Mining Visual Data

Nistor GROZAVU Nicoleta ROGOVSCHI
LIPN-UMR 7030, Paris 13 University,
99, av. J-B Clément, 93430 Villetaneuse, France
email: Nistor.Grozavu@lipn.univ-paris13.fr email: Nicoleta.Rogovschi@lipn.univ-paris13.fr

***Abstract* — The Internet offers to its users an ever-increasing number of information. Among those, the multimodal data (images, text, video, sound) are widely requested by users, and there is a strong need for effective ways to process and to manage it, respectively. Most of existed algorithms/frameworks are doing only images annotations and the search is doing by these annotations, or combined with some clustering results, but they do not allow a rapid browse of these images. In this paper, an image retrieval system is presented, including detailed descriptions of used lwo-SOM approach and a novel interactive learning using user information/response. Also, we show the use of unsupervised learning for images, we do not dispose of the labels, and we will not take into account the corresponding text for the images. The used DataSet contains 17812 images extracted from wikipedia pages, each of which is described by it colors and texture.**

(Some figures in this paper are in color only in the electronic version)

***Index Terms* — Clustering, Visual Data, Self-Organizing Maps, weighting.**

I. INTRODUCTION

The multimodal data has properties which made difficult its exploitation using statistical classical methods:

- Heterogeneous data : some physical image characterization are associated to some words extracted from a text jointed to the image, or sound and image which came from a video, or sound and text derived from an audio record;
- The data are usually mixed : one can have at the same time quantitative data (images characteristics/ features), and binary data (the words which correspond to the images);
- The number of features to describe this data is usually very important/big and can be in order of some thousands features;

Producing visual data/content in digital form, even the visualization of the numerical data is becoming more and more common and affordable.

Images DataSets are becoming more common and widely used as visual information is produced at a rapidly growing rate. Creating images and storing them became an easily and very used process for general use.

Consequently, the digital visual libraries are growing and there is a strong need of adequate solutions to process this data and to extract relevant information from it.

The traditional text-based approaches to image retrieval have proven out to be inadequate for many purposes. In some occasions, image databases have associated captions or other text describing the image content and these annotations can be used to greatly assist image search. Manually annotating large databases takes, however, a lot of effort and raises the possibility of different interpretations of the image content. As a result, content-based image retrieval (CBIR) has received considerable research and commercial interest in the recent years. One of the challenge is to automate the process of image retrieval and to make it separately from text annotation [5].

One of the most interesting and used technique for data reduction and visualization in machine learning are the Self-Organizing Maps (SOM) proposed by Kohonen in 1998. This approach was used for image retrieval system called PicSOM [5] which use the tree structured SOM (TS-SOM) [4].

In this work we propose a novel technique which propose to use the *lwo*-SOM [1] to attempt a 3D visualization and browsing of the dataset. Also, we incorporate an interactive learning approach based on the users/experts information and on the computing of the Euclidian distances matrix. This technique is similar to the annotation which was used in different way, like in [2] which compute mixture models for each image and uses the Mallows distance to construct a matrix to be used by the clustering algorithms.

Contrary to the most of the feature extraction techniques where the main access to the images is made through query, we will use an autonomous approach which auto-organize the structure of the dataset using the learned map.

The system is also capable to receive new data after the clustering and to place it in the computed space in the map.

In this paper, we will study a specific image collection of 17812 images extracted from Wikipedia pages. The images are accompanied with keyword-type annotations which specify a subset of available keywords for each image.

The problem of clustering and weighting constitute an important part of the design of good learning algorithms. The generalization performance of these algorithms can be significantly degraded if irrelevant variables are used. This negative effect increases in the case of unsupervised learning where no class labels are given. In this case, the problem is that not all variables are important. Some of the variables may be redundant, some may be irrelevant, and

some can even degrade clustering results. Our purpose is to weight the most important features in order to allow a better organization of the map and, if needed to detect/select the relevant features. Continuous weighting provides a richer feature relevance representation. Hence, it is clear that the clustering and variable weighting task are coupled, and applying these tasks in sequence can degrade the performance of the learning system. Consequently, it is necessary to develop a simultaneous algorithm of clustering and variables weighting. The models that interest us in this paper are those that could make at the same time the dimensionality reduction and clustering using Self-Organizing Maps (SOM, [3]) in order to perform the information extraction. SOM models are often used for visualization and unsupervised topological clustering. Its allow projection in small spaces that are generally two dimensional.

The rest of this paper is organized as follows: We show in section 3 the used local weighting approach *lwo*-SOM based on classical SOM after the presentation of the used technique for images feature extraction in Section 2. In section 4 we describe our proposed framework for images clustering and browsing, and we show the results using this technique on the wikipedia images. Finally we offer some concluding comments of proposed method and the further research.

II. Images and Features extraction

The images (17812) were extracted from the wikipedia web site from the tourism compartments by the Xerox Research Center [6] during the Infom@gic project. Each image has a web link and a set of words attached to it.

The Fisher Kernels approach was used to obtain a numerical transformation of images.

The used technique is an approach which was inspired by the bag-of-words used in text categorization referred to as the bag-of-keypatches or bag-of-visterms (BOV). Given a visual vocabulary, the idea is to characterize an image with the number of occurrences of each visual word.

The gradient of the log-likelihood transforms a variable length sample X into a fixed length vector whose size is only dependent on the number of parameters.

Perronnin F. and Dance C. proposed to apply Fisher kernels on visual vocabularies, where the vocabularies of visual words are represented by means of a GMM (Gaussian Mixture Models).

All the images were resized to contain approximately the same number of pixels, so, the same number of features was extracted from all images (between 500 and 600 for each feature type). The first features are based on local histograms of orientations; the second ones are simple local RGB statistics [6].

After obtaining of these features vectors, the features dimension were reduced to 6400 using an GMM algorithm and the Fisher Kernel described bellow.

Fisher kernels have been introduced to combine the benefits of generative and discriminative approaches. Let p be a pdf whose parameters are denoted λ. Then one can characterize the samples $X = x_t, t = 1...T$ with the following gradient vector:

$$\nabla_\lambda \log p(X|\lambda)$$

We note that in this work we used the numerical data for each image; the transformation was made by the Xerox Research Center.

III. Process and manage the visual data

For the Framework for images Clustering and Browsing we use the local weighting Self-Organizing Map *lwo*-SOM [1] which use a weighting technique to weight the observation x with the weight vector π before computing the Euclidian distance. In this case the SOM cost function is rewritten as follows:

$$R_{lwo}(\chi, W, \Pi) = \sum_{i=1}^{N} \sum_{j=1}^{|W|} K_{j,\chi(x_i)} \left\| \pi_j x_i - w_j \right\|^2 \qquad (1)$$

The minimization of $R_{lwo}(\chi, W, \Pi)$ is done by iteratively repeating the following three steps until stabilization. After the initialization step of prototype set W and the associated weights set Π, at each training step *(t+1)*, an observation x_i is randomly chosen from the input data set and the following operations are repeated:

- Minimize $R_{lwo}(\chi, \hat{W}, \hat{\Pi})$ with respect to χ by fixing W and Π. Each weighted observation $(\pi_j x_i)$ is assigned to the closest prototype w_j using the assignment function defined as follows:

$$\chi(x_i) = \arg\min_{i \le j \le |w|} \left(\left\| \pi_j x_i - w_j \right\|^2 \right)$$

- Minimize $R_{lwo}(\hat{\chi}, W, \hat{\Pi})$ with respect to W by fixing χ and Π. The prototype vectors are updated using the gradient stochastic expression:

$$w_j(t+1) = w_j(t) + \varepsilon(t) K_{j,\chi(x_i)} (\pi_j x_i - w_j(t))$$

Minimize $R_{lwo}(\hat{\chi}, \hat{W}, \Pi)$ with respect to Π by fixing χ and W.

The update rule for the weight vector $\pi_j(t+1)$ is:

$$\pi_j(t+1) = \pi_j(t) + \varepsilon(t) K_{j,\chi(x_i)} x_i (\pi_j(t) x_i - w_j(t))$$

As in the traditional Kohonen's stochastic learning algorithm, we denote by $\varepsilon(t)$ the learning rate at time t. The training is usually performed in two phases. In the first phase, a large initials learning rate $\varepsilon(0)$ and a large neighborhood radius $T_{\max}$. In the second phase both learning rate and neighborhood are small right from the beginning.

IV. The Framework for Images Self Organizing Map

After obtaining the *lwo*-SOM map, we construct the distance matrix NxC, computing the Euclidian distance

between each sample (image) $i=1...N$ and each prototype $j=1...C$ of the map:

$$\chi(D_m)=\left(\left\|\pi_j x_i - w_j\right\|^2\right)$$

The matrix D_m is sorted in order to have all the samples (images) structured by levels: from the best matching unit until the last unit. Using this sorted matrix, we can construct now the map and visualize images which correspond to these units. For each unit where is a corresponding image from the data set. Each map unit/cell has several images/samples/observations which were captured during the learning process. So we have two ways to manage this images dataset: to cluster and to browse it, presented in the next sections.

1. Clustering the images

Firstly, we learn a map size 13x13 (169 cells), and secondly, for each cell we display the corresponding image from the dataset. Also, the framework will give the possibility that when choosing the interest image, the map will show all others images which were captured by this cell/neuron (Figure 1).

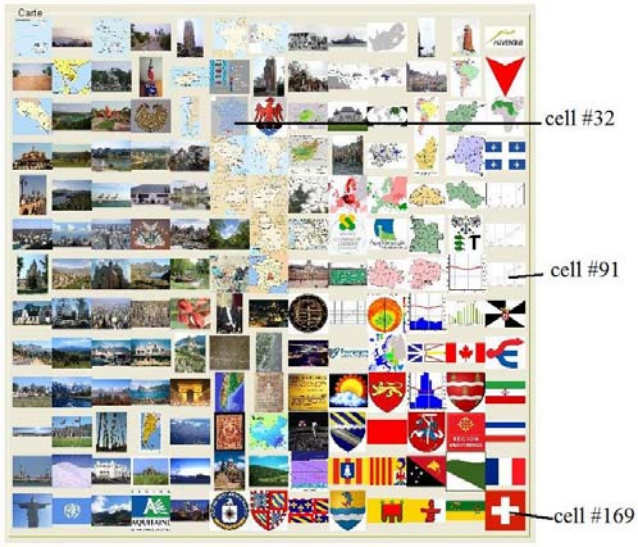

Figure1. Images SOM Map (13x13 size)

On the obtained map (Figure 1), we can detect 6 clusters: one in the top left corner which correspond to the clear blue images (like images which sky); a cluster in the left bottom part of the map - images with the more darkly blue color. We observe that these two clusters are neighborhoods on the map because they are correlated (clear blue and dark blue). Another correlated cluster to these two is the cluster situated on the bottom which contains images with blue/black color. On the right bottom corner of the map there is a cluster which contains red/yellow images represented the flags. The right top cluster has more white images representing graphs and geographical maps; and the last cluster is situated in the middle of the map and is representing by brown color images. Like wee can see, the topolgy of the map is well defined and the neighborhoods cells on the map are correlated between them. This technique gives the possibility to the user to have a small-space representation for the entire dataset.

Now, we will analyze the captured images by the cells 169, 91 and 32 like we can see on the figure 1. For the cell 169 (Figure 2) we can observe that all the captured images are reds and represents the flags (only one image are green). The 91th cell captured all white images representing the graphs presented in the figure 3.

Figure 2. Captured images by the 169th cell

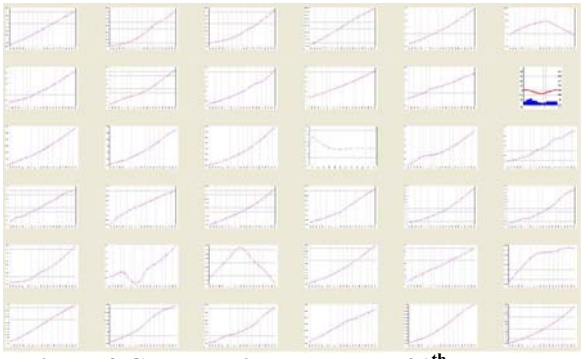

Figure 3 Captured images by the 91th cell

More difficult is to find the correlated images for the 32th cell there not all the images are highly correlated, this means that the expert annotation could not coincide with our result. In this case, the system (framework) must be able to use the user/expert information and to change its results after the learning process - the interactive learning.

Figure 4. Captured images by the cell 32

2. Browsing

The second scenario is to browse the images data set by levels. Firstly, we visualize the map with the best matching units (the most representative images) and then, we can chose the next level to visualize (or to skip some levels) until we are satisfied of the result. This process is doing in a 3D visualization by displaying the maps with the

corresponding captured images step by step like shown in the figure 5.

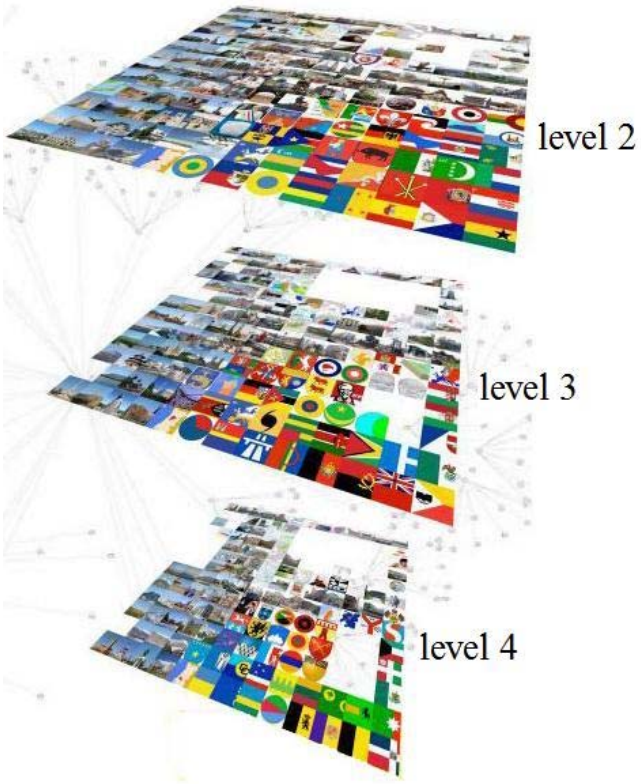

Figure 5. Images DataSet browsing using *lwo*-SOM technique.

Most of the image retrieval systems do not support browsing, likely because it is difficult to define and implement. Rather, these systems force/ask the user to specify what they are looking for with a query. If the user's task is not compatible with the images annotations made by another users/experts, the system will not be able to help the user learn what kind of images can be found. So, the problem, in this case is the text annotations, and the no-organizations between images during this task. Our purpose is to automate the browsing task using not only the annotated text, but also the similar images detected during the unsupervised learning. The idea, is to present a images map to the user in order to detect not only the searched image, but also the similar images from the map (neighboring cells using the Euclidian distance) (Figure 5). Furthermore, a cell from the map (the best matching unit) can be used to represent many others similar pictures, and will accurately suggest the kinds of pictures that will be found by exploring that cluster.

The figure 5 shows the map with the best matching units (first level), and the next 3 levels of the maps. For each map the neighborhoods displayed images are correlated between them, and one can detect also some cells which are empty, because there are cells which captured only 1, 2, or 3 images. So displaying the map level which is greater then the size of the captured images vector for a cell, the respective cell will display an empty (white) image to show that where are not more correlated images to the last one.

3. Interactive Learning

As already stated, sometimes we can have images which don't correspond to the user/expert criteria, and, in this case, the framework gives the possibility to the user to choose the class/cell where the image should be placed. Our system will compute the minimum Euclidian distance between the image *i* and the corresponding cell/prototype *j*. Then, the new image is placed in the respective cell and to find its place in the cluster *j* the corresponding vector of images is re-sorted. This process is useful also to cluster the new images which arrive in a real-time to the dataset/framework by computing the corresponding distance: firstly between it and all the prototypes/cells, and secondly - to compute the distance inside of the cell between it and all the images to find its place.

4. Search

A second important facility for the images libraries is the images retrieval based on user queries. In literature exists many algorithms to resolve this problem, like computing the probability of each candidate image of emitting the query items, computing the images rank, and many others. Our goal is to use the jointed images words, but instead of displaying all the images which has this word jointed, our system displays to the user the level of the map where is situated the searching image. With this, the user will have the possibility to choose another correlated image even if it doesn't has the searching word jointed to it.

V. CONCLUSION

In this paper we presented a novel solution for manage and process visual datasets. We used the *lwo*-SOM [1] which allows us to do a better classification of the data and to obtain more correlated images on the map. We propose two scenarios: a clustering and a browsing schema which could be done simultaneously with the interactive learning. Also, we propose an original solution for the searching on the images libraries using the annotated text only to find the corresponding level of the map, and then to use the correlation between images on the map and inside the cell to display the information, in order to avoid the eventual noisy (worst annotated text).

ACKNOWLEDGMENTS

This work was supported by Cap Digital under the Infom@gic Project.

REFERENCES

[1] N. Grozavu, Y. Bennani, M. Lebbah. From variable weighting to cluster characterization in topographic unsupervised learning. IJCNN, Atlanta, USA, 2009.
[2] C. Julien, and L. Saitta. Image databases browsing by unsupervised learning. ISMIS, 2008.
[3] T. Kohonen, Self-Organizing Maps. Springer Berlin, 2001.
[4] P. Koikkalainen. Progress with the tree-structured self-organizing map. In Proc. 11th Europ. Conf. Artificial Intell., 1994.
[5] M. Koskela. Interactive image retrieval using self-organizing maps. Dissertation Repport, 2003.
[6] F. Perronin and C. Dance. Fisher kernels on visual vocabularies for image categorization. page 1-8, 2007.