Document Reconstruction Based on Feature Matching

Carlos Solana, Edson Justino, Luiz S. Oliveira, and Flavio Bortolozzi
Pontificia Universidade Catolica do Parana
Programa de Pos-Graduacao em Informatica Aplicada
Curitiba, Parana, BRAZIL
{csolana,justino,soares,fborto}@ppgia.pucpr.br

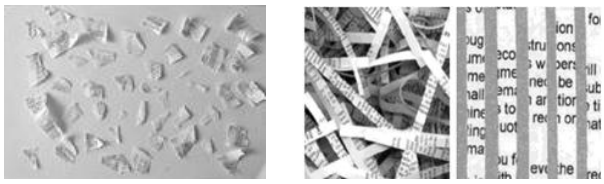
Abstract

We describe a procedure for reconstructing documents that have been shredded by hand, a problem that often arises in forensics. The proposed method first applies a polygonal approximation in order to reduce the complexity of the boundaries and then extracts relevant features of the polygon to carry out the local reconstruction. In this way the overall complexity can be dramatically reduced because few features are used to perform the matching. The ambiguities resulting from the local reconstruction are resolved and the pieces are merged together as we search for a global solution. We demonstrated through comprehensive experiments that this feature-matching-based procedure produces interesting results for the problem of document reconstruction.

1 Introduction

Questioned Document Examination (QDE) is a subfield of forensic sciences and it is related to the Federal, Civil, Law Enforcement, and Justice areas. The task of document examination is to compare a questioned document, using a scientific method to a series of known standards, for example, signature verification, handwriting identification, etc. In order to perform a reliable analysis, forensic document examiner must count on well preserved documents.

However, very often questioned documents suffer damages at several levels, such as, torn edges, moisture, obliteration, charring, and shredding. In the latter case, shredding can be performed by a machine or by hand (Figure 1). In both cases, documents need to be reconstructed so that forensic examiners can analyze them. The amount of time necessary to reconstruct a document depends on the size and the number of fragments, and it can be measured in days or even weeks. Sometimes some fragments of the document can be missing, and for this reason the document can be only partially reconstructed. Even then, the manual effort of the forensic examiner, which is tedious and laborious, can be alleviated.



(a) (b)

Figure 1. Different kinds of shredding.

One problem faced when reconstructing documents by hand lies in its the manipulation. The physical reconstruction of a document modifies some aspects of the original document because products like glue and adhesive tape are added into it. This type of manipulation if known as destructive analysis.

In this paper we focus on the reconstruction of documents shredded by hand, which is similar to the *automatic assembly of jigsaw puzzle*. Puzzle pieces are often represented by their boundary curves and local shape matching is usually achieved by curve matching. However, matching between two pieces usually occurs over only a fraction of their correspondent boundaries, and for this reason a partial curve matching is necessary.

Wolfson in [11] describes two curve matching algorithms where the boundaries are represented by shape feature strings which are obtained by polygonal approximation. The matching stage finds the longest common substring and it is solved by geometric hashing. The algorithms described by Wolfson are pretty fast, and for this reason are used by most puzzle solving methods. According to Kong and Kimia in [5], these algorithms fail when the number of puzzle pieces become larger, though. Other methods for curve matching have been proposed in the literature [9, 3] to perform matching at fine scale, however, their expensive

computational cost compromises their application for puzzle solving.

Kong and Kimia [5] propose re-sampling the boundaries by using a polygonal approximation in order to reduce the complexity of the curve matching. They make a coarse alignment using dynamic programming on the reduced version of the boundaries. Thereafter, they apply dynamic programming again into the original boundaries to get a fine-scale alignment. A similar approach is applied by H. Leitão and J. Stolfi [8], where they compare curvature of the fragments, at progressively increasing the scales of resolution, using an incremental dynamic programming sequence matching algorithm. Dynamic programming for puzzle solving has been used also by Bunke and Kaufmann [2] and Bunke and Buehler [1].

Many times local shape analysis produces ambiguous matches and it gets worse as the number of puzzle pieces increases. In order to eliminate the ambiguity in the global picture a global search technique is required. Wolfson et al in [12] report an algorithm to solve large puzzles, but with some constraints regarding the shape of the puzzle pieces. They show good results, but this kind of strategy is not practical in many real applications.

Uluk and Toroslu [10] search all pairs of pieces and the best match is selected and merged to form a new piece. The algorithm is repeated until there is only one piece left. Since invalid matching may occur during the merging process, a backtracking procedure is considered. If the backtracking is used very often, then the idea becomes computationally expensive. Alternatively, computers may be used only in the local shape analysis stage and an human can be used to assist the global search [7, 8]. This is known as "human in the loop" evaluation, a concept that has been largely applied to critical systems [6].

In this work we propose a local reconstruction based on two steps. First of all, we apply a polygonal approximation in order to reduce the complexity of the boundaries and overcome specific problems faced in document reconstruction. Most of the works found in the literature exploit the fact that ordinary puzzle pieces have smooth edges and well defined corners. However, we demonstrate that pieces of paper shredded by hand does not follow this pattern. Then, the second step consists in extracting relevant features of the polygon and using them to make the local reconstruction. In this way the overall complexity can be dramatically reduced because few features are used to perform the matching.

The ambiguities resulting from the local reconstruction are resolved and the pieces are merged together as we search for a global solution. We demonstrated by comprehensive experiments that this feature-matching-based procedure produces interesting results for the problem of document reconstruction. A global search is considered to reconstruct the entire document.

The remaining of this work is organized as follows: Section 2.1 presents an overview of the proposed methodology, Section 2.2 describes the feature set we have used to carry out the local matching. Section 2.3 shows how we compute the similarity between the polygons as well as the global search algorithm. Finally, Section 3 reports the experimental results and Section 4 present some perspectives of future works and concludes this work.

2 The Proposed Methodology

Our methodology is composed of three major steps as depicted in Figure 2. Initially, each piece of the document is pre-processed through polygonal approximation in order to reduce complexity of the boundaries. Then, a set of features is extracted from each polygon in order to carry out the matching. In the following sections we describe in details each component of the methodology.

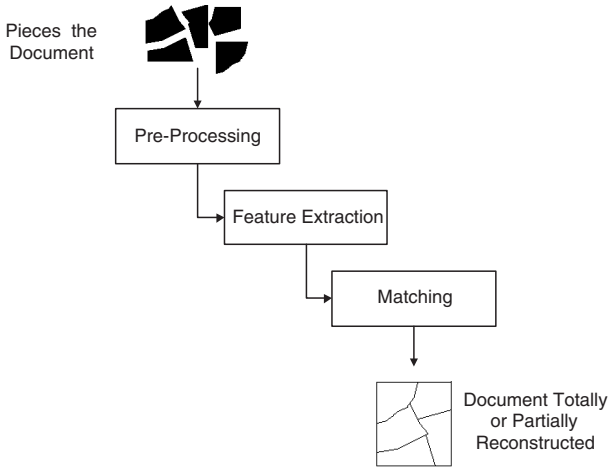


Figure 2. The block diagram of the proposed methodology.

2.1 Pre-Processing

Traditional puzzle solving algorithms usually take into account smooth edges and well defined corners. However, dealing with shredded documents is quite more complex. The act of shredding a piece of paper by hand often produces some irregularities in the boundaries, which makes it impossible to get a perfect curve matching. Figure 3 shows an example of this problem.

It can be observed from Figure 3b, that the fragment has two boundaries: the inner and the outer boundaries. The problem lies in the fact that when acquiring the images of this kind of fragments, the inner boundary is lost, and it is

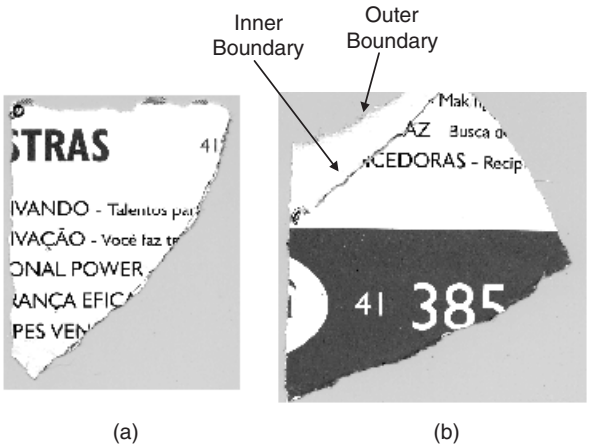


Figure 3. Inner and outer boundaries produced by shredding.

easy to see that the outer boundaries of the fragments (a) and (b) do not match perfectly.

In order to overcome this kind of problem, we have tested different algorithms, and the one that brought the best results was the well-known Douglas-Peucker (DP) algorithm [4]. This algorithm implements a polyline simplification and it is used extensively for both computer graphics and geographic information systems.

The DP algorithm uses the closeness of a vertex to an edge segment. This algorithm works from the top down by starting with a crude initial guess at a simplified polyline, namely the single edge joining the first and last vertices of the polyline. Then the remaining vertices are tested for closeness to that edge. If there are vertices further than a specified tolerance, $T > 0$, away from the edge, then the vertex furthest from it is added the simplification. This creates a new guess for the simplified polyline. Using recursion, this process continues for each edge of the current guess until all vertices of the original polyline are within tolerance of the simplification.

More specifically, in the DP algorithm, the two extreme endpoints of a polyline are connected with a straight line as the initial rough approximation of the polyline. Then, how well it approximates the whole polyline is determined by computing the distances from all intermediate polyline vertices to that (finite) line segment. If all these distances are less than the specified tolerance T, then the approximation is good, the endpoints are retained, and the other vertices are eliminated. However, if any of these distances exceeds the T tolerance, then the approximation is not good enough. In this case, we choose the point that is furthest away as a new vertex subdividing the original polyline into two (shorter) polylines.

This procedure is repeated recursively on these two shorter polylines. If at any time, all of the intermediate distances are less than the T threshold, then all the intermediate points are eliminated. The routine continues until all possible points have been eliminated. Figure 4 shows two different levels of approximation.

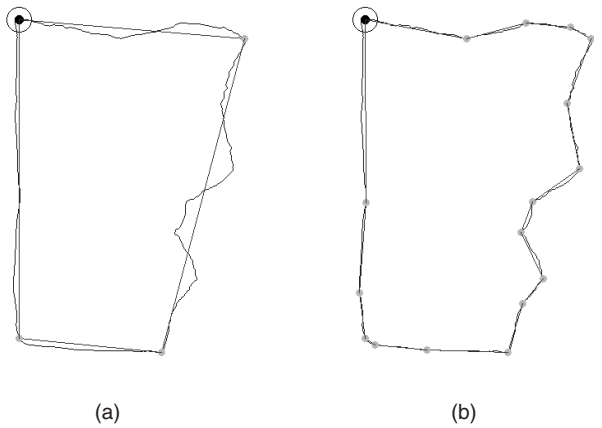


Figure 4. Inner and outer boundaries produced by shredding.

2.2 Feature Extraction

After the complexity reduction through polygonal approximation, the next step consists in extracting features to carry out the local matching. The feature extraction can be seen also as a complexity reduction process, since it converts the polygon in a sequence of features. Here, we propose a simple feature set that can be used to carry out the local matching.

The first feature is the angle of each vertex with respect its two neighbors. Consider for example the vertices A and B in the polygon depicted in Figure 5. The angle α is given by

$$cos(\alpha) = \frac{uv}{|u||v|} \quad (1)$$

We also verify whether such an angle is convex or concave. For example, in Figure 5, vertex B has a convex angle while vertex C has a concave one. To complete our feature set, we compute the distances between the vertex and its neighbors (next and previous in a clockwise sense). Such distances are achieved by means of the well-known Euclidean distance. Table 1 describes the feature vector extracted from the polygon depicted in Figure 5. The last two features are the coordinates of the vertex in the image.

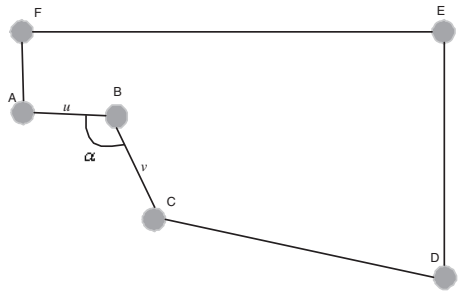


Figure 5. Angle features extracted from the polygon.

Table 1. Description of the feature vector.

Vertex	Angle	Distances		X	Y
		Next	Previous		
A	270	40.0	45.0	10	70
B	120	45.0	43.6	55	67
C	200	43.6	115.7	67	25
D	245	115.7	11.0	180	0
E	270	110.0	170.0	180	110
F	270	170.0	40.0	10	110

This table can be read as follows: The angle of the vertex B, which is computed by using vertex A and C, is 120 degrees. The Euclidean distances between B and its neighbors A and C are 45.0 and 43.6, respectively. The coordinates of the vertex B in the image are (55,67).

2.3 Matching

2.3.1 Computing the Similarity Between Polygons

The feature vector described so far allows us to compute a degree of similarity, which is used to measure the quality of the matching between two fragments of the document.

First of all, we verify the complementarity between the angles of the two vertices being compared. Generally speaking, both angles must sum up 360 degrees, as illustrated in Figure 6. Of course we have to consider some degrees of freedom, since they could not sum up 360 due to the estimations made during the polygonal approximation. If the complementarity is verified like in Figure 6, then $W_{angles} = 1$.

Thereafter the distances between the vertex and their neighbors are compared as illustrated in Figure 7. D_{p1} is the Euclidean distance between the vertex A_1 and its previous neighbor C_1. D_{n1} the Euclidean distance between the vertex A_1 and its next neighbor B_1. The distances D_{p2} and

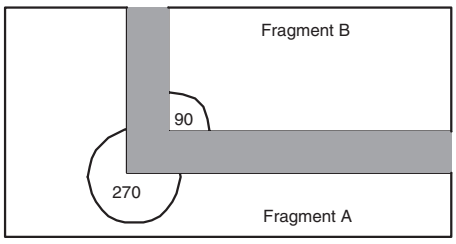


Figure 6. Similarity between angles.

D_{n2} are computed in the same way. After computing such distances, a measure of similarity $W_{matching}$ is calculated by using Equation 2.

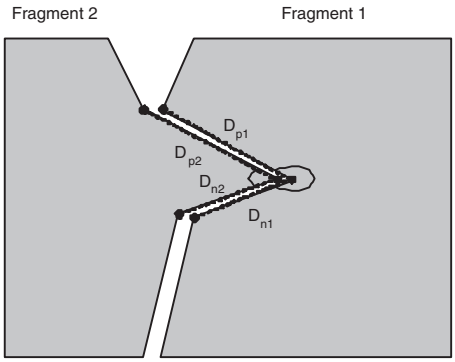


Figure 7. Distance features extracted from the polygon.

$$W_{matching} = \begin{cases} 1 & \text{if } [(D_{p1} \simeq D_{p2} \\ & \text{OR } D_{n1} \simeq D_{n2}) \\ & \text{AND } W_{angle} = 1] \\ & \\ 5 & \text{if } [(D_{p1} \simeq D_{p2} \\ & \text{AND } D_{n1} \simeq D_{n2}) \\ & \text{AND } W_{angle} = 1] \end{cases} \quad (2)$$

It is clear from Equation 2 that the weight is much more relevant when both distances are similar. These values were determined empirically through several experiments.

Finally, we consider the relevance of the matching regarding the perimeter of the fragment using the following rules:

- If the contour matched represents more than $1/5$ of the perimeter of the fragment, then $W_{matching} = W_{matching} + 2$.
- If the contour matched represents more than $1/10$ of the perimeter of the fragment, then $W_{matching} = W_{matching} + 1$.

- Otherwise, $W_{matching}$ is not increased.

After several experiments, we realized that these simple rules allow a more reliable identification of the relevant matchings.

2.3.2 Global Search

Once the metric to measure a matching has been defined, the next step consists in reconstructing the entire document. As stated somewhere else, this global search also eliminates the ambiguities resulting from the local reconstruction described in the previous section.

The method applied here is based on the algorithm proposed by Leitão and Stolfi [8], which tries to match two pieces at time. Let us consider a shredded document $D = \{F_1, F_2, \ldots, F_n\}$ composed of n fragments. The algorithm compares the fragment F_1 with all the other fragments searching for the best matching, i.e., the match that maximizes the $W_{matching}$ defined previously. Then, the fragments F_i and F_j that maximizes $W_{matching}$ are merged forming a new fragment F_{ij}.

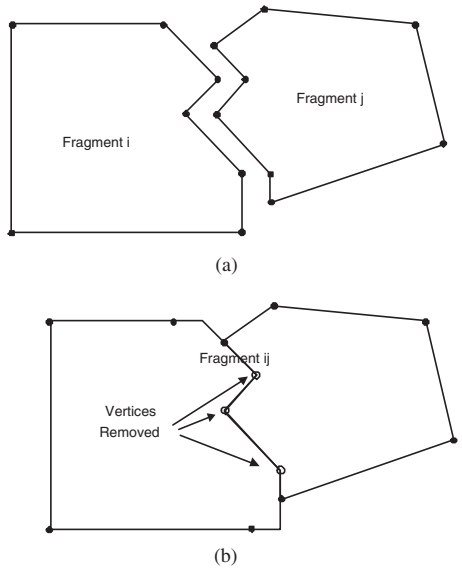


Figure 8. Best matching (a) Fragments i and j and (b) New fragment F_{ij} where three vertices were removed.

The feature vector of the new fragment F_{ij} is then modified by removing the vertices matched. Figure 8 shows this merging and the vertices removed as well.

After merging, the process starts again but now the document has $n - 1$ fragments. It ends when the number of fragments is 1 or none fragments have been merged. This procedure is described in Algorithm 1.

Algorithm 1 Global Search

```
1:  D = {F_1, F_2, ..., F_n}
2:  repeat
3:     best = NULL
4:     for i = 2 to n do
5:        Compute all possible W_matching for F_1 and F_i
6:        if there is a W_matching > 0 then
7:           best = i that maximizes W_matching
8:        end if
9:     end for
10:    if best ≠ NULL then
11:       F_new = F_1 ⋃ F_best
12:       Remove F_1 and F_best from D.
13:       Insert F_new into D
14:       n = n − 1
15:    end if
16: until n = 1 OR best ≠ NULL
17: return F_new
```

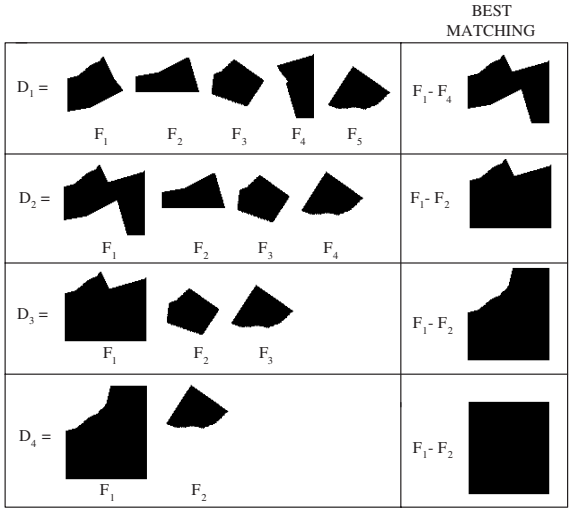


Figure 9. Steps of the document reconstruction.

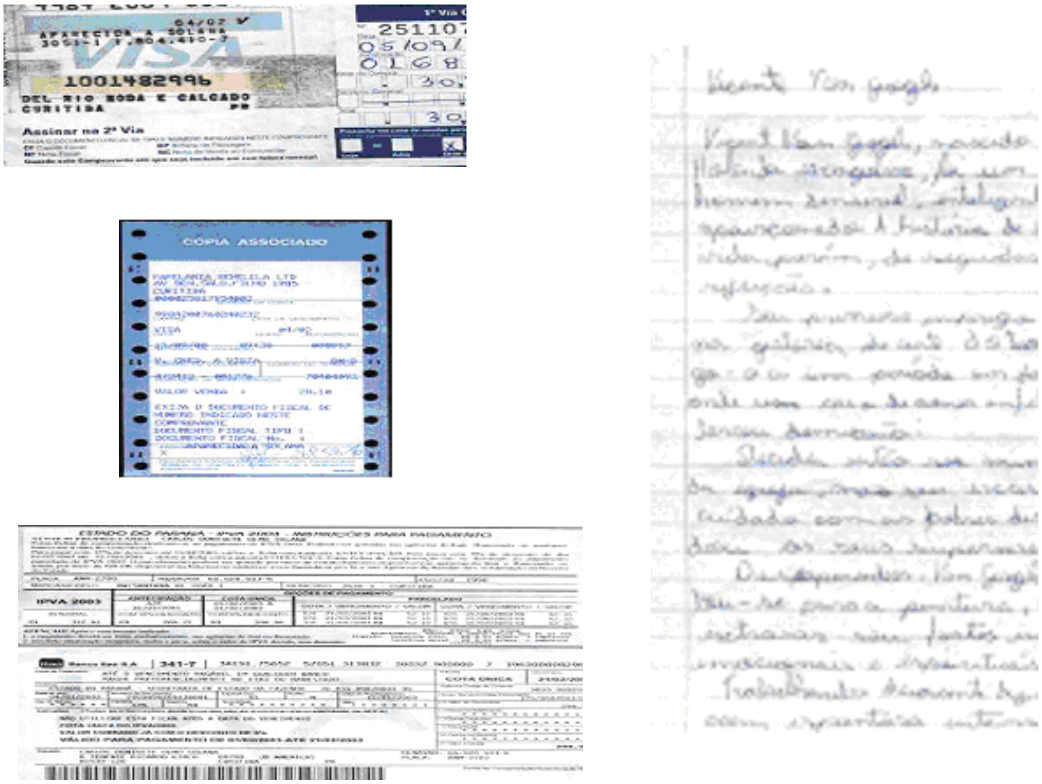


Figure 10. Examples of the documents in the database

Analyzing the algorithm, we can observe that it returns the document either totally or partially reconstructed. Figure 9 shows the results of the global search for a 5-fragment document. In such a case, the document was totally reconstructed.

3 Experiments

In order to validate the proposed methodology, we have built a database of shredded documents. Firstly, we have collected 100 diversified documents containing handwriting, machine printed, images, and graphs. Then, the documents were shredded into 3 to 16 fragments. It is worth of remark that the documents where randomly shredded, in other words, we have not used any criteria to perform this task. The idea was to create a database as close as possible of a real database. After shredding, the fragments of the documents were labelled and digitalized in 150 dpi, gray-level. Figure 10 shows some examples of the documents in the database, while Figure 11 shows an example of a document totally reconstructed

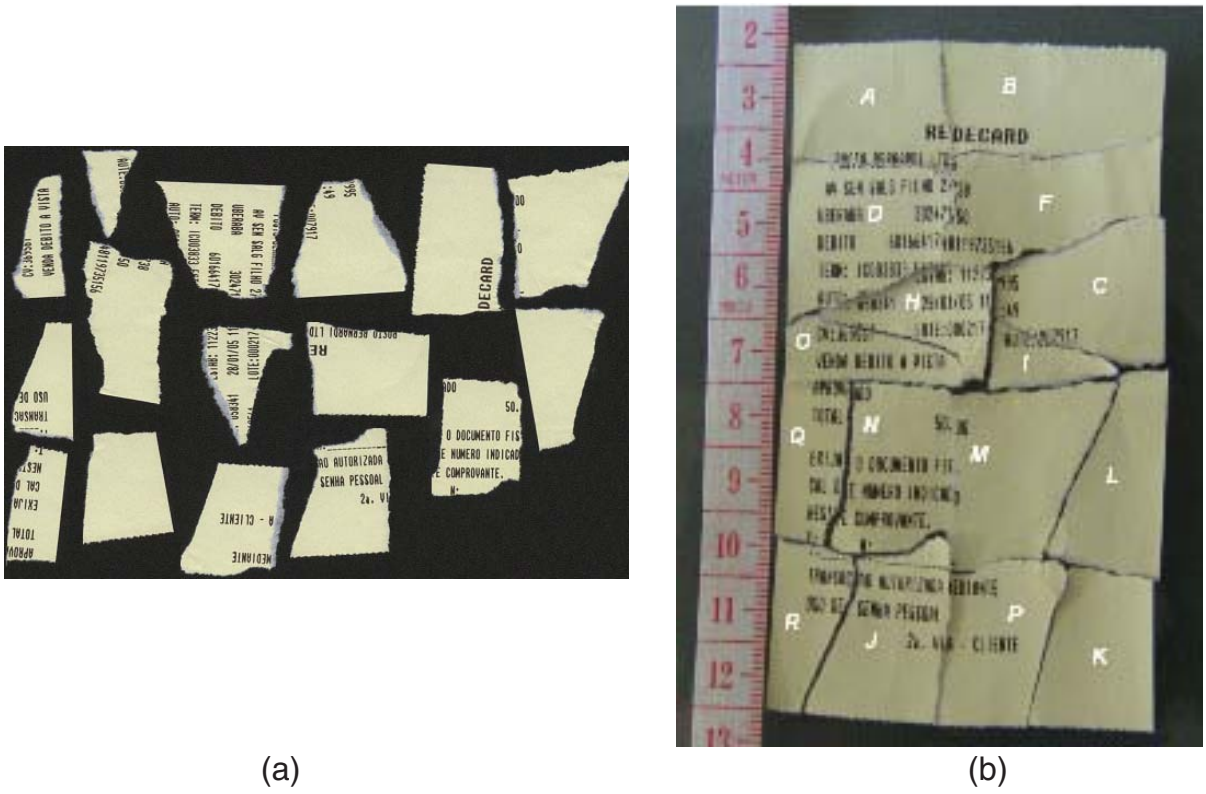
We have used 10% of the database to train the system (fine-tuning the parameters) and the remaining were used to performance evaluation. Figure 12 reports the average performance of the system as the number of fragments increases. As we can see, the performance drops as the number of fragments gets bigger. This is a drawback of the polygonal approximation. It allows us to reduce considerably the complexity of the matching process, but on the other hand, such an approximation makes it difficult to keep the same level of performance for document with large number of fragments.

A possible solution for that lies in moving from coarse to fine-scale alignment so that the effects of the approximation could be minimized. In the long run, it is a question of choosing the best trade-off between complexity and performance.

As we have mentioned before, the results produced by this kind of system are very useful even when the documents are partially reconstructed. In such cases an human will dispend considerably less efforts to finish reconstructing the document than starting from scratch.

4 Conclusion and Future Works

In this paper we have proposed a methodology for document reconstruction based on feature matching. It takes two steps where the former makes a approximation in order



(a) (b)

Figure 11. Examples of a document totally reconstructed: (a) Fragments and (b) Document reconstructed.

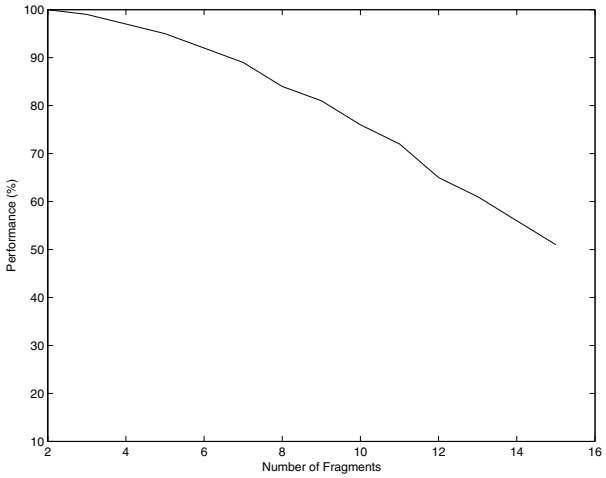


Figure 12. Performance of the proposed methodology in reconstructing documents shredded by hand.

to reduce the complexity of the boundaries and overcome specific problems faced in document reconstruction and the latter extracts relevant features of the polygon and uses them to make the local reconstruction.

The results we have shown demonstrates that the methodology, in spite of the fact of using few features, is able to reconstructed documents shredded by hand. As discussed previously, the performance drops as the number of fragments gets bigger due to the scale used during the polygonal approximation. This issue can be addressed by choosing the most important aspects for the application, i.e., reducing complexity or improving performance. It is worth of remark that both views are important. A less complex system, like the one presented here, could be applied initially and them a more complex one, hence more time consuming, could be applied to resolve final confusions.

As future works, we plan to make some experiments in this sense and also to analyze the performance of the system for other kinds of shredding as well.

References

[1] H. Bunke and U. Buehler. Applications of approximate string matching to 2-D shape recognition. *Pattern Recognition*, 26:1797–1824, 1993.

[2] H. Bunke and G. Kaufmann. Jigsaw puzzle solving using approximate string matching and best-first search. In *Procs of the 5th International Conference on Computer Analysis of Images and Patterns*, pages 299–308, 1993.

[3] I. Cohen, N. Ayache, , and P. Sulger. Tracking points on deformable objects using curvature information. In *Procs of the Second European Conference on Computer Vision*, pages 458–466, Santa Margherita Ligure, Italy, 1992.

[4] D. Douglas and T. Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *The Canadian Cartographer*, 10:112–122, 1973.

[5] W. Kong and B. Kimia. On solving 2d and 3d puzzles under curve matching. In *Procs of the Conference on Computer Vision and Pattern Recognition*, pages 583–590, Kauai, Hawaii, USA, 2001.

[6] A. Korzyk and W. Yurcik. On integrating human-in-the-loop supervision into critical infrastructure process control systems. In *Procs of the Advanced Simulation Technologies Conference*, San Diego, USA, 2002.

[7] H. C. G. Leitao and J. Stolfi. Digitalization and reconstruction of archaeological artifacts. In *Procs of the 14th Brazilian Symposium on Computer Graphics an Image Processing*, page 382, Florianopolis, SC, Brazil, 2001.

[8] H. C. G. Leitao and J. Stolfi. A multiscale method for the reassembly of two-dimensional fragmented objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24:1239–1251, 2002.

[9] T. B. Sebastian, J. J. Crisco, P. N. Klein, and B. Kimia. Constructing 2d curve atlases. In *Procs of the Mathematical Methods in Biomedical Image Analysis*, pages 70–77, 2000.

[10] G. Uoluk and I. Toroslu. Automatic reconstruction of broken 3-d surface objects. *Computer & Graphics*, 23:573–582, 1999.

[11] H. Wolfson. On curve matching. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12:483–489, 1990.

[12] H. Wolfson, E.Schonberg, A. Kalvin, and Y. Lamdan. Solving jigsaw puzzles by computer vision. *Annals of Operations Research*, 12:51–64, 1988.