FEATURES FOR THE RECONSTRUCTION OF SHREDDED NOTEBOOK PAPER

A. Ukovich and G. Ramponi

IPL - DEEI, University of Trieste, v. A. Valerio, 10, I-34100 Trieste, Italy

ABSTRACT

In this paper the selection of suitable content-based features for the reconstruction of shredded notebook paper is examined. An algorithm for the detection of squared paper is proposed, based on the Hough transform. Experimental results show a good performance of the algorithm.

1. INTRODUCTION

The reconstruction of shredded documents is a problem that may be encountered in the field of forensics and investigation science. The aim of the reconstruction is to uncover evidence of destroyed documents. If the number of documents to recover is large and the remnants of different documents are mixed, as is usually the case, the reconstruction by hand is a very difficult and time-consuming process. A computer-aided reconstruction using image processing algorithms can significantly improve the problem solution.

The problem can be approached as a particular case of the *jigsaw puzzle* problem [1], in which the pieces of the puzzle, when appropriately re-assembled, do not form one connected planar region (as in the classical definition), but a number of connected planar regions equal to the number of shredded sheets. The description of the piece appearance by means of numerical features is used for the automatic re-assembly of jigsaw puzzles. Shape features, commonly used for the automatic re-assembly of jigsaw puzzles, do not give much help in the case of shredded documents, where most remnants have approximately the same shape. Rather, features describing the piece content are needed.

Typical office documents may include printed paper, handwritten notebook paper, and graphics. Both general content-based features, suitable to any of these kinds of office documents, such as standard MPEG-7 descriptors, and features specifically designed for the different categories of documents need to be used. In a previous work we concentrated on printed documents [6]. In this paper we address the problem of finding suitable features for handwritten notebook paper. These features will be used in the general system for the reconstruction of shredded documents. An example of remnants of shredded notebook paper is shown in Figure 1.

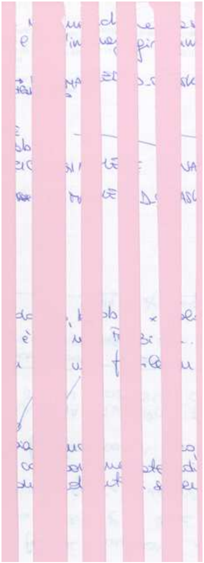

Fig. 1. Notebook remnants with handwritten text.

In Section 2 the distinctive features of notebook paper are examined, and a set of possible features to describe their remnants is selected. In Section 3 the squared pattern detection feature is presented. In Section 4 some experimental results are shown.

2. FEATURES FOR NOTEBOOK PAPER

In order to define appropriate features for the description of shredded remnants of notebook paper, some considerations have to be done on their general appearance and on their variability. In general, the features used for the description of remnants of printed text are not always suitable to the description of notebook remnants. For example, the paper used for printing a document is usually white, and different documents are printed on the same kind of paper in terms of size and of pulp type. On the contrary, notebook paper has usually the following distinctive features:

- the paper used can be of different kinds (and different thickness), resulting in different color and/or texture

This work has been supported by Andor s.r.l., Trieste, Italy.

- the handwritten text can have different color ink
- different writers produce different handwriting style
- squared paper or lined paper can be used.

Moreover, from our observations, the shredded remnants of notebook paper are more likely to have different sizes. This happens for two reasons: the first one is that common notebooks have various standard (e.g. $A4$ and $A5$) and non-standard sizes, the second one is that, in the case of notebook paper with low thickness, a jamming in the shredded machine is likely to occur, producing not regular or torn remnants.

On the base of these considerations on the distinctive features of remnants of notebook paper, the content-based features selected for the aim of reconstruction can be divided in the following categories: color features, features for the detection of squared/lined paper, features for the handwriting style description.

In the following, some suitable features for describing color and the handwriting style, respectively, are examined. For the detection of squared paper, features able to detect geometrical patterns should be used. An algorithm for the detection of squared paper, based on the Hough transform, is described in Section 3.

2.1. Color Features

Color features have the aim of describing the kind of paper and the color of ink of the handwritten text. For the description of the ink color, a segmentation needs to be performed to divide the textual part from the background one. The ink color is a distinctive feature of the pen or pencil used for writing.

There are a number of different color descriptors proposed in the literature [4] [5]. A suitable color descriptor for our problem does not need to include spatial information of the color distribution, because we just want to extract the overall paper color and the overall ink color (we suppose the same pen/pencil has been used on the same page). For this reason, descriptors such as color layout are not adequate. General color histograms can be used. However, the dimensionality of a color histogram feature is too high for a good classification. Either the MPEG-7 Scalable Color can be used, where the histogram information is reduced to a small number of coefficients, or descriptors of the dominant color. Indeed, since the histogram is representing either the background (paper) or the foreground (ink), it has approximately one dominant color. In order to capture the information on the paper pulp type, color correlograms [3] can be used on the segmented background, since they are able to describe the color-texture appearance.

2.2. Handwriting style description features

Concerning the handwriting style description, the studies done in the fields of handwriting recognition, writer identification and handwriting classification can help in choosing suitable features. However, we have to remember that a remnant usually does not contain entire words, in the frequent case in which the shredding direction is orthogonal to the text direction, but rather a few characters for each line of text. This is a very poor information for a good description of the writing style. Moreover, while in the case of remnants of printed paper the layout of the remnants in terms of the spatial arrangement of lines of text and paragraphs is expected to be quite the same in remnants which are adjacent in the original document, it is more difficult to have such a regular arrangement in notebook remnants written by hand, in particular in the case the paper is not squared. For this reason, descriptors able to detect general preferential directions of the characters of the handwritten text need to be selected. General texture descriptors [2], such as grey level co-occurrence matrix, Fourier descriptors, invariant moments, are not able to capture the directionality of the writing, since they focus on more general properties such as coarseness and contrast. Edge descriptors should be selected, such as the MPEG-7 Edge Histogram descriptor [4].

The Edge Histogram descriptor evaluates the spatial distribution of edges into four directions, vertical, horizontal, 45 deg. diagonal, 135 deg. diagonal, and one isotropic orientation (for non-directional edges). Local edge histograms are computed dividing the original image into 16 sub-images, independently of the original image size. Each sub-image is further divided into image blocks, and edges are computed using 2×2 partitions of each image block. The image blocks whose edge strength exceeds a threshold are used for computing the histogram. The frequency of occurrence of each of the five edge class is computed for each of the 16 sub-images. The values in the histogram are normalized to $[0, 1]$ and are nonlinearly quantized resulting in a 3 bits/bin representation.

In our case, since the images of the remnants have a very small width compared to the height, the initial division of the image into 16 sub-images organized in a 4×4 mesh is not suitable. A modification of the descriptor, with a division into 16×1 sub-images (16 vertically arranged blocks) is requested.

3. THE SQUARED PAPER DETECTION FEATURE

A first important distinction which helps the reconstruction is between squared and blank notebook paper. The detection of squares and lines in the remnants is not trivial, due to the low contrast of the printing. The feature which we resort to for this detection relies on the Hough Transform [2].

3.1. Hough transform

According to the transform, each pixel having coordinates (x, y) in the original image is transformed into a curve in the Hough space (θ, ρ). For the detection of straight lines, the parametric equation performing the transform is the following:

$$\rho = x\cos(\theta) + y\sin(\theta) \tag{1}$$

An accumulating matrix $H(\theta, \rho)$ is defined in the quantized Hough space, where $\theta = \{\theta_1, \ldots, \theta_M\}$ and $\rho = \{\rho_1, \ldots, \rho_N\}$. For each pixel in position (x_i, y_i) in the original image having gray value $v(x_i, y_i)$ and for each angle θ_m belonging to $\{\theta_1, \ldots, \theta_M\}$, ρ_m is calculated in this way:

$$\rho_m = x_i\cos(\theta_m) + y_i\sin(\theta_m), \tag{2}$$

and a contribution equal to $v(x_i, y_i)$ is added in the position (θ_m, ρ_m) of the accumulating matrix:

$$H(\theta_m, \rho_m) = H(\theta_m, \rho_m) + v(x_i, y_i). \tag{3}$$

A straight line in the original image corresponds to a set of sinusoidal curves in the Hough space (one for each pixel which the straight line is made of) intersecting in the same point (θ_m, ρ_m). In the Hough two-dimensional histogram this results in a high value $H(\theta_m, \rho_m)$, with the value θ_m indicating the orientation of the straight line and ρ_m indicating its distance from the image origin (x_0, y_0). The detection of straight lines in the original image becomes the detection of peaks in the Hough two-dimensional histogram. In Figure 3 an example is shown.

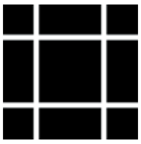

Fig. 2. Synthetic squared pattern

3.2. Square pattern

A pattern of squares, positioned with any orientation in the image, such as the one in Figure 2, produces in the Hough space some bright points with θ coordinate having just two possible values, θ_1 and θ_2, where $abs(\theta_1 - \theta_2) = 90$. This happens because a squared pattern is nothing else than a set of straight lines orthogonal to each other. The other pixels in the original image that do not belong to any straight line pattern produce in the Hough space some curves which do not intersect with each other or intersect in random positions. This means they are sparse and they create low and dispersed values in the Hough accumulating matrix. In our detection process a nonlinear operator $H_1 = F(H)$ is applied to the transformed data to separate such two different contributions. The details of the operator are not described here, due to a nondisclosure agreement.

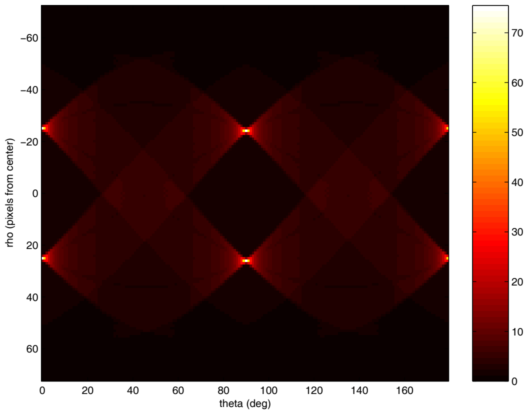

Fig. 3. Hough accumulating matrix $H(\theta_m, \rho_m)$ of the image in Figure 2. Peaks can be detected for $\theta = 0$ and $\theta = 90$

If the Hough Transform is performed from θ_1 to θ_M degree (for example $\theta_1 = 0$ and $\theta_M = 179$) with a step equal to 1 degree, then the square pattern can be detected looking for the maxima values in the array g:

$$g(\theta_i) = \sum_{\rho_j} H_1(\theta_i, \rho_j) + \sum_{\rho_j} H_1(\theta_{i+90}, \rho_j), \tag{4}$$

where θ_i belongs to $\{\theta_1, \ldots, \theta_{\frac{M}{2}}\}$

A square pattern or a line pattern in the original image exists if and only if there is a strong peak in $g(\theta_i)$. The square pattern can then be detected by means of an index, defined as:

$$SP = max(g) \tag{5}$$

3.3. Shredded notebook paper

In the case of the remnants of shredded documents, their typical shape, a strip as in Figure 1, could yield misleading results, if the Hough Transform is performed on the whole image at once. Indeed, since the image width is very low with respect to the image height, in the Hough space the contributions of θ in the horizontal direction are much less dispersed than the contributions of θ in the vertical direction, because in the horizontal direction just a few ρ values are possible. For this reason the Hough transform is evaluated on image sub-blocks of approximately square shape in the vertical remnant direction, and the contributions are then added.

4. EXPERIMENTAL RESULTS

Experiments have been conducted to evaluate the performance of the squared pattern detection feature described in Section 3. A data set of remnants has been created, including squared paper remnants and non squared paper remnants. The remnants have a large variability in terms of handwriting styles, percentage of handwritten text, orientation of the direction of writing, paper used, squared paper used. The description of the data set is shown in Table 1. Note that 6 different types of square paper are present in the dataset. Since these types include thick or thin printed squares, blue and black, and different square sizes, we believe this is a representative set of the possible squared patterns on notebook paper.

number of remnants	38
number of squared paper remnants	16
number of non squared paper remnants	22
number of different documents used	18
average number of remnants of the same document	2
number of different types of squared paper	6

Table 1. Dataset characteristics

The same parameters can be used for all the remnants, because the sub-blocks in which the remnant image is divided into have all approximately the same size. Moreover, the possible squared patterns of notebook paper have a size that does not vary much. Indeed, the square side size of notebook paper is usually 0.4 or 0.5 cm. This means that in the same sub-block approximately the same amount of straight lines can be encountered, thus the peaks in the $H(\theta_m, \rho_m)$ space have a value that does not depend on the size of the remnants or on the type of document.

The squared paper detection features values obtained for the data set remnants is shown in Figure 4. There are very low values, below $SP = 50$, as well as higher and more dispersed values. Two clusters are detected selecting the threshold $SP = 50$.

The classification result for the squared paper detection feature is shown in Table 2. All the remnants are correctly classified.

	squared paper	non squared paper
cluster 1	16	0
cluster 2	0	22

Table 2. Classification results for the feature SP

Some considerations need to be done on the distribution of the SP feature for the squared paper remnants. The values are high and quite disperse (as in Figure 4). This is due to the fact that different types of squared paper have been used for the experiments. The notebook paper for which the square pattern is weak results in low values of SP, while if the square pattern is very pronounced high values of SP are obtained. This information can be used for a further classification, since the square pattern remnants having similar SP values are very likely to belong to the same original document.

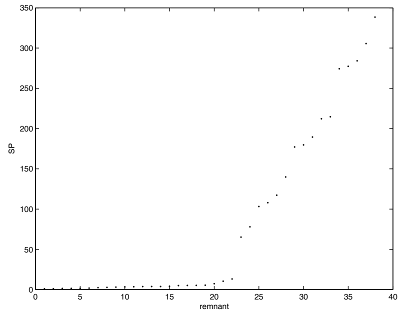

Fig. 4. Distribution of the feature squared pattern detection (SP) among the data set (ordered for increasing values).

5. CONCLUSIONS AND FUTURE WORK

In this paper the suitable features for the description of shredded notebook paper remnants for the purpose of reconstruction have been analyzed. An algorithm for the detection of squared paper based on the Hough transform has been proposed. Experiments using this algorithm on a set of remnants coming from different documents have shown a good performance. The color properties of the squared pattern feature have not been exploited in this preliminary study: the acquired RGB data have been converted to luminance values. A more effective classification of the different pages can be achieved taking also into account the hue of the printed pattern. Further classification experiments will be conducted using the color features, the handwriting style description features and the squared paper detection feature together.

6. REFERENCES

[1] H. Freeman and L. Garder, "Apictorial Jigsaw Puzzles: The Computer Solution of a Problem in Pattern Recognition," *IEEE Trans. on Electronic Computers*, vol.13, pp.118-127, 1964.

[2] R. C. Gonzalez and R. E. Woods, "Digital Image Processing," Addison-Wesley, Reading, MA, USA, 1992.

[3] Jing Huang, S.R. Kumar, M. Mitra, Wei-Jing Zhu, R. Zabih "Image indexing using color correlograms," *Proc. CVPR '97*, pp.762-768, 1997.

[4] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and Texture Descriptors," *Circuits and Systems for Video Technology, IEEE Trans. on*, vol.11, no.6, pp703-715, 2001.

[5] Y. Rui, T. Huang, and S. Chang, "Image Retrieval: Current Techniques, Promising Directions and Open Issues," *Journal of Visual Communication and Image Representation*, vol.10, no.4, pp.39-62, 1999.

[6] A. Ukovich and G. Ramponi, "System Architecture for The Digital Recovery of Shredded Documents," *in Proc. IS&T/SPIE Electronic Imaging 2005*, 2005.