NAVIDOMASS: Structural-based approaches towards handling historical documents

Salim Jouili, Mickaël Coustaty, Salvatore Tabbone, Jean-Marc Ogier

NAVIDOMASS: Structural-based approaches towards handling historical documents

Salim Jouili*, Mickael Coustaty†, Salvatore Tabbone* and Jean-Marc Ogier†
*LORIA-INRIA UMR 7503,
BP 239, 54506 Vandoeuvre-lès-Nancy Cedex, France
{salim.jouili, tabbone}@loria.fr
†L3i Laboratory
Avenue Michel Crepeau, 17042 La Rochelle, France,
{mcoustat, jmogier}@univ-lr.fr

***Abstract*—In the context of the NAVIDOMASS project, the problematic of this paper concerns the clustering of historical document images. We propose a structural-based framework to handle the ancient ornamental letters data-sets. The contribution, firstly, consists of examining the structural (i.e. graph) representation of the ornamental letters, secondly, the graph matching problem is applied to the resulted graph-based representations. In addition, a comparison between the structural (graphs) and statistical (generic Fourier descriptor) techniques is drawn.**

I. Introduction

Driven by rapidly changing amounts of digitized historical document, specific pattern recognition systems could undergoing a strategy shift. Indeed historical document images, such as ancient ornamental letters (i.e. decorated initials), are particularly difficult for the recognition process since it contains a lot of information (e.g. texture, decorated background, letters). Figure 1 illustrates some samples of ancient ornamental letters. One can remark that each image is a mixture of simple components such that the initial and the background pattern. In addition, some degradations linked to the state of original paper and the digitalization can be observed. To deal with this kind of properties, pattern recognition systems require specific techniques which take into account these characteristics.

The overall goal of the NAVIDOMASS[1] project is to develop such rigorous techniques of pattern recognition supporting the ancient documents specificity. In this paper, we consider the historical document representation analysis within the NAVIDOMASS context. Generally in pattern recognition, the document representation can be broadly divided into statistical and structural methods [2]. In the former, the document is represented by a feature vector, and in the latter, a data structure (e.g. graphs or trees) are used to describe objects and their relationships in the document. The classical recognition systems are often limited to work with a statistical representation due to the need of computing distances between documents (feature vectors) or finding a representative of a cluster of documents. However, when a numerical feature vector is used to represent the document, all structural information is discarded although the structural representation is more powerful in terms of its representational abilities [2]. Since we deal with ornamental letters images which are complex images, the structural approaches seem to be more suitable for the representation task. Guided by these observations, we propose in this paper a structural-based framework to handle the ancient ornamental letters data-sets.

Our contribution is divided into two axes: first of all we discuss the structural description of documents. Here, several ways can be considered. Indeed to represent an ornamental letter as a graph, one can use techniques used for graphical symbols, 3d objects or shapes [13]. Here, we use a new technique developed for the purpose of structural representation of ornamental letters. The proposed approach works as follows: firstly, we separate image into three components that are easier to process. The first component contains all shapes, the second component contains textures and the third one noise of image. We will not consider these two last components in this paper. In order to extract shapes from the first component, we used a Zipf Law that extract the most frequent pattern of image. The last step can be resumed by a selection of the biggest connected component. In addition, a classical Region Adjacency Graph (RAG) technique is also used to provide a comparison with the proposed technique. Secondly, we consider the distance between graphs. That is to say the optimal approximation of the graph edit distance since this problem is considered to be NP-Complete. Here, an improvement of the graph matching approach based on node signature and developed in the context of the NAVIDOMASS project is compared with the well-known graph edit distance approximation recently introduced by Riesen and Bunke [12].

In addition, even if the statistical and structural approaches

[1]This work is partially supported by the French National Research Agency project NAVIDOMASS referenced under ANR-06-MCDA-012 and Lorraine region. For more details and resources see http://navidomass.univ-lr.fr

comparison is not the aim of this paper, we also use the generic Fourier descriptor (GFD) which is well known for its good performance [14]. This comparison shows the general behavior of structural approaches vs a statistical one. The experiments are performed in a clustering context. We used the well-known K-means algorithm to cluster the lettrines data-set. The results are evaluated using two cluster indices validity: The GK-index [7] and the Dunn Index [5].

II. Ornamental Letters Database

In this paper, we are interested in clustering images extracted from documents of the fifteenth and sixteenth century. Among the huge mass of old documents available, we are particularly interested in ornamental letters, also called *lettrines*. They correspond to images widely used in books and very reused over time that begin a chapter or a paragraph (see Figure 1 for examples). The *Centre d'Etude Supérieur de la Renaissance*[2] provides us paper's documents from the Renaissance compound of two major types of particularities: the support and the period. Our database is composed of more than 4000 images degraded by time and printed by wood stamps.

(a) (b) (c) (d)

Figure 1. Samples of ancient ornamental letters images

III. Two description approaches

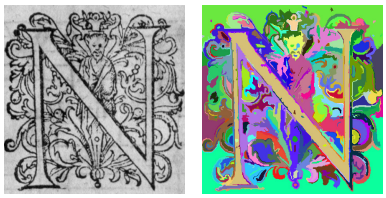
A. Region Adjacency Graph (RAG)

Each lettrine is represented by a Region Adjacency Graph.The nodes correspond to the homogeneous regions in the lettrine obtained by making use of the graph-based image segmentation [6] with a user-based parameterization technique. In figure 2, an illustration of the segmentation technique is applied to a lettrine. Each node is attributed with a vector of features which describe the area, perimeter and centroid coordinates of the corresponding region in the image. An edge between two nodes describes the adjacency relationships of the associated regions to the involved nodes.

B. Our description approach

1) Extraction and segmentation of images: Our aim is to catch the pure geometrical component in an image independently of texture and noise. For this, in [4], authors propose a decomposition model which splits an image into three components: the first one, u, containing the structure of the image (see Figure 3(b) for an example),

[2]CESR - University of Tours



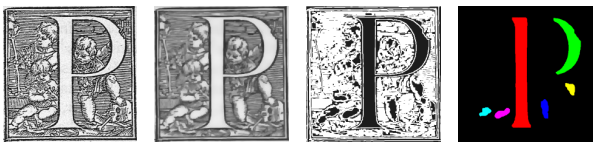
(a) Original image (b) Segmented image

Figure 2. Example of an image and its corresponding region-based segmentation

a second one, v, the texture, and the third one, w, the noise. For a better comprehension of different spaces, see [1]. In this article, we are particularly interested to the first one which capture regions with low variation of gray levels. From this image, we apply a Zipf law to extract areas.

Image simplification and segmentation. The first layer obtained by decomposition contain all shapes. In order to extract them and to select the most interesting, we used a *Zipf Law*. *Zipf Law* relies on the frequency and on the rank of appearance of words in a text. This law has been transposed on images [11] by taking sub-images as patterns and by calculating frequency and rank of these patterns. This method is a three steps process (see [3] for details):

- Simplify image by applying a 3-means on gray level histogram to reduce number of patterns (the choice of three can be explained by the fact that images are composed of three elements : background, foreground and motive)
- Seek for patterns of size three by three to obtain their frequency and their rank (that can be resume to a count of each pattern that permit to know their frequency and their rank)
- Classify patterns according to the evolution law of the frequency compared to their rank. From the precedent step, three straight lines are computed to estimate the main parameters of Zipf laws that interfere. The first one corresponds to the most frequent patterns (shapes of image) and an example of result is presented in Figure. 3(c)



(a) (b) (c) (d)

Figure 3. Example of an image and representation of different treatments: a) Original image; b) Region with low variation of gray levels; c) Shapes segmented using Zipf law (in black); d) Six larger connected component

2) Description of images: Once shapes have been extracted, one can seek connected components of binarized image. When we observe all the connected components in Figure 3(c), we can see that the most important shapes have particular characteristics (based on size, location, center of mass and eccentricity). A selection of connected components in accordance with these parameters permit to obtain region of interest of drop caps. An example of extracted connected components can be seen in Figure 3(d). From all these extracted shapes, we build a complete graph where nodes correspond to region of interest. Each node is described by a quadruplet of information which contains the coordinates of the center of mass, area and eccentricity of each shape. We are actually working on a deeper description by adding relation between shapes (distance and angle) on the vertices of graph.

IV. Graph Matching for indexing

Matching by minimizing edit distance gauges the distance between graphs by counting the least cost of edit operations needed to make two graphs isomorphic. A standard set of edit operations is given by insertions, deletions and substitutions. These edit operations are applied on both edges and nodes. In addition, a certain cost is associated with each of these operations. Obviously, for every pair of graphs A and B there exists different sequence of edit operations transforming A into B. However, the computation of the edit distance between two graphs involves not only finding a sequence of edit operations to transform one graph to the other, but also finding such a sequence that possesses the minimum total cost. Formally, The graph edit distance between two graphs A and B is given by:

$$d(A,B) = \min_{(e_1,..,e_k)\in\gamma(A,B)} \sum_{i=1}^{k} c(e_i)$$

where $\gamma(A,B)$ denotes the sequences of edit operations transforming A into B and $c(e_i)$ denotes the cost of the edit operation e_i.

In order to compute an optimal graph edit distance, several techniques have been proposed. In this paper, we consider an improvement of our previous work on the approximation of graph edit distance [8], [9] and the work of Riesen and bunke [12].

A. Riesen-Bunke approach

Riesen and al. [12] consider the approximation of the graph edit distance as an instance of an assignment problem. The method computes the edit distance between two graphs based on a bipartite graph matching by means of the Hungarian algorithm and provides sub-optimal edit distance results. Let A=(V_a,E_a) and B=(V_b,E_b) be two graphs, the authors formulate the assignment problem by $\mid V_a \mid + \mid V_b \mid \times \mid V_a \mid + \mid V_b \mid$ matrix. In this matrix, we can observe four parts; the first represents the costs of all possible node substitutions ($\mid V_b \mid \times \mid V_a \mid$) these costs are computed by the euclidean distance or the string edit distance (depends on the type of the involved labels), the second part represents the costs of all possible node deletions ($\mid V_b \mid \times \mid V_a \mid$) these costs are considered as constant values. The third part of the matrix represents the costs of all possible node insertions ($\mid V_b \mid \times \mid V_a \mid$) these costs are also considered as constant values. Finally, the last part corresponds to zeros ($\epsilon \rightarrow \epsilon$ substitution costs). Then, they apply the Hungarian algorithm to this matrix to define the optimal matching between A and B.

B. Our approach

In a previous work, the problem of the computation of an approximation of graph edit distance is solved by means of node signatures [9], [8], each node is associated with a multi-subsets where the subsets are a collection of degree of the node, the attributes of the node and the adjacent edge attributes. Here we add the degrees of the adjacent nodes to the signature for more structural information about the involved node. Given a graph G=(V,E,α,β), the node signature of $n_i \in V$ is defined as follows:

$$\gamma(n_i) = \left\{ \alpha_i,\ \theta(n_i),\ \{\theta(n_j)\}_{\forall ij\in E},\ \{\beta_{ij}\}_{\forall ij\in E} \right\}$$

where

- α_i the attribute of the node n_i.
- $\theta(n_i)$ the degree of n_i.
- $\{\theta(n_j)\}_{\forall ij\in E}$ the degrees set of the nodes adjacent to n_i.
- $\{\beta_{ij}\}_{\forall ij\in E}$ the attributes set of the incident edges to n_i.

Therefore, the need of computing a distance between two node signatures is very essential in further handling these signatures. Thus, a node signature distance is introduced based on the Heterogeneous Euclidean Overlap Metric (HEOM) which handles numeric and symbolic attributes. The overall distance between two heterogeneous vectors $\boldsymbol{i}$ and $\boldsymbol{j}$ is given by the function HEOM($\boldsymbol{i,j}$):

$$HEOM(i,j) = \sqrt{\sum_{a=0}^{A} \delta(i_a,j_a)^2} \qquad (1)$$

where a refers to one attributes of A and $\delta(i_a,j_a)$ is defined as:

$$\delta(i_a,j_a) = \begin{cases} 1 & \text{if } i_a \text{ or } j_a \text{ is missed} \\ \mathit{Overlap}(i_a,j_a) & \text{if } a \text{ is symbolic} \\ \mathit{rn\_diff}_a(i_a,j_a) & \text{if } a \text{ is numeric} \end{cases}$$

Afterwards, using the node signatures and this distance, the graph edit distance problem is reformulated as an instance of the assignment problem, which can be solved by the Hungarian method [10].

Description	Distance	GK-Index	Dunn Index
RAG	NS♦	-0.0016	0.100
	R&B★	-	-
OD®	NS♦	-0.008	**0.862**
	R&B★	**-0.0013**	0.447
GFD	Euclidean	-0.0019	0.056

♦Proposed graph matching based on node signature
★Riesen and Bunke method in [12]
®Proposed graph representation of lettrines using the decomposition model

Table I
CLUSTERING RESULTS

V. CLUSTERING EVALUATION

In the experiment, three instances of the lettrine data-set are considered: a graph-based representation using the RAG technique (cf. §3.1), a graph-based representation using the proposed technique (cf. §3.2) and a features vector representation using the GFD descriptor. The k-means algorithm is applied to the three instances of the lettrine data-set with the same parameter (k=4 which is the number of lettrine's classes provided by the historians). Both the Riesen-Bunke approach and the proposed graph matching technique are used to compute distances between graphs. Nevertheless, to define the constant cost of node and edge deletions for the Riesen-Bunke method, we processed empirically by choosen a subset from the lettrine data-set and computing clustering validation indices for a set of cost values, than we take the best ones achieved on the selected subset. Let us remind that high values of the Dunn and GK indices correspond to the better clustering. The achieved results are reported in Table I, and show that the better results based on the Dunn index are provided by the combination between the proposed graph-based description and the graph matching by means node signatures. Based on the GK-index the better results are provided by the combination of the Riesen & Bunke graph matching method and the proposed structural description. However, unlike our graph matching technique, the Riesen & Bunke graph matching method can not successfully achieve the experiments with the RAG description in reasonable time (7 days of running without results). This is due to the very large graphs provided by the RAG description (the mean number of nodes by a graph is 108.3). Consequently, the results of the proposed representation seems more appropriate than the RAG. In the other side, the results provided from the use of the GFD descriptor are the lower which show that the structural-based approaches are more appropriate than the statistical one to handle images of ancient document, especially the lettrines.

VI. CONCLUSION

In this paper, we discussed the structural description of historical documents, especially the images of ornamental letters (Lettrines). A new technique for Lettrine's graph-based representation is introduced and compared with a classical Region Adjacency Graph technique. In addition, our previous work regarding the graph matching is improved by the definition of a new component in the proposed signature. The experimental results have shown that the combination of the proposed graph representation technique and a suited graph matching technique provides a better performance than the statistical technique (GFD).

REFERENCES

[1] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher. Structure-texture image decomposition - modeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67(1):111–136, 2006.

[2] H. Bunke, S. Günter, and X. Jiang. Towards bridging the gap between statistical and structural pattern recognition: Two new concepts in graph matching. In *ICAPR*, pages 1–11, 2001.

[3] M. Coustaty, J.-M. Ogier, R. Pareti, and N. Vincent. Drop Caps Decomposition For Indexing - A New Letter Extraction Method. In *10th ICDAR*, pages 476–480.

[4] S. Dubois, M. Lugiez, R. Péteri, and M. Ménard. Adding a noise component to a color decomposition model for improving color texture extraction. *Proceedings CGIV 2008 and MCS08*, pages 394–398, 2008.

[5] J. C. Dunn. Well separated clusters and optimal fuzzy-partitions. *Journal of Cybernetics*, 4:95–104, 1974.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[7] L. Goodman and W. Kruskal. Measures of association for cross-classifications. *J. Amer. Statist. Asso.*, 1954.

[8] S. Jouili, I. Mili, and S. Tabbone. Attributed graph matching using local descriptions. In *11th Int. Conf. on ACVIS, LNCS 5807*, pages 89–99, 2009.

[9] S. Jouili and S. Tabbone. Graph matching based on node signatures. In *7th IAPR-TC-15 Int. Workshop on GbRPR, LNCS 5534*, pages 154–163, 2009.

[10] H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

[11] R. Pareti and N. Vincent. Ancient initial letters indexing. In *18th Int. Conf. on Pattern Recognition*, pages 756–759, Washington, DC, USA, 2006. IEEE Computer Society.

[12] K. Riesen and H. Bunke. Approximate graph edit distance computation by means of bipartite graph matching. *Image Vision Comput.*, 27(7):950–959, 2009.

[13] K. Riesen, M. Neuhaus, and H. Bunke. Graph embedding in vector spaces by means of prototype selection. *IAPR Workshop on GbRPR, LNCS 4538*, pages 383–393, 2007.

[14] D. Zhang and G. Lu. Shape-based image retrieval using generic fourier descriptor. *Signal Processing: Image Communication*, 17:825–848, 2002.