Combining Force Histogram and Discrete Lines to extract Dashed Lines

Isabelle Debled-Rennesson\textsuperscript{1} and Laurent Wendling\textsuperscript{2}

\begin{itemize}
\item \textsuperscript{1}LORIA Campus Scientifique, Université Henri Poincaré 54506 Vandœuvre-lès-Nancy France
\item \textsuperscript{2}LIPADE, Université Paris Descartes, 45, rue des Saints-Prés 75270 Paris Cedex 06
debled@loria.fr; wendling@parisdescartes.fr
\end{itemize}

Abstract

A new method to extract dashed lines in technical documents is proposed in this paper by combining force histogram and discrete lines. The aim is to study the spatial location of couples of connected components using force histogram and to refine the recognition by considering surrounding discrete lines. This new model is fast and it allows a good extraction of occulted patterns in presence of noise. Efficient common methods require several thresholds to process with technical documents. The proposed method requires only few thresholds which can be automatically set from data.

1. Introduction

It is well-known that a dashed line brings precious information for the understanding of the document (separating parts, associated text boxes, etc) [1, 2]. In many recognition systems, it is important to have an accurate and powerful operator related to the retrieval of such typical lines. Most of the algorithms rely on some characteristics common to all dashed lines. Three main hypotheses are generally taken into account: It exists a minimum number of dashes having approximately the same length, they are regularly spaced and they follow a virtual line. Extraction can be carried out either directly on the pixel image, using directional mathematical morphology operators [3], or on the vectors set by the raster-to-vector conversion [4]. A powerful approach has been proposed by Dori et al. [5] in 1995. This method offers a satisfying solution for this problem in most of the cases. The underlying mechanism is a sequential stepwise recovery of components that meet certain continuity conditions relating to common characteristics of dashed lines. The method starts by extracting keys, that is segments which are smaller than a given threshold and which have at least one free extremum. The main loop consists in choosing a key as the start of a new dashed line hypothesis, and in trying to extend this hypothesis in both directions, by adding other segments belonging to the same virtual line. This search is done in a search area whose width is the double of the current key width, and whose length is the maximal distance allowed between two segments belonging to a same dashed line. Dosch et al. [2] have proposed some improvements to the basic method by studying connection points and the merging dashed segments by propagating them following a distance threshold. Even if results are satisfying in many cases, methods depend on well-known raster-to-vector method drawbacks especially in presence of noise. Furthermore distortions imply the delicate location of patterns to be found. As a consequence, numerous thresholds are generally manually set while depending both on the scale of documents and on the structure of patterns to be handled. Finally it is not easy to assess the accuracy of extracted primitives from data without human parameter setting. Here the problem is tackled by calculating force histogram between pairs of labeled components to provide kernel pattern to the calculation of discrete lines in order to propagate the dashed lines to efficiently process occluded components. This new promising generic method is fast and relies on few thresholds which can be directly set from document analysis.

2. Force Histogram

In this section, the computation of a histogram of forces [6] is recalled. Let $\varphi_r$ be the map from $R$ into $R_+$, null on $R_-$, such that:

$$\forall d \in R^*_+, \varphi_r(d) = 1/d^r$$

with $r$ the kind of force. The method is further based on the handling of segments to decrease the computation time, that is the calculation of the attraction force $f_r$ of a segment with regard to another. Let $I$ and $J$ be two segments beared by the same line of angle $\theta$, $|I|$ and $|J|$
their length and $D^0_{i,j}$ the distance between $I$ and $J$.

$$f_r(|I|, D^0_{i,j}, |J|) = \int_{D^0_{i,j}+|J|}^{D^0_{i,j}+|I|} \int_{0}^{f_r(u-v)dvdu}$$

When two objects $A$ and $B$ are handled following a direction $\theta$, a pencil of parallel lines, of angle $\theta$, which entirely describes these objects, can be defined. Let us take one line, denoted $D^0_{\eta}$. The two sets of segments borne by this line correspond to: $A_\theta(\eta) = \cup \{I_i\}_{i=1,n}$ and $B_\theta(\eta) = \cup \{J_j\}_{j=1,m}$. The mutual attraction of these segments is done by:

$$F(\theta, A_\theta(\eta), B_\theta(\eta)) = \sum_{i \in 1..n} \sum_{j \in 1..m} f_r(|I_i|, D^0_{i,j}, |J_j|)$$

All the pencils of lines $D^0_{\eta}$ which entirely describe $A$ and $B$ are then considered. In the discrete case, the histogram corresponds to a set of angles and the calculation of $F^{AB}(\theta)$ is assumed to an evaluation of the forces exerted by an object with regard to another in the direction $\theta$. The calculation of $F^{AB}_{r}$ onto a set of angles $\theta_i (\theta_i \in [-\pi, +\pi])$ defines a signature, denoted $F^{AB}$, following a set of directions. By axiomatic definitions of the function $F$, the following properties are checked: translation as the information of the object is processed independently of its location in the frame of the plane, symmetry as the forces exerted are the same ones following two opposite directions, scale factor if the signatures are normalized (only the shape of the histogram is considered) and rotation (after shifts), because the approach is isotropic. These third properties are well adapted to the studied problem: invariance to translation is required as segments can be randomly located in the document, symmetry is useful to consider the shape of matched histogram and homothety is suitable when handling with segments of different size. In this context, rotation is not informative because segments are included in the same optimal bounding discrete line. Nonetheless such property may be interesting in further investigation dedicated to extract circular dashed structures.

It has been shown that force histogram supersedes histogram of angles and it is well suited to assess spatial relations between regions [7]. The whole histogram can be assumed to a set of features integrating distance and spatial location between pairs of components ($F^{AB}$). It can also provide a compact description while considering only a component alone ($F^{AA}$) [6]. Basic method has been adapted here to manage with $n$ connected components instead of between a couple of objects.

3. Discrete Lines

The arithmetical definition of discrete lines [8] is used in our method to recognize dashed lines:

**Definition 1** A discrete line $D(a, b, \mu, \omega)$, whose main vector is $(b, a)$, lower bound $\mu$ and thickness $\omega$ (with $a$, $b$, $\mu$ and $\omega$ being integer such that $gcd(a, b) = 1$) is the set of integer points $(x, y)$ verifying $\mu \leq ax - by < \mu + \omega$.

The following definitions [9] permit in the proposed method to obtain the discrete lines, possibly thick, containing the dashed lines. Let us consider a sequence of points $S_b$ and we consider $|a| \leq |b|$ to simplify the writing:

**Definition 2** The discrete line $D(a, b, \mu, \omega)$ is said bounding for $S_b$ if all points of $S_b$ belong to $D$.

**Definition 3** A bounding discrete line of $S_b$ is said optimal (see Fig. 1) if its vertical distance (i.e. $\frac{\omega}{\max(|a|, |b|)}$) is minimal, i.e. if its vertical distance is equal to the vertical distance of $conv(S_b)$, the convex hull of $S_b$.

A linear algorithm was proposed in [9] to incrementally obtain the characteristics of the optimal bounding discrete line of a sequence of points. On the one hand, it relies on the linear and incremental computation of the convex hull of the scanned segment and, on the other hand, it relies on the arithmetical and geometrical properties of discrete lines.

![Figure 1](https://via.placeholder.com/150)

**Figure 1.** $D(5, 8, -8, 11)$, optimal bounding line (vertical distance $= \frac{16}{8} = 1.25$) of gray point sequence.

This algorithm is used in our method to detect the optimal bounding lines of the selected parts of dashed lines obtained after the computation of a histogram of forces between components of an image.

4. Dashed Line Extraction

The proposed method is defined from common dashed line hypotheses previously stated[5]: first hypothesis is taken into account by considering minimal
sequence of three connected components and length between them is included in both calculation and comparison of force histograms, the regular spacing between components relies also on the analysis of force histogram which directly integrates spatial location and the calculation of surrounding discrete lines allows to model the last hypothesis. The overall process is summarized in Fig. 2.

**Figure 2. System description.**

Let us describe main steps of the process.

**Step 1.** A common labeling algorithm is applied to the whole document to extract the set of connected components, noted \( L \).

**Step 2.** Due to the specificity of technical data, it is not useful to focus on each couple of components but only on pairs of closer connected components to decrease the processing time. For each pair \((A, B) \in L^2\) with \( A \neq B \) a force histogram \( F^{AB} \) is calculated and for any component \( A \in L \), another one called \( F^{AA} \).

**Step 3.** The aim is to extract kernel patterns of cardinality 2 or 3, noted \( KP_2 \) or \( KP_3 \). For each element \( u \in L \), we consider its two closest (following minimal distance \( d \)) components \( v \) and \( w \in L \). If these components have both a similar shape and \( d(u, v) \approx d(u, w) \) we obtain \( F^{uv} \approx F^{uw} \) assuming that \( u \) is the middle segment. In this case \((u, v, w) \in KP_3\). Obviously a threshold is required to make decision but it can also be estimated by the distribution of force histogram distances between closest components as dashed lines are, by definition, rather regular. A set of \( KP_2 \) is also calculated on remaining set of components, assuming that \( KP_2 \cap KP_3 = \emptyset \) to process with dashed line occulted components. Local histogram calculated on close components are compared using a basic similarity ratio, that is \( F^{uv} \approx F^{uw} \) if they are similar. In this case \((u, v) \in KP_2\). Each element of \( KP_2 \) and \( KP_3 \) is assumed to be a key kernel.

**Step 4.** The optimal bounding discrete line of each key kernel is computed [9] and the characteristics of these discrete lines are recorded. It permits the construction of one mask by key kernel, \( k \), containing all the points of the optimal bounding discrete line of \( k \). A large dashed line can be constituted by several key kernels and if we consider all the associated discrete lines, this may not overlap exactly. So to be more robust to noise and to distortion, we propose to merge intersecting discrete lines to provide local looking up area around key kernels. The **local mask area** defined from the discrete lines \( D_i \) and \( D_j \) is set as follows.

\[
\Delta_{ij} = D_i \cup D_j \text{ if } |D_i \cap D_j| \geq \min(|D_i|, |D_j|)
\]

and so on for other discrete lines corresponding to possible dashed lines.

**Step 5.** Then we propagate each kernel pattern to the left and to the right by considering the corresponding local mask area as a mask allowing the detection of possible occulted components (included in elements of \( L \)). The propagation is made by an assessment of force histogram between the couple defined by extremity basic component of the propagated kernel pattern and new basic segment candidate which is compared with the previous couple. If the result is similar, the process is run again. Finally series of achieved basic segments greater to three are assumed to be dashed lines.

### 5. Experimental Study

This methodology has obviously no interest considering clean documents (or close to skeleton) where basic geometric features are suitable to characterize dashed lines. Our approach has been tested on real documents coming from different sources (and resolution) like architectural drawings, EDF networks, electronic drawing... Similar (or close) results can be achieved using classical methods but requiring a manual setting – which is not easy to evaluate the cost! – relative to the structure of the processed document. Such methods are interesting to study amount of data of same nature and issued of similar acquisition step. As said previously, this lost of robustness is delicate considering documents of different nature and scale. Furthermore underlying methods rely on the quality of the extracted skeleton and the unification step. Once again the aim of this paper is to propose a new concept able to easily handle with broad technical images of document. Two significative results are provided in figures 3 and 4. These examples show main problematic cases due to classical raster-to-vector method[10] giving rise to bad quality of junctions and artifacts (see surrounding circles).
It is easy to show that the complexity of force histogram is in $O(pn\sqrt{n})$ with $n$ the number of points of the image. But it is rather in $O(pn)$ considering the process of closer components in a labeled document. $p$ is set to 128 directions. Experimental studies [6] have shown that more processed directions have low influence to the global shape of the histogram. Comparison between histogram of forces requires obviously the use of a metric to decide how close the matched objects are. However a lower value gives only rise to more kernels which are removed following the ratio $\omega - \frac{1}{\max(|a|,|b|)}$, called width, of the obtained optimal bounding discrete lines. Moreover the discrete line width has been automatically set from the analysis of the distribution of distance values after applying a distance transform on pairs of components (width of around 4 has been found for the previous examples).

6. Conclusion

An original method to extract dashed lines has been proposed in this paper. Achieved results are very promising, this method is robust and relies on few thresholds which can be automatically set by considering the particular aspect of technical documents. Currently the extension of the method to more complex repetitive patterns of dashed lines is under consideration by introducing grammar to represent the kernel to propagate. Further investigations will be carried out to extract dashed circles from combining both curvature profile and force histogram.

References