Visual Analysis System for Features and Distances Qualitative Assessment: Application to Word Image Matching

Frédéric Rayar, Tanmoy Mondal, Sabine Barrat, Fatma Bouali, Gilles Venturini

Visual Analysis System for Features and Distances Qualitative Assessment: Application to Word Image Matching

Frédéric Rayar*, Tanmoy Mondal*,
Sabine Barrat*, Fatma Bouali*†, Gilles Venturini*
*Université François Rabelais, LI EA-6300, Tours, France
Email: name@univ-tours.fr
†Université de Lille 2, IUT, Dpt STID, Lille, France
Email: fatma.bouali@univ-lille2.fr

***Abstract*—In this paper, a visual analysis system to qualitatively assess the features and distance functions that are used for calculating dissimilarity between two word images is presented. Computation of dissimilarity between two images is the prerequisite for image matching, indexing and retrieval problems. First, the features are extracted from the word images and a distance between each image to others is computed and represented in a matrix form. Then, based on this distance matrix, a proximity graph is built to structure the set of word images and highlight their topology. The proposed visual analysis system is a web based platform that allows visualisation and interactions on the obtained graph. This interactive visualisation tool inherently helps users to quickly analyse and understand the relevance and robustness of selected features and corresponding distance function in a unsupervised way, *i.e.* without any ground truth. Experiments are performed on a handwritten dataset of segmented words. Three types of features and four distance functions are considered to describe and compare the word images. Theses material are leveraged to evaluate the relevance of the built graph, and the usefulness of the platform.**

I. Introduction

In Document Image Analysis (DIA) field, describing and comparing image entities, such as words or characters, are challenging tasks. Indeed, the choice of features and distances is a crucial step and has a deep impact on many applications like optical character recognition, word-spotting or writer identification. One can find many papers where new features or distances are proposed to address new challenges: difficult scripts, historical documents or camera-based input.

In most of these works, an evaluation of the proposed features or distances is done to assess the quality of the contributions regarding an objective and the studied dataset(s). Both qualitative and quantitative experiments can be performed. In this process of evaluation, one usually performs the following steps: *(i)* use a public accessible dataset or create one, *(ii)* create the ground-truth if not already available, *(iii)* extract the proposed features, *(iv)* then use a common classifier and cross-validation to generate some metrics, for instance the accuracy or precision-recall. However, obtaining data with ground-truth which requires quite a lot of time and effort. Furthermore, after using this kind of procedure, it is difficult to look into what went wrong and why, like misclassification or erroneous recognition. Thus, one may miss a chance to reassess the used features or distances and improve their relevance.

Visual analytic is a paradigm that combines automated analysis methods with interactive visual interfaces to allow one to reason and understand complex datasets. Only a few works leverage this paradigm to analyse documents entities. For example, [1] and [2] use graphs to study the distribution of handwritten digits, while [3] and [4] study font distribution. In [5], multidimensional scaling is used to visualise the diversity of japanese handwritten hiragana. In these works, features and distances are chosen and not discussed afterwards.

In this paper, we propose a visual analysis system that allows a fast and qualitative assessment of word images features and distances. First, features are extracted from word images and a distance matrix is computed. This matrix is leveraged to structure the word images in a relative neighbourhood graph (RNG). The RNG allows to highlight the topology of the data: it gives the global distribution and underlines the similarity between the data. Second, a web platform is used to visualise and interact with the graph. The global visualisation is a good qualitative clue to assess the quality of the features or distances. One can also zoom to study local neighbourhood of a node or explore the graph neighbour by neighbour while having a visual feedback. This interaction allows one to discuss some adjacencies that can be erroneous and put into question the choice of features or distances.

The rest of the paper is organised as follows: Section II presents the used graph for structuring the data and outlines some specifications of the proposed visual analysis system. Section III presents the material that is used to evaluate the proposed visual system. In Section IV, we present the performed experiments to evaluate the relevance of the built

graph and the usefulness of the platform with some use cases. Finally, we conclude our study in Section V.

II. Methods and platform

A. Relative neighbourhood graph

The relative neighbourhood graph has been introduced in the work of Toussaint [6]. The construction of this graph is based on the notion of *relatively close* neighbours, that defines two vertices as relative neighbours if they are at least as close to each other as they are to any other vertices. From this definition, we can define $RNG = (V, E)$ as the graph built from the vertices of D where distinct vertices p and q of D are connected by an edge $\overline{pq}$ if and only if they are relative neighbours. Thus,

$$E(RNG) = \{\overline{pq} \mid p, q \in D, p \neq q,$$
$$\delta(p, q) \leq max(\delta(p, r), \delta(q, r)), \forall r \in D \backslash \{p, q\}.$$

where $\delta : D \times D \rightarrow \mathbb{R}$ is a distance function. An illustration of the *relative neighbourhood* of two vertices $p, q \in \mathbb{R}^2$ is given in Figure 1.

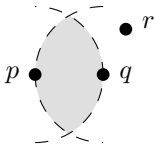

Figure 1: Relative neighbourhood (grey area) of two vertices $p, q \in \mathbb{R}^2$. If no other vertex lays in this neighbourhood, then p and q are relative neighbours.

The choice of the RNG is justified as follows: on one hand, the main drawback of the RNG is its construction. The classical and brute-force construction has a complexity of $O(n^3)$. However this issue has been addressed [7], [8]. On the other hand, the RNG is a connected graph that highlights the topology of the data and embeds local information about vertices neighbourhood. This graph has been used in [2], [3] and [4] over the minimum spanning tree [9], due to its properties. Furthermore, the connectivity property guarantees that each word images can be reachable during a content-based exploration of the graph.

B. Visual analysis system description

The proposed visual analysis system allows to visualise and interact with graphs of images. The graph that is described in the previous section highlights the topology of the studied data. Thus, in our case, one should expect the following observations: *(i)* similar word images may be linked by an edge, *(ii)* on the contrary, dissimilar word images should not be linked, or at least by a long edge. This last phenomenon occurs because of the connectivity property of the selected proximity graph. Figure 2 shows the platform interface.

The proposed visual analysis system has been realised using web technologies, namely HTML5, CSS3 et Javascript. This choice is explained as follows:

Figure 2: Interface of the proposed visual analysis system.

1) The platform target is image analysis researchers, but also experts from others fields such as health or digital humanities. Indeed, these later could bring a different point of view of the network analysis and thus raise different questions about the features and distance. Nowadays, a majority of users that are not experts in computer science can still manage well web navigation. Thus, such users are familiar with web browsers. We think that presenting the system as a light web platform, would make users more disposed to exploit it.
2) The platform is only client-side and do not use any server. This can be justified by two arguments: *(i)* no upload of the pictures to a server is needed, operation that may cost time and *(ii)* as the images are not sent nor stored in an external server, we respect the potential confidentiality or license issues that are related to the images.

Some of the visualisation and interaction specifications of the platform are given below:

- First, the graph is laid out and shown as a whole. The user can quickly have a glimpse on the global distribution of the word images and their topology. If the features and distance are discriminant, one can already perceive some connected components.
- The user can display the word images related to each node. Thus, the network can be interpreted more easily.
- To have a better view of the word images, the user can move the pointer above a node: a larger thumbnail is displayed.
- When a node is selected, a zoom is performed and the focus is set on it. The adjacencies of this node are highlighted. Thus, the user can study local neighbourhood of a node.
- In addition, when a node is selected, related information is displayed on the right of the graph (*cf.* Figure 2). Among these details, the larger thumbnail is displayed, and clicking on it allows to display the image at its original size. The relative neighbours of the selected graph are also displayed, and the user can perform a step by step exploration.
- Also, a history of the user navigation is stored and displayed, allowing him to return to any previous step of his exploration.

III. Material description

A. Dataset

The Bentham Dataset [10] consists of a series of documents from the Bentham collection. It has been prepared in the tranScriptorium project [1]. This dataset mainly includes manuscripts that were written by Jeremy Bentham (1748-1832) himself over a period of sixty years, as well as written copies by Bentham's secretarial staff. In our experiments, 100 word images have been selected and ground-truthed. The ground-truth has been leveraged only to evaluate the relevance of the graph that is used.

B. Features Extraction

Once the segmented words are obtained from the datasets, the next task is to extract the useful features from them. Both gray scale and binary height normalised images are used to extract features. In the following section, we describe three different categories of features: namely column-based features, histogram of gradient (HOG) based features and block level improved HOG features. These features, that are leveraged in the literature for word image description, are used throughout our experimental process.

1) Column-based features: we leverage here a set of statistical column-based features, used previously for handwriting recognition [11]. Although these features can be outperformed in terms of accuracy by more complex features (*e.g.* gradient based features, graph similarity features, *etc.*), they remain quite interesting due to their less computational cost and comparative accuracy. Here, we have chosen 8 features, $F_1, F_2, \ldots, F_8$ to define each pixel column. The description of the features F_i is given below in Table I. Thus, for an image with a pixel width of N, sequences of feature vectors are obtained by moving from the left to right over the word image. Please see [12] for the details of this column-based features.

2) Slit style HOG based features: the slit style HOG (SSHOG) [13] is a specially modified version of HOG [14], to make it suitable for *word spotting* applications. A fixed sized slit window is slid over the image in an horizontal direction for extracting features from each slit. Please see [13] for more details on SSHOG.

3) Block style HOG based features: the classical HOG descriptor was improved by Felzenszwalb et al. [15]. In these experiments, we also use Felzenszwalb's implementation. For a given height normalised word image of $M \times N$ pixels, the image is divided into fixed size cells of size c pixels. We extract HOG descriptor, that consists of 31 individual features, from each cell. Thus, we get a $m \times n$ matrix of HOG descriptors, where $m = \frac{M}{c}$ and $n = \frac{N}{c}$. Finally, we create a $(31 * m) \times n$ matrix, where for each column, we concatenate the m HOG descriptors of the $m \times n$ matrix.

[1] http://transcriptorium.eu/ icdar15kws/data.html

C. Distance functions

After extracting the features from each image, four distance functions are considered for calculating the dissimilarity between the images. The ideas behind these algorithms are explained in this section.

1) Dynamic Time Warping (DTW): this technique [16] is a dynamic programming (DP) based approach for calculating the optimal correspondence between two feature sequences X and Y. To align these two sequences, we construct a matrix, where each element of the matrix corresponds to the squared distance between elements of the sequences. Then, DTW computes a path cost matrix $\mathfrak{P}$ using dynamic programming.

2) Itakura Parallelogram: to speed up DTW and to avoid pathological matching, constraints are widely imposed for the calculation of warping path. It reduces time complexity by limiting the number of cells that are evaluated in the cost matrix. Itakura band [16] is a global constraint that gives an efficient trade-off between accuracy and speed, when it is defined properly.

3) Pseudo Local DTW (LDTW): this approach extends the DTW algorithm to perform pseudo-local alignment using a specific DP-path [17]. It applies different DP paths at different location of path cost matrix ($\mathfrak{P}$) for handling stretching and compression of individual points in time series data.

4) CDP: this technique [18] is able to perform subsequence matching (full query in longer target) and to locate multiple occurrences of the query in the target. Even so, this algorithm works well with properly segmented words.

IV. Results and discussion

A. Graph relevance

In order to evaluate the relevance of the graph, we have used the *Bentham* dataset. The main objective is to check whether the observations that can be done with a classical information retrieval metric can also be done using the graph. Using the ground-truth, the mean average precision (mAP) of each feature and distance pair have been computed. Furthermore, the RNG has been built for each of the mentioned pairs, and a graph metric has been calculated.

First, let us define the mean average precision. Given a query, we define Rel as the set of relevant similar word images with regard to the query and Ret as the set of ranked retrieval results from the dataset. The precision at k, noted $P@k$ is obtained by computing the precision by considering only the k top most results that are returned by the system. The mAP is the average of the precision at k for each relevant answers in the ranked retrieval results. Let $r(k)$ be the binary function on the relevance of the k^{th} item in the returned ranked list, the mAP is calculated as follows:

$$mAP = \frac{\sum_{k=1}^{|Ret|} P@k \times r(k)}{|Rel|}$$

Table I: Extracted features from the word images, considering an image with N columns and M rows

Sr. No	Feature set description
F1.	Projection profile of sequence
F2.	Background-to-ink transition in pixel column
F3.	Upper profile of sequence
F4.	Lower profile
F5.	Distance between upper and lower profile
F6.	Number of foreground pixels in pixel columns
F7.	Center of gravity (C.G.) of the column obtained from the foreground pixels
F8.	Transition at C.G. obtained from F7

Table II gives the average mAP values that have been computed for each presented feature and distance pair in Section III. Each image have been considered as a query, and searched in the remaining 99 word images. Then, the average mAP on the hundred queries has been computed.

Table II: Mean average precision for the Bentham dataset over studied features and distances.

	CDP	DTW	Itakura	LDTW
column-based	0.23	0.22	0.26	0.23
SSHOG	0.03	0.11	0.09	0.10
BlockHOG	0.19	0.26	0.28	0.24

Second, we define the metric that has been considered to evaluate the computed graphs. Let us consider a graph $G = (V, E)$, where V and E are the set of nodes and edges of G, respectively. For each node $n \in V$, we compute the precision $P(n)$ given by $P(n) = \frac{TP(n)}{RN(n)}$, where $TP(n)$ corresponds to the number of relative neighbours of n that have the same class as n and $RN(n)$ corresponds to the number of relative neighbours of $n \in G$. Then we compute the average precision $P(G)$ of the graph with

$$P(G) = \sum_{n \in V} \frac{P(n)}{|V|},$$

where $|V|$ is the number of nodes of G. Table III gives the average precision of the computed graphs for each feature and distance pair that are presented in Section III.

Table III: Graph average precision for the Bentham dataset over studied features and distances.

	CDP	DTW	Itakura	LDTW
column-based	0.37	0.32	0.30	0.38
SSHOG	0.12	0.21	0.20	0.26
BlockHOG	0.32	0.39	0.41	0.37

As we can see in Table II and Table III, the graph metric allows one to make the same observations than the mean average precision. For instance, regardless of the distance function, one can state that SSHOG performs less well than the column-based feature or the BlockHOG. As well, if we choose one set of features, we can rank the distance functions and decide which one is more prone to perform well with the selected features.

Thus, the generated graphs highlight the quality of the chosen features and distances as well as a classic information retrieval metric.

B. *Visual analysis system usefulness*

In this section, we illustrate the usefulness of the proposed visual analysis system with a set of use cases. These use cases are scenarios that could be helpful to DIA experts when assessing the quality of the features, the distance function or both. Here, the experiments are done in an unsupervised way, *i.e.*, the ground-truth of the word images is not presented. A discussion about some limitations of the platform is also presented at the end of this section.

1) Global visualisation: the first visualisation that is presented to the user is the whole graph. It could be laid out thanks to a drawing algorithm, namely *Force Directed* algorithm, to underline the topology of the graph, and hence the dataset. Leveraging this visualisation, one can quickly have a glimpse of the distribution of the images. On one hand, *communities* (*i.e.* group of nodes that are densely connected between them) may highlight very similar word images, *wrt.* the chosen features and distance pair. On the other hand, a graph with no structure may be synonym of a non discriminative pair. In Figure 3, one can observe a graph that displays a community and other dense groups of nodes while in Figure 4, no immediate structure appears.

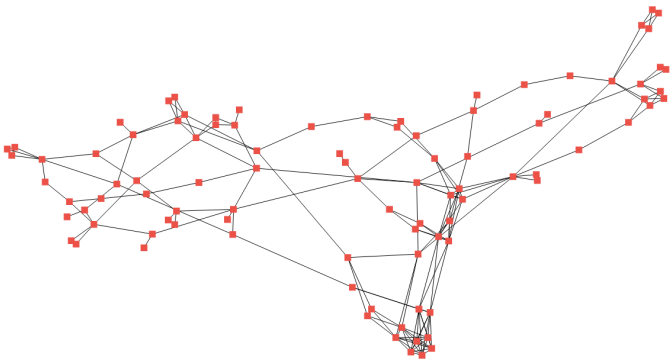

Figure 3: RNG drawing using BlockHOG features and Itakura distance. One can observe a community in the bottom of the graph.

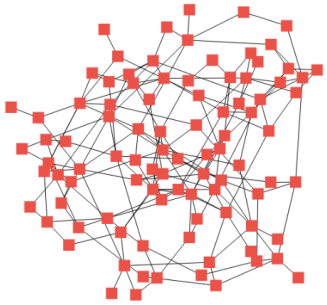

Figure 4: RNG drawing using SSHOG features and Itakura distance. One can observe that no specific graph structure is highlighted.

Figure 6: Using the same features and distance function as in Figure 5, one can observe that an irrelevant adjacency has been created.

2) Local neighbourhood study: one of the main goals of the proposed platform is to allow the user to locally study parts of the graph. Thus, using available interactions, one can zoom on a region of interest of the graph, and select a given node. Then, one can visualise and make observations on the neighbourhood of the selected node. Figure 5 illustrates a case where a word image is correctly linked to its similar images. At the same time, it is linked to other images which are not really similar. This could be either due to the choice of the features and distance pair or because it is a consequence of the graph connectivity property. In Figure 6, we observe a situation where a node has only one relative neighbour and the related word images are not similar at all. Here the bad segmentation of the bottom word image could be an explanation.

Figure 5: RNG drawing using BlockHOG features and LDTW distance. The center word image is relevantly linked to similar word images.

3) Feature evaluation: the proposed visual analysis system can be used to evaluated features only. Indeed, if one leverages only one distance function, he could visualise the generated graph over different features. Then, observations can be made either on the whole graphs (*e.g.* which features generate a graph with interesting structures) or be more thoroughly on a local level. Figure 7 illustrates a local study. The Itakura distance has been selected, and we observed the relative neighbours of a given word image ("*Majesty*") on the three RNG built considering the three features that are mentioned in Section III. One can observe that the precision is not the same. For BlockHOG features(*cf.* Figure 7a), two similar images have been correctly linked. Regarding the column-based features (*cf.* Figure 7b), only one among two relative neighbours is relevant. Finally, SSHOG features (*cf.* Figure 7c) perform less well: they link only one similar word image among three and this relevant neighbour is the farthest among the three. Thus, based on such local mining on several word images of the graph, one can draw conclusions about the relevance and robustness of the features.

4) Distance evaluation: in the same way, the proposed visual analysis system can also be used to evaluated distance functions only. Indeed, if one selects a feature, he could visualise the generated graph considering different distance functions. Thus, it is possible to quickly select the most relevant distance function for a given feature but also the most relevant feature for a given distance function. Note that this observation shall be done only on the considered dataset and must be tested over others datasets to assess its relevance.

5) Limitations: during our experiments, we underlined some limitations of the proposed visual analysis system. First, one should not draw some conclusions based only on the global topology of the graph. Indeed, the display of the graph depends on the choice of the drawing algorithm. Several drawing algorithms have been proposed in the graph community [19]. In this work, we have chosen for a *Force Directed* algorithm, that allows to highlight the graph structure. Second, regarding the number of considered nodes in the qualitative evaluations. Indeed, it is possible to build the RNG for large collection of images [8] in acceptable times. This may be useful to study the distribution of large datasets. However, an issue occurs with the visualisation. Indeed, one can face problems such as the overlapping of nodes, hence word images. These occlusions may be an issue for a thorough local study. Furthermore, one must be aware that the interactions will not be as smooth, due to memory issue.

(a) (b) (c)

Figure 7: Ranked relative neighbours of a given word image : "*Majesty*" using the same distance, namely Itakura, and using (a) BlockHOG features, (b) column-based faetures and (c) SSHOG features. One can observe different precisions depending on which features are considered.

V. Conclusion

In this paper, we proposed a visual analysis system that allows one to perform a fast and qualitative assessment of features and distance functions of document entity images. A proximity graph is used to structure and visualise the set of images in a custom web platform. Experiments are done considering the evaluation of features and distance functions for word images. The graph relevance is underlined and the usefulness of the proposed visual system is highlighted by a set of use cases. Future works will mainly focus on the improvement of the platform. Indeed, one can easily think about more visualisations and interactions, such as multi-faceted thumbnails to improve the evaluation of either only the features or only the distance function. Feature selection interactions could also be embedded to further understand the impact of features. Besides, graph partitioning techniques could be used to highlight possible communities.

Acknowledgment

The authors wish to thank Dr. Nicolas Ragot for his feedback on this work.

References

[1] S. Uchida, R. Ishida, A. Yoshida, W. Cai, and Y. Feng, "Character image patterns as big data," in *ICFHR*, 2012, pp. 479–484.

[2] M. Goto, R. Ishida, Y. Feng, and S. Uchida, "Analyzing the distribution of a large-scale character pattern set using relative neighborhood graph," in *12th International Conference on Document Analysis and Recognition, Washington, DC, USA, August 25-28, 2013*, 2013, pp. 3–7.

[3] C. Nakamoto, R. Huang, S. Koizumi, R. Ishida, Y. Feng, and S. Uchida, "Font distribution observation by network-based analysis," in *Camera-Based Document Analysis and Recognition - 5th International Workshop, CBDAR 2013, Washington, DC, USA, August 23, 2013*, 2013, pp. 83–97.

[4] S. Uchida, Y. Egashira, and K. Sato, "Exploring the world of fonts for discovering the most standard fonts and the missing fonts," in *13th International Conference on Document Analysis and Recognition, Nancy, France, August 23-26, 2015*, 2015.

[5] Y. Akao, A. Yamamoto, and Y. Higashikawa, "Feasibility study of visualizing diversity of japanese hiragana handwritings by mds of earth mover's distance toward assisting forensic experts in writer verification," in *11th International Workshop on Document Analysis Systems, DAS 2014, Tours, France, April 7-10, 2014*, 2014, pp. 26–30.

[6] G. T. Toussaint, "The relative neighbourhood graph of a finite planar set," *Pattern Recognition*, vol. 12, pp. 261–268, 1980.

[7] M. Goto, R. Ishida, and S. Uchida, "Preselection of support vector candidates by relative neighborhood graph for large-scale character recognition," in *13th International Conference on Document Analysis and Recognition, Nancy, France, August 23-26, 2015*, 2015.

[8] F. Rayar, S. Barrat, F. Bouali, and G. Venturini, "An approximate proximity graph incremental construction for large image collections indexing," in *Proceedings of Foundations of Intelligent System 22nd International Symposium, ISMIS 2015*, 2015.

[9] O. Boruvka, " O Jistém Problému Minimálním (About a Certain Minimal Problem) (in Czech, German summary)," *Práce Mor. Prírodoved. Spol. v Brne III*, vol. 3, 1926.

[10] B. Gatos, G. Louloudis, T. Causer, K. Grint, V. Romero, J. A. Sánchez, A. H. Toselli, and E. Vidal, "Ground-Truth production in the tranScriptorium project," in *11th IAPR International Workshop on Document Analysis Systems (DAS)*, 2014.

[11] U. Marti and H. Bunke, "Using a statistical language model to improve the performance of an HMM-based cursive handwriting recognition system," *International Journal of Pattern Recognition*, vol. 15, pp. 65–90, 2001.

[12] T. Mondal, N. Ragot, J.-y. Ramel, and U. Pal, "Flexible Sequence Matching Technique: Application to Word Spotting in Degraded Documents," in *ICFHR*. IEEE, Sep. 2014, pp. 210–215.

[13] K. Terasawa and Y. Tanaka, "Slit Style HOG Feature for Document Image Word-Spotting," *ICDAR*, pp. 116–120, 2009.

[14] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 886–893.

[15] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminative Trained Part Based Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[16] T. Albrecht, "Dynamic Time Warping (DTW)," pp. 69–85, 2009.

[17] J. Listgarten, R. M. Neal, S. T. Roweis, and A. Emili, "Multiple Alignment of Continuous Time Series," *Advances in Neural Information Processing Systems*, vol. 17, no. 17, pp. 817–824, 2005.

[18] R. Oka, "Spotting method for classification of real world data," *The Computer Journal*, vol. 41, no. 8, pp. 1–6, 1998.

[19] I. Herman, I. C. Society, G. Melançon, and M. S. Marshall, "Graph visualization and navigation in information visualization: a survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, pp. 24–43, 2000.