Using Earth Mover's Distance in the Bag-of-Visual-Words Model for Mathematical Symbol Retrieval

(Article begins on next page)

Using Earth Mover's Distance in the Bag-of-Visual-Words Model for Mathematical Symbol Retrieval

Simone Marinai, Beatrice Miotti, and Giovanni Soda

Dipartimento di Sistemi e Informatica
Università di Firenze, Italy
simone.marinai@unifi.it

Abstract—In this paper, the Earth Mover's Distance (EMD) is used as a similarity measure in the mathematical symbol retrieval task. The approach is based on the Bag-of-Visual-Words model. In our case the features extracted from each symbol are clustered by means of Self-Organizing Maps (SOM) and then occurrences of features in the clusters are accumulated in a vector of visual words. The comparison between the latter vectors is performed with the EMD which naturally allows to incorporate the topological organization of SOM clusters in the distance computation.

The proposed approach is experimentally tested in a mathematical symbol retrieval task and compared with the cosine similarity and with some variants that have been recently proposed.

Keywords-Bag of Visual Words; Earth Mover's Distance; Self Organizing Map;

I. Introduction

The digitization of collections in libraries has been widely addressed in the last years. Different types of documents call for different approaches. For instance printed documents can be processed by Optical Character Recognition (OCR) tools to be indexed on the basis of their textual content. Scientific and technical documents are particularly difficult to handle and several tools, designed to recognize the mathematical expressions in printed documents, have been proposed. In [1] the limits of commercial OCR tools for the recognition of mathematical expressions are discussed. OCR engines recognize the textual parts with a great accuracy, but are less accurate with mathematical expressions. It is therefore appropriate to search for mathematical expressions by using document image retrieval approaches based on the retrieval of mathematical symbols [2].

In this paper, we propose to use the Earth Mover's Distance to compute the similarity of mathematical symbols represented in the bag of visual words paradigm. The symbols are described with Shape Context (SC) that are clustered to reduce the number of different SC to compare. Symbols are then indexed by computing the occurrences of SCs in each cluster. In our approach the clustering is performed with the Self-Organizing Map (SOM) that spatially organizes the cluster centers to reflect their similarity in the input space.

The SOM properties are exploited in the retrieval in order to allow an inexact match between symbols that are described with similar, but not identical, SCs. This is achieved by using the Earth Mover's Distance to compute the similarity between distributions of SCs assigned to the symbols.

The paper is organized as follows. In Section II we summarize the overall approach adopted for mathematical symbol retrieval. The similarity measures considered are described in Section III and the Earth Mover's Distance is discussed in Section IV. The experimental results are summarized in Section V while the conclusions are drawn in Section VI.

II. Mathematical Symbol Indexing

Referring to Fig. 1, the first step of both symbol indexing and retrieval is the features extraction, where symbols are represented by Shape Context (SC) descriptors [3]. Let P be the set of contour points of a symbol. If the symbol is composed by more than one connected component then P contains the contours of all the components. The SC for each point p_i in P is computed by considering the relative position of the other points in P that are accumulated in a coarse histogram h_i whose bins are uniform in log-polar space. Let m be the cardinality of P and p_j be one of the remaining m–1 points in P. The point p_j is assigned to one bin according to the logarithm of the Euclidean distance between p_i and p_j and to the direction of the link between p_i and p_j. The histogram h_i is defined to be the Shape Context of p_i.

The m Shape Context vectors describe the whole symbol. Shape contexts are invariant to translations and can be modified to be scale and rotation invariant [3]. In case of mathematical symbols, the rotation can be misleading, e.g. confusing ∪ and ∩. To deal with small symbols the SC are computed by counting all the points belonging to each bin instead of considering only the points in P [4].

The subsequent steps of the process (Fig. 1) are the clustering and the indexing. The comparison between the shape contexts can provide a very accurate evaluation of the similarity between symbols. However, the computational
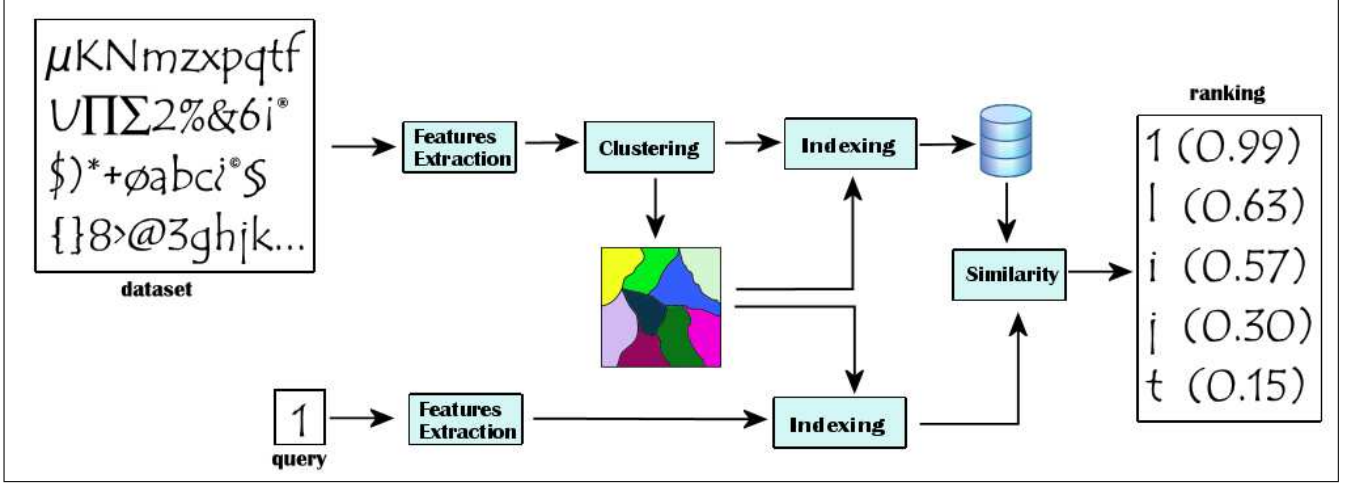
Figure 1: Schema of the mathematical symbol retrieval process. Indexed symbols are represented with Shape Contexts that are then clustered with an SOM map. The indexing is made by computing the occurrences of SCs in the SOM units.

cost of a pairwise comparison is excessive and cannot be considered when dealing with large data-sets. One solution of this problem relies on a symbol representation that exploits techniques adopted in the vector space model of Information Retrieval. One vector quantization is first performed by clustering (in most cases with K-means) the SCs and then labeling each SC with the index of the cluster it belongs to. Following the textual analogy, each cluster is considered as a "visual-word" and each symbol is represented on the basis of the frequencies of each "visual-word" in its description [5].

In this work we use the Self Organizing Map (SOM) to perform the vector quantization [2]. In contrast with K-means, the SOM clusters are topologically ordered and more similar shape contexts are mapped on close neurons in the map. The symbol S_i is then represented as a vector V_i of size C (the number of SOM units) whose elements contain the occurrences of SCs in each cluster. According to the vector space model, the indexed dataset is often normalized with the $tf \cdot idf$ weighting schema.

In Fig. 2 the indexing process is described starting from the phase of symbol sampling and feature extraction, and then showing how the visual dictionary, in this case the SOM map, is used to assign each shape context to the cluster it belongs to. From the portion of SOM shown on the right we can notice that similar clusters are closer in the two dimensional map. Each cluster is represented with its centroid SC that is graphically depicted as a circular mask where we fill the bins in the Shape Context with a gray level that is darker for higher values in the corresponding SC bins.

Each cluster can be considered as one "visual-word" in the visual dictionary and SCs in a symbol are labeled according to this dictionary. Neighboring SOM units in the figure are in general more similar, with smooth variations between close units. For instance, centroids in the last row represent

SCs that correspond to lower corners (e.g. occurring in the bottom of a 'V'). Moving from left to right we can notice a smooth rotation of the represented corner.

III. SIMILARITY

The similarity between objects represented with the vector space model is often computed with the cosine of the angle between vectors (e.g. [6]). Alternative approaches have been proposed for instance in [7] where the similarity is computed by means of the Chi-square distance. The cosine similarity is defined by:

$$sim(V_q, V_i) = \frac{\sum\limits_{j=1,C} v_{j,q} \cdot v_{j,i}}{|\vec{V_q}| \cdot |\vec{V_i}|} \qquad (1)$$

where V_q and V_i are the query and a generic indexed vector, respectively. One limit of this strategy comes out when very similar SCs are mapped on similar, but different, clusters.

To take advantage of the SOM topology we recently proposed some changes to this similarity [2] [4] that are shortly summarized in the following. The basic idea is to allow a partial match for elements $(v_{j,q})$ of V_q that have no counterpart in V_i. In this case we look for elements in V_i that correspond to neighbors of $V_{j,i}$ in the SOM. The modified similarity *sim'* can be expressed by:

$$sim'(V_q, V_i) = sim(V_q, V_i) +$$
$$+ \frac{\sum\limits_{(j|v_{j,i}=0)} v_{j,q} \cdot \dfrac{F(v_{j,i})}{\alpha}}{|\vec{V_q}| \cdot |\vec{V_i}|} \qquad (2)$$
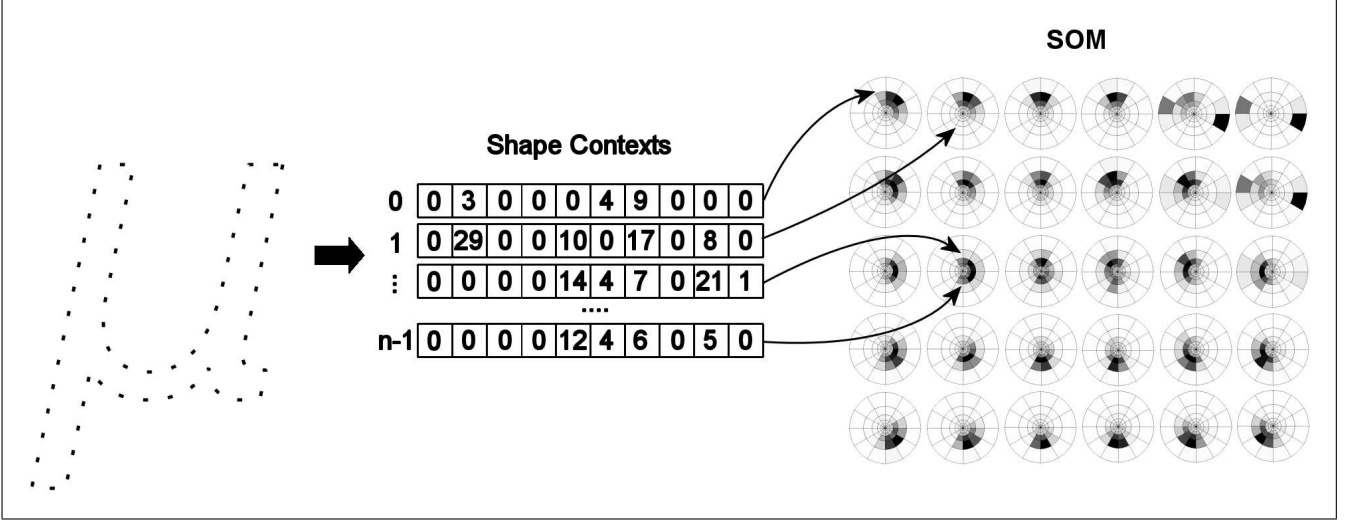
Figure 2: Example of the vector quantization process. On the left some points in P are shown. For each of the n points a Shape Context is computed and then assigned to one SOM unit. In the SOM, closest units correspond to most similar SCs.

where α is an empirical parameter and $F(\cdot)$ is a function that returns one unit $v_{w,i}$ among the four (or eight) nearest neighbors of $v_{j,i}$. This addend of sim' is considered for each element j such that $v_{j,q} \neq 0$ and $v_{j,i} = 0$. There are four variants of sim'. In the similarity called sim_{S4} (sim_{S8}) we analyze the 4 (8) neighbors of $v_{j,i}$. $F(v_{j,i})$ returns the value $v_{w,i}$ of the unit w in the SOM neighborhood of unit j with the highest $v_{w,i}$. In this case $\alpha = 4$ (or $\alpha = 8$).

The similarity called sim_{E4} is computed analyzing the 4 neighbors of $v_{j,i}$, but in this case $F(v_{j,i})$ returns the value $v_{w,i}$ of the unit w in the SOM neighborhood of unit j with the minimum distance in the $\mathbb{R}^C$ vectorial space with respect to $v_{j,i}$. An extension to the eight neighbors, called sim_{E8} has been considered as well. In the latter two similarities we have $\alpha = 4$.

IV. Earth Mover's Distance

In this work, we propose to use the EMD to compute the similarity for SOM-based clustering. This distance is compared to the cosine similarity and to its variants explained in detail in [8].

The Earth Mover's Distance (EMD) is a method to evaluate the dissimilarity between two multi-dimensional distributions which are often used in computer vision to summarize different image features [9]. In image retrieval an image can be represented, for instance, by the distribution of pixel intensities. These distributions can be summarized with clustering algorithms, which reduce the feature space in a fixed number of bins. Each cluster c_j is associated with a weight w_j that indicates the size of the cluster (e.g. the occurrences of features in each cluster). The EMD describes the cost that must be paid to transform one distribution

(considered as a mass of earth spread in space) into the other (considered as a collection of holes in the same space). The EMD measures the least amount of work needed to fill the holes with earth, considering that a unit of work corresponds to transporting a unit of earth by a unit of distance (*ground distance*). The EMD evaluation is based on the solution of the transportation problem which consists in finding the least expensive flow from one distribution to another according to some constraints.

In a more formal way we can express the transportation problem as follows [10]. Let $S = \{w_{s_1}, \ldots, w_{s_m}\}$ be the first distribution with m elements s_i ; $Q = \{w_{q_1}, \ldots w_{q_n}\}$ be the second distribution with n elements; and $D = [d_{ij}]$ be the ground distance matrix where d_{ij} is the distance between the element s_i and q_j.

The flow $\mathbf{F}$ that minimizes the overall cost is computed by Eq. (3) where f_{ij} is the flow between s_i and q_j.

$$\mathbf{F} = \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} \cdot d_{ij} \tag{3}$$

and it is subjected to the following constraints:

$$\begin{cases} f_{ij} \geqslant 0 & \text{for } 1 \leqslant i \leqslant m, \ 1 \leqslant j \leqslant n \\ \sum_{j=1}^{n} f_{ij} \leqslant w_{s_i} & \text{for } 1 \leqslant i \leqslant m \\ \sum_{i=1}^{m} f_{ij} \leqslant w_{q_j} & \text{for } 1 \leqslant j \leqslant n \\ \sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} = \min(\sum_{i=1}^{m} w_{s_i} \cdot \sum_{j=1}^{n} w_{q_j}) \end{cases} \tag{4}$$

The first constraint indicates that the items can be moved only from S to Q and not vice versa. The next two constraints are related to the amount of mass which can be sent from the elements in S (it must not exceed the weight values) and to the amount which can be received by elements in Q (again limited by the weights). The last constraint forces to move the maximum amount of mass as possible. After solving the transportation problem and computing the total flow **F**, the EMD is defined as the work normalized by the total flow:

$$\textbf{EMD}(S, Q) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij} \cdot d_{ij}}{\sum_{i=1}^{m} \sum_{j=1}^{n} f_{ij}} \qquad (5)$$

The EMD is a robust method to compare multidimensional distributions of features. It is a true metric if the ground distance is metric and if the total weights of the two signatures are equals.

The choice of the ground distance depends on the problem. For instance, in [9], Rubner et al. propose to use the Euclidean distance in the CIE Lab color space as measure of similarity among colors, and the Euclidean distance in the log-polar space to compare texture features in the frequency domain.

In this work we use the EMD distance to measure the similarity between the vector representations of mathematical symbol images that are obtained by SOM clustering of SC. The ground distance is based on the distance between SCs and is computed on the SOM map to take into account the SOM topology. The occurrences of SCs in each vector represent the weight as described in Eq. (5).

In our experiments we considered three ground distances to be applied on the map. Let (x_i, y_i) and (x_j, y_j) be the coordinates of the centroids c_i and c_j in $\mathbb{R}^2$ in the SOM. We considered:

- The Euclidean distance (L_2):

$$d(i, j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \qquad (6)$$

- The squared Euclidean distance (L_{22}):

$$d(i, j) = (x_i - x_j)^2 + (y_i - y_j)^2 \qquad (7)$$

- The Euclidean distance between centroids c_i and c_j in $\mathbb{R}^C$ (L_{2E}) where C is the dimension of the feature

vectors:

$$d_{i,j} = \sqrt{\sum_{k=0}^{C} |c_{i_k} - c_{j_k}|^2} \qquad (8)$$

To bound the influence of farthest units, when $d(i, j) > D_{max}$ we set $d(i, j) = \overline{D}$ where $\overline{D}$ is an high value (e.g. $\overline{D} = 100$).

V. EXPERIMENTS

In our experiments, we used the Infty-CDB3 dataset gathered in the context of the Infty Project [11]. Infty-CDB3 is a collection of isolated alphanumeric characters and mathematical symbols, splitted into two data sets that contain a total of $259, 389$ patterns.

Before indexing the data, we computed the SC clusters on a set of $22, 923$ symbols, belonging to 53 pages randomly selected from the whole dataset. From each symbol we extracted around 50 SCs so that we used a total of $1, 102, 049$ feature vectors for clustering. We then indexed all the $259, 357$ symbols in the two datasets (very small symbols have been removed from the collection). In some preliminary tests we compared alternative approaches that can be used to index the data and build the SOM [4]. In the tests discussed in this paper, we used, without re-training, the SOM obtained from these experiments [4].

The experiments have been carried out to compare the results obtained with the EMD distance, applied to the SOM map, with the cosine similarity and its modified versions. The retrieval has been performed on 392 random queries. The Precision at 0% Recall (that is obtained by interpolation of the P-R plot) and the area under the P-R plot are reported in Table I for various cases. sim is the cosine similarity, sim_{E4} is the best method among those not using the EMD (Section III). For the EMD-based similarity, the three ground distances have been tested: L_2 is the Euclidean distance, L_{22} is the squared Euclidean distance, L_{2E} is the Euclidean distance between centroids. In addition, some experiments have been performed with the $tf \cdot idf$ normalization (TF) and some without this normalization (NTF). In the results summarized in the table, the notation $\overline{D}$ indicates the use of the upper bound to the distance ($D_{max} = 5$).

Table I: Precision at 0 % Recall and Area under the curve (AUC) of Precision - Recall. See Table III for a summary of acronyms.

Similarity	PR 0	AUC
sim	97.86	2338.14
sim_{E4}	97.97	2362.45
L_{2_TF}	97.25	2375.18
L_{2E_TF}	98.11	3297.85
L_{2E_NTF}	**98.23**	**3313.17**
$L_{22\_\overline{D}\_NTF}$	95.58	2566.12

By analyzing Table I we can notice that the best results are obtained with L_{2E} both with (L_{2E_TF}) and without (L_{2E_NTF}) the $tf \cdot idf$ normalization. In all the cases the AUC values are larger when using EMD with respect to the other similarity measures. Considering the best EMD case the AUC value has been improved of about the 40 % with respect to the best results obtained with the other similarities.

To reduce the computational time with the EMD similarity it is possible to consider a two-step retrieval. In the first step the standard cosine similarity is considered. The top 100 elements of the ranked list are then re-ordered with a refinement step that uses the EMD similarity. The results of this experiment are shown in Table II. Even if it is not possible to equalize the best result of Table I (obtained using only the EMD similarity) we can improve in any cases the results obtained with the cosine similarity. In particular the results obtained with the cosine similarity sim_{E4} are improved of about the 2.16 %.

Table II: Results of the two phases retrieval. See Table III for a summary of acronyms.

Step 1	Step 2	PR 0	Area
sim	L_{2_TF}	97.40	2338.76
sim	$L_{22\_\overline{D}\_NTF}$	97.22	2342.43
sim_{E4}	L_{22_TF}	94.33	2293.70
sim_{E4}	L_{2_TF}	97.36	2358.62
sim_{E4}	L_{2E_TF}	98.10	**2421.08**

Table III: Acronyms used in the experiments.

Acronym	Similarity Measure
sim	Standard cosine similarity: Eq. 1.
sim_{E4}	Similarity computed with Eq. 2.
L_{2_TF}	EMD with Euclidean distance; $tf \cdot idf$ normalization.
L_{22_TF}	EMD with squared Euclidean distance; $tf \cdot idf$ normalization.
L_{2E_TF}	EMD with Euclidean distance between centroids; $tf \cdot idf$ normalization.
L_{2E_NTF}	EMD with Euclidean distance between centroids; no $tf \cdot idf$ normalization.
$L_{22\_\overline{D}\_NTF}$	EMD with bounded squared Euclidean distance; no $tf \cdot idf$ normalization.

VI. Conclusions

In this paper, we described the integration of the Earth Mover's Distance with a bag of visual word representation of symbols based on SOM clustering. This procedure is applied to a problem of mathematical symbol retrieval. With the proposed approach we can take advantage of the topological organization of SOM clusters and find a correspondence between similar but not identical visual words in different indexed symbols. The distance between SOM clusters is used as ground distance in the EMD algorithm.

We are now testing the EMD-based approach on other domains such as writer identification with promising results that confirm the findings reported in this paper.

References

[1] M. Suzuki, F. Tamari, R. Fukuda, S. Uchida, and T. Kanahori, "Infty: an integrated OCR system for mathematical documents," in *ACM Symposium DocEng*. New York, USA: ACM, 2003, pp. 95–104.

[2] S. Marinai, B. Miotti, and G. Soda, "Mathematical symbol indexing using topologically ordered clusters of shape contexts," in *Proc. Int.l Conference on Document Analysis and Recognition*, 2009, pp. 1041–1045.

[3] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, Apr 2002.

[4] S. Marinai, B. Miotti, and G. Soda, "Mathematical symbol indexing," in *Proc. Int'l Conf. on of the Italian Association for Artificial Intelligence*. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 102–111.

[5] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. Int'l Workshop on Multimedia Information Retrieval*. New York, USA: ACM, 2007, pp. 197–206.

[6] T.-O. Nguyen, S. Tabbone, and O. R. Terrades, "Symbol descriptor based on shape context and vector model of information retrieval," in *Proc. IAPR Int'l Workshop on Document Analysis Systems*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 191–197.

[7] G. Tan, C. Viard-Gaudin, and A. Kot, "Information retrieval model for online handwritten script identification," in *Proc. Int.l Conference on Document Analysis and Recognition*, 2009, pp. 336–340.

[8] S. Marinai, B. Miotti, and G. Soda, "Mathematical symbol indexing for digital libraries," in *Digital Libraries*, ser. Communications in Computer and Information Science. Springer, 2010, vol. 91, pp. 113–124.

[9] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *ICCV*, 1998, pp. 59–66.

[10] ——, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, pp. 99–121, 2000.

[11] M. Suzuki, S. Uchida, and A. Nomura, "A ground-truthed mathematical character and symbol image database," in *Proc. Int.l Conference on Document Analysis and Recognition*, 2005, pp. 675 – 679 Vol. 2.