with the global processing of this method.

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 (Gatos et al., 2006).

5.3. 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 Fig. 8).

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 (Zhang and Lu, 2002) was done.

Table 2 presents the percentage of well classified drop caps. Considering this criterion, our approach performs better than the others. This result can be explained by the side effects of our method which extract specific image objects. As a consequence irrelevant components (belonging to the texture part of the drop cap) are removed because of their size. In many cases, the letter of the dropcap is clearly disconnected from other parts (see Fig. 8), leading to a more accurate recognition process. It should however be noted that the recognition rate could be further increased by using a specific strategy for drop caps recognition (as in Naegel and Wendling, 2009a).

5.4. Extension to color documents

Extending the component-tree concept to color images is not straightforward because color space is not totally ordered. Therefore, different orderings may be used to use component-trees with color images. A comparison between orderings that could be used to compute component-trees of multivalued images has been proposed in (Naegel and Passat, 2009).

In order to highlight structures of interest (i.e. colored characters) according to some predetermined knowledge, one possibility is to use reduced ordering, which consists in reducing each color triplet (red,green,blue) to a scalar value by using a function $r : T \rightarrow \mathbb{R}$, where T denote the set of RGB colors. We consider the distance to a reference value $c \in T$, i.e. $r(T) = d(c,T)$, where $d(\ldots)$ denote the Euclidean distance. By choosing $r$ close to the “color” of the structures of interest, a grey-level image where the structures of interest are highlighted (i.e. bright on dark background) is obtained.

Let $F : E \rightarrow T$ a color image defined in the RGB space. For each reference color is computed the corresponding reduced grey-level image $F_{c}$, where: $F_{c}(x) = r(F(x))$. Following the methodology described in Section 4, a binarization is obtained for each reference color $c$. The final binarization result is obtained by taking the union of all binarizations. The choice of reference colors can be determined automatically by using a clustering of the color space (Badekas et al., 2006).

Fig. 9 presents the results of our method obtained on color maps from the MediaTeam Oulu Document Database. These results show that our method allows to easily locate characters parts.

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 relies on 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


http://www.mediateam.oulu.fi/downloads/MTDB.