Shredded document reconstruction using MPEG-7 standard descriptors

Anna Ukovich, Giovanni Ramponi, Haralambos Doulaverakis, Yiannis Kompatsiaris

***Abstract*— The recovery of paper documents which have been disposed of is an important issue in many contexts, such as forensics and investigation sciences. The automatization of the process by means of image processing techniques can give a considerable help in the problem solution. We propose in this paper an overall architecture for the reconstruction of strip-cut shredded documents, paying particular attention to the possibility of using MPEG-7 descriptors for the strip content description.**

I. Introduction

One of the tasks which forensics and investigation science have to deal with is the recovery of shredded documents. Documents may be torn by hand, but more often are destroyed using a suitable mechanical equipment, which cuts them into thin strips or even, with a cross-cut shredder, into small rectangles. The reconstruction of documents that have been shredded by a office strip-shredder is a difficult and time-consuming task to be performed by a human operator, and can become an insoluble problem when the number of shredded documents is large. The problem could be considered as a particular case of *jigsaw puzzle*. The automatization of the process by means of image processing techniques can give a considerable help in the problem solution.

Computer vision methods for the solution of jigsaw puzzles have been proposed since 1963 [1]. In the latest years, jigsaw puzzle approaches have been adopted in the field of archaeology and art conservation, for the computer-aided reconstruction of two- and three-dimensional fragmented objects [2] [3] [4].

In the literature, most of the solutions proposed for the assembly of jigsaw puzzles define a model for the piece contour and perform the matching based on the contour shape. In some contributions this problem is called ”apictorial jigsaw puzzle”. However, a human being attempting to assemble a puzzle does not consider just the information on the piece contour, but rather tries to find matching pieces on the base of their content, be it considered as color, shape or texture appearance. There are some works exploring the use of puzzle piece color information, together with contour shape information, to improve the automatic puzzle solver [5] [6].

The necessity of using the piece content information is made stronger by the consideration that in the case of shredded document reconstruction the puzzle pieces have almost all the same, rectangular, shape, thus the contour shape does not provide the necessary information for the matching. To define content features, the techniques developed in the last years for content-based image retrieval (CBIR) systems [7] [8] [9] represent a good starting point. In order to find a solution for the shredded document reconstruction problem in the context of CBIR systems, the concept of match should be replaced by the concept of similarity. We will assume that strips that are found to be similar by a CBIR system will have a high probability to be part of the same region (document).

We propose in this paper an overall architecture for the reconstruction of shredded documents, paying particular attention to the possibility of using MPEG-7 descriptors for the strip content description. We hypothesize that the documents have been cut by a strip-cut shredder. To our best knowledge, this is the first attempt in the literature to solve the shredded document reconstruction problem. The paper is organized as follows. In Section II the shredded document reconstruction problem is defined. In Section III the general system architecture is presented. In Section IV the MPEG-7 descriptors are described and in Section V the experimental results are reported and discussed.

II. Problem definition

With reference to publications in which the jigsaw puzzle problem has been either defined [1] [5] or redefined for a particular application [2], a set of ”puzzle rules” can be determined for the case of shredded documents. As in [3] we will distinguish between an ideal case of shredded remnants and the real, observed, case. In the ideal case we define the following rules:

- a piece (strip) is a one- or double-sided connected planar region
- the strip contour can be segmented into four sides separated by four corners
- the four sides of the contour represent a rectangle, that has the same dimensions for all the strips
- the set of pieces (strips), when properly assembled, fit together forming not one, but a number of regions (documents)
- two pieces that mate share a common border segment
- two corners of two adjacent matching pieces coincide when assembled
- there are no gaps between correctly matching pieces
- a piece matches up to two other pieces
- the match occurs only along the two longer sides of the contour

A. Ukovich and G.Ramponi are with the Dipartimento di Elettrotecnica, Elettronica e Informatica, University of Trieste, via A. Valerio, 10, I-34127 Trieste Italy, tel. +39 040 5587140, fax +39 040 5583460 (e-mail: {aukovich,ramponi}@units.it)

H. Doulaverakis and Y. Kompatsiaris are with the Informatics and Telematics Institute , 1st km of Panorama - Thermi Road, P.O. Box 361, GR-57001 Thessaloniki, Greece, tel. +30 2310 464160, fax +30 2310 464164 (e-mail: {ikom,doulaver}@iti.gr)

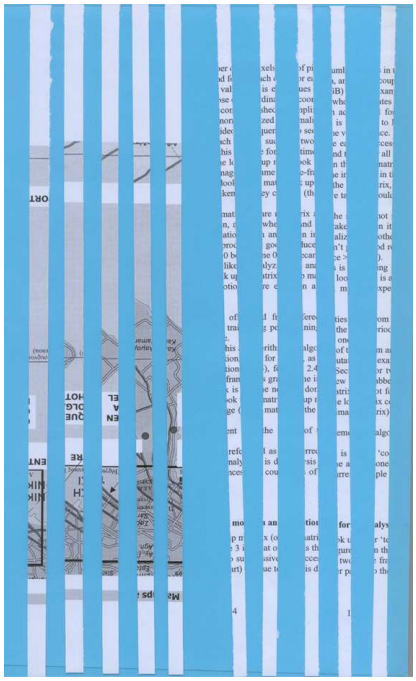

Fig. 1. Sample remnants of two different documents after the acquisition with a scanner

- the solution to the problem is unique
- there are frame pieces, i.e. pieces belonging to the border of the document (two for each reconstructed document)
- when double-sided pieces are considered, if two pieces match in one side they match in the other one too

These hypotheses are not all verified in the real, observed, case. The shape and contour appearance of real shredded document strips has been accurately described in the work by Brassil [10]. From this work and from our observations we can state that:

- the remnants (the strips) are not exactly rectangular
- the remnants do not have all the same shape
- two of the individual piece sides (the shorter) are flat, as well as the other two (the longer) are approximately flat or present a slight curvature
- the contour is slightly torn when shredded
- there can be small gaps between correctly matching pieces, due to the fact that the contour is slightly torn
- after shredding, the strips are further manipulated for disposal, and can be torn or folded
- some strips have an irregular shape, due to the fact that a jamming may occur and the shredding is performed in two or more steps

In Figure 1 an example of remnants of shredded documents is shown.

III. System overview

The system architecture we propose for the shredded documents reconstruction consists of different parts, that are described here.

First of all, the strips are acquired by a scanner. The background is chosen in an appropriate way, in order to have a high contrast with the strips. Thus, the process of strip-background segmentation is easy. More than one strip is acquired at once, taking care that the strips be separated one from the other, as in Fig. 1.

A preprocessing of the images acquired by the scanner is necessary to eliminate some noise of the acquisition process (that in particular occurs on the strip border, due also to the fact that strips are slightly torn when shredded), to segment the strips from the background and to divide the strips in different files (one file for each strip). After this step we have a database of images, each containing a single strip.

Since an exhaustive search is computationally very expensive, in particular if a large number of documents has to be reconstructed, a first clustering of the strips is necessary before the actual strip matching process starts. If we suppose to have N strips in the database after the two previous steps, then a first clustering has the aim of grouping together similar strips into subsets S_i, with $i = 1, \ldots, n$. The number of strips in the subset will be larger than the number of strips resulting from a shredded document, in a such a way that in the same subset strips belonging to similar documents will be grouped. For example, a subset will be made of strips containing color text, another one of strips containing handwritten text.

The matching process is done within the subset S_i obtained in the previous step. At this point an exhaustive search can be conducted in order to find the contiguous strips belonging to the same document. Once the strips belonging to the same document have been grouped together and ordered, algorithms are used to virtually reconstruct the original document. To evaluate the affine transformation of the strips, cross-correlation algorithms can be used, similarly to [11].

For the remnant clustering and matching, content-based image retrieval techniques are used. The grouping and the matching is done evaluating the similarity among the strips on the base of some general features, commonly used in content-based image retrieval systems, as well as some domain-specific features for the shredded documents.

The general features include color, texture and shape features. Color features allow to distinguish among color shredded documents. Texture features are expected to be useful in the case of text documents, since the text distribution in the strip could be regarded as a texture. Since in the real strips a slight curvature is often observed due to the shredding process, as explained in Section II, information on the shape of the strips could be useful in the strip matching process. As general features, we have considered the MPEG-7 descriptors, as described in the next Section.

Domain specific features include strip border features, OCR features and language dependent grammatical rules. Indeed, color, shape, texture features on the border region need to be considered for a correct matching of the document remnants. This is what the two operators proposed in [6], [5] do, comparing color features of regions close to the border and at the same position along the strip. OCR (Optical Character Recognition) features and language-dependent grammatical rules are useful since shredded strips usually come from office documents, which often contain text. If the used font is small enough and the document is cut in the direction orthogonal to

the text direction, we may be able to identify portions of lines of text, each line made of a sequence of letters.

IV. MPEG-7 DESCRIPTORS CONSIDERED

The MPEG-7 descriptors used in the experiments are three color descriptors [13], Scalable Color, Color Structure, Color Layout, two texture descriptors [13], Edge Histogram and Homogeneous Texture, and two shape descriptors [14], Contour Shape and Region Shape. They are shortly described below.

Scalable Color It consists of the image histogram in the hue-saturation-value (HSV) color space, quantized with a nonlinear quantization and encoded with a Haar Transform.

Color Structure If we use $c_0, c_1, c_2, \ldots, c_{M-1}$ HMMD color space quantized colors, the color structure histogram is $h(m), m = 0, 1, \ldots, M-1$, with each bin representing the number of $8x8$-structuring elements in the image containing at least one pixel with color c_m. The structuring element spatial extent is determined by:

$$\begin{aligned} p &= max\{0, round(0.5log_2WH - 8)\} \\ K &= 2^p, E = 8K, \end{aligned} \tag{1}$$

where W, H are the image width and height, respectively, $E \times E$ is the spatial extent of the structuring element, K is the subsampling factor.

Color Layout It extracts the average color, in the YCrCb color space, of 64 (8×8) image blocks, and it encodes this color using a DCT.

Edge Histogram It describes the spatial distribution of the edges in the image. Five edges categories are considered: vertical, horizontal, 45 deg., 135 deg., isotropic (nonorientation specific). Three levels of localization (scale) are considered.

Homogeneous Texture It provides a quantitative description of homogeneous texture regions in the image. It is obtained by filtering the image with a bank of orientation- and scale-sensitive filters, and computing the mean and standard deviation of the filtered outputs in the frequency domain.

Contour Shape It describes the contour of a region and it is based on the curvature scale-space (CSS) description of the contour shape.

Region Shape It consists of the Angular Radial Transformation (ART) coefficients F_{nm}; if f is an image intensity function in polar coordinates and V_{nm} is the ART basis function of order n and m:

$$\begin{aligned} F_{nm} &= < V_{nm}(\rho, \theta), f(\rho, \theta) > \\ &= \int_0^{2\pi} \int_0^1 V^*_{nm}(\rho, \theta), f(\rho, \theta) \rho \, d\rho \, d\theta. \end{aligned} \tag{2}$$

V. EXPERIMENT RESULTS

In this Section we will report the experiments done using the MPEG-7 descriptors and the MPEG-7 XM (eXperimentation Model) software [12]. This experiments have the aim of evaluating the performance of standard features, commonly used in content-based image retrieval systems, when applied to the particular case of an image database of remnants of shredded documents.

The data set used for the experiments include 9 typical office documents. The name of the document and its characteristics are shown in Table I. For each document 5 shredded remnants have been taken, with the exception of *text3color*, for which 8 strips have been considered. The total number of images in the database is thus 48. By now the database is small, but we are planning to run further experiments on a larger number of images.

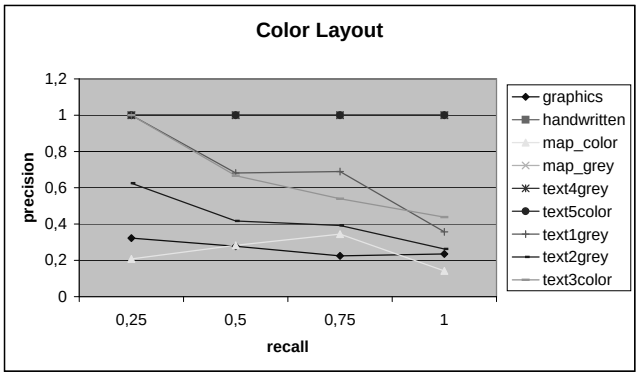

Fig. 2. Precision-recall curves for the Color Layout descriptor.

NAME	CONTENT	TYPE	COLOR
text1grey	report	text	B&W
text2grey	paper	text	B&W
text4grey	report	text	B&W
text3color	manual	text	color
map_grey	city map	graphics	grey
map_color	city map	graphics	color
graphics_grey	block diagram	graphics	grey
text5color	leaflet	text	color
handwritten	block notes	text	blue ink

TABLE I
DATASET CHARACTERISTICS

The discrimination power of each feature has been evaluated separately, using the MPEG-7 XM software. The performance of each feature has been analyzed using the precision-recall curves. Since we expect some features to work well with some kind of documents, and other with other kinds of documents, we have obtained one precision-recall curve for each document, instead of evaluating the overall performance of the feature in the whole database. The results for two descriptors, Color Layout and Scalable Color, are shown in Figures 2 and 3 as an example.

The results of the experiments were in general satisfactory for the three color descriptors, when considering the retrieval on color documents. It is interesting to observe the difference in performance between the Scalable Color descriptor and the Color Layout descriptor. The first one is a color histogram, and gives information about the overall color appearance of the strip. The second localizes the color information spatially in the image. The document *map_color* is a city map in color. The descriptor Scalable Color, as it is shown in Fig. 3, has a precision value constant and equal to 1 for this document, as well as for the Color Layout in Fig. 2 the precision is very low. This is due to the fact that the map has in each strip almost the same colors (thus the histograms used by the Scalable Color descriptor are the same), but the spatial distribution of these colors varies (because the map is detailed) and the

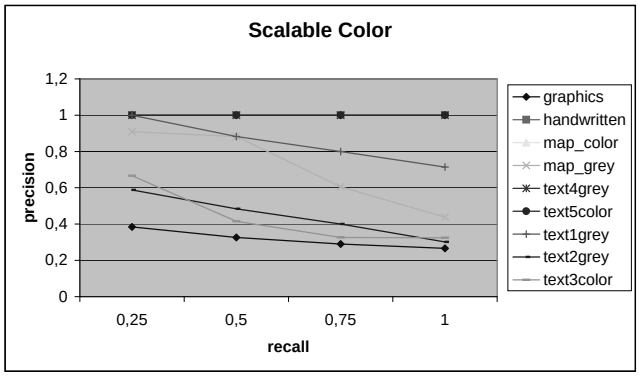

Fig. 3. Precision-recall curves for the Scalable Color descriptor.

Color Layout descriptor does not have a good performance. Conversely, for the document *text3color*, the Color Layout descriptor gives a curve in the precision-recall diagram that is higher than the one in the Scalable Color descriptor. The *text3color* document is a text document consisting of three colors, black, blue and red. In the document, red and blue are used for the titles, while black is used for the text inside the paragraphs. Since the strips are obtained cutting the document in the direction orthogonal to the text lines, it results that the blue and red parts are positioned in the same spatial positions in the shredded strips, thus a color descriptor with spatial color information, as Color Layout, performs better.

The two texture descriptors gave in general low precision-recall curves, indicating that they are not able to capture the texture appearance of the documents used in the experiments. In particular for the Edge Histogram descriptor, the reason could be the particular size of the document remnant images. Indeed, the images are very long (around 4000 pixel in height for a 400 dpi acquisition) and very narrow (the width is in general less than 200 pixels). Since the Edge Histogram operator considers three different scales dividing progressively the image into nxn sub-image blocks, it works well with images with similar width and height values, but it could have problems with the very long and narrow images of the strips.

The two shape descriptors describe the shape and contour appearance of the region that we selected to be the entire strip. Retrieval results were not particularly satisfactory, with the exception of those documents, *text4grey* and *text5color*, for which the curvature described in section II is more evident.

VI. Conclusions and future work

In this preliminary paper we have analyzed the possible use of content-based image retrieval techniques for the shredded document reconstruction task. We have characterized the shredded remnants as pieces of a particular jigsaw puzzle and we have described the general system architecture. Experiments obtained by using the standard MPEG-7 descriptors demonstrate that the features commonly used in general purpose content-based image retrieval systems can be used for this task as well, in particular for the color descriptors. The MPEG-7 texture descriptors used did not give the expected results, indicating the need of finding other texture descriptors free from problems of image dimension. The shape descriptors are useful only in the case of a strong curvature in the cut document remnants. Some domain specific features, such as OCR and features describing the content of the strip in the region close to the two horizontal borders, need also to be explored. Further results will be presented in the final version of the present submission.

VII. Acknowledgements

This research has been partially supported by the COST 276 project, within a Short Term Scientific Mission of one of the authors at the Institute of Telematics and Informatics of Thessaloniki, Greece, where the MPEG-7 experiments have been conducted. This research is conducted in the framework of the SCHEMA NoE (IST-2001-32795) [15].

References

[1] H. Freeman and L. Garder, "Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition," *IEEE Transactions on Electronic Computers*, vol. 13, pp. 118–127, April 1964.

[2] C. Papaodysseus, T. Panagopoulos, M. Exarhos, C. Triantafillou, D. Fragoulis, and C. Doumas, "Contour-shape based reconstruction of fragmented, 1600 b.c. wall paintings," *IEEE Transactions on Signal Processing*, vol. 50, no. 6, pp. 1277–1288, 2002.

[3] H. da Gama Leitao and J. Stolfi, "A multiscale method for the reassembly of two-dimensional fragmented objects," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 1239–1251, September 2002.

[4] E.-A. K. G. Papaioannou and T. Theoharis, "Virtual archaeologist: Assembling the past," *IEEE Computer Graphics and Applications*, vol. 21, no. 2, pp. 53–59, 2001.

[5] D. Kosiba, P. Devaux, S. Balasubramanian, T. Gandhi, and R. Kasturi, "An automatic jigsaw puzzle solver," in *Pattern Recognition, 1994. Vol. 1 - Conference A: Computer Vision and Image Processing., Proceedings of the 12th IAPR International Conference on*, vol. 1, pp. 616–618, 1994.

[6] M. Chung, M. Fleck, and D. Forsyth, "Jigsaw puzzle solver using shape and color," in *Proc. ICSP 98*, p. 877880, 1998.

[7] 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.

[8] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, 2000.

[9] T. Sikora, "The mpeg-7 visual standard for content description-an overview," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 696–702, 2001.

[10] J. Brassil, "Tracing the source of a shredded document," tech. rep., HP Labs 2002 Technical Reports, 2002.

[11] F. Stanco, L. Tenze, G. Ramponi, and A. D. Polo, "Virtual restoration of fragmented glass plate photographs," in *In proceedings of IEEE-Melecon 2004*, pp. 243–246, 2004.

[12] *MPEG-7 XM software*. `http://www.lis.ei.tum.de/research/bv/topics/mmdb/e_mpeg7.html`.

[13] B. S. Manjunath, J.-R. Ohm, V. V. Vasudevan, and A. Yamada, "Color and texture descriptors," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 703–715, 2001.

[14] M. Bober, "Mpeg-7 visual shape descriptors," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 11, no. 6, pp. 716–719, 2001.

[15] *SCHEMA NoE*. `http://www.schema-ist.org/SCHEMA/`.