Towards an Images Dataset Processing trough Supervised and Unsupervised Learning

Nicoleta ROGOVSCHI Nistor GROZAVU
LIPADE, Paris-Descartes University, France
nicoleta.rogovschi@parisdescartes.fr
LIPN, Paris 13 University, France
nistor.grozavu@lipn.univ-paris13.fr

***Abstract* – 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 most of them do not allow a quick browsing of these images. Even if the search is very quickly, but if the number of images is very large, the system must give the possibility to the user to browse this data. In this paper we investigate the use of the supervised learning to classify an images dataset and the unsupervised learning to browse the images. In our proposed schema, we used both PCA and LDA to transform the feature space and then to classify the dataset. We used this technique for all five datasets available on the challenge web site of The German Traffic Sign Recognition Benchmark: HOG1, HOG2, HOG3, HueHIst and Haar [7]. Finnaly we used a voting approach to find the consensus for all five partitions. Also, an application to the images browsing is shown using the topological unsupervised learning.**

***Index Terms* – content-based image retrieval, topological learning, clustering, self-organizing maps.**

I. INTRODUCTION

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 German Traffic Sign Recognition Benchmark competition task [7] is a multi-class classification problem. The dataset consists of 39209 images where 26640 are for training and 1569 images are for the test.

Five pre-calculated features sets were available for the Challenge: three sets of HOG features, Haar-like features and Hue Histograms having the size:

- HOG1: 1568 features;
- HOG2: 1568 features;
- HOG3: 2916 features;
- HueHist: 256 features;
- Haar: 11584 features;

The first phase to do when deal with large dataset is to transform the features space and to detect the irrelevant variables.

The second step is to apply a classification approach on the new dataset to learn a model and to affect the test data.

Finally, the last phase is the fusion of all the results (classification of the all) in order to obtain a global classification result combining all five pre-calculated features.

We tested several supervised learning approaches to obtain high classification accuracy as: neural networks based methods, and feature transformation techniques as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).

We observed that we can increase the classification result if the feature space is transform using a principal component analysis technique.

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 challenges is to automate the process of image retrieval and to make it separately from text annotation [5].

One of the most interests 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 proposes to use the *lwo*-SOM [1] to attempt a 3D visualization and browsing of the dataset.

The rest of this paper is organized as follows: We show in section 2 the used feature transformation and dimensionality reduction approach. The supervised learning and the fusion technique used in the proposed method (section 5) are presented in sections 3 and 4. In section 5.A we describe the proposed unsupervised learning for images clustering and browsing, and we show the results using this technique on the Wikipedia images. Finally we offer some concluding comments of the proposed method and the further research.

II. FEATURES TRANSFORMATION AND DIMENSIONALITY REDUCTION

Principal component analysis (PCA) is a popular data processing and dimension reduction technique. As an un-supervised learning method, PCA has numerous applications such as handwritten classification, human face recognition, etc.

There is a strong link between the self-organizing maps (SOM) and PCA, as they have the same goal, i.e. to reduce the dimension and to visualize the dataset. This is why; we will use the both SOM and PCA as a pre-processing step for our model.

The PCA algorithm is presented as following:

Let the data **X** be a **n**x**m** matrix, where **n** and **m** are the number of observations and the number of variables, respectively.

The PCA estimation problem can be equivalently formulated as the following optimization problem, in which the sum of estimation errors from all variables is minimized:

$$R_{PCA}(\hat{\alpha}, \hat{z}_i) = \arg\min \sum_{i=1}^{N} (x_i - \widehat{x_i})^T (x_i - \hat{x}_i)$$

$$with \quad \hat{x}_i = \hat{\alpha}\hat{z}_i, \quad and \quad \widehat{\alpha^T}\hat{\alpha} = I$$

where xi and $\hat{x}_i$ are the i-th measured and estimated observation, and $\hat{z}_i$ represents the the estimated principal component corresponding to the observation xi.

In order to detect the number of eigenvalues values, we use the Cattell's Scree Test which is a graphical method first proposed by [8].

The basic idea of the Scree test is to generate, for a principal components analysis (PCA), a curve associated with eigenvalues, allowing random behavior to be identified (a simple line plot). Cattell suggests finding the place where the smooth decrease of eigenvalues appears to level off to the right of the plot. To the right of this point, presumably, one finds only "factorial scree". Non graphical solutions to the Cattell scree test are also proposed: an acceleration factor and the optimal coordinates index. The acceleration factor indicates where the elbow of the scree plot appears. It corresponds to the acceleration of the curve, i.e. the second derivative. Frequently this scree is appearing where the slope of the hill changes drastically to generate the scree. It is why many researches choose the criterion eigenvalue where the slope changes quickly to determine the number of components for a PCA. It is what Cattell named the elbow. So, they look for the place where the positive acceleration of the curve is at his maximum. Cattell's scree test and Bartlett's chi-square test for the number of factors to be retained from a factor analysis are shown to be based on the same rationale, with the former reflecting subject sampling variability, and the latter reflecting variable sampling variability. In the Cattell scree method, we can interpret the eigenvalues as the degree of relevance of each factor axis. The concept of covariance or correlation matrix is not appearing and is not necessary. Therefore, this method is not specific to PCA or a factorial analysis. The number of variables retained is equal to the number of values preceding this 'scree'. We therefore needed to identify the point of maximum deceleration in the curve.

Figure 1 shows an example of a curve generated using a data vector.

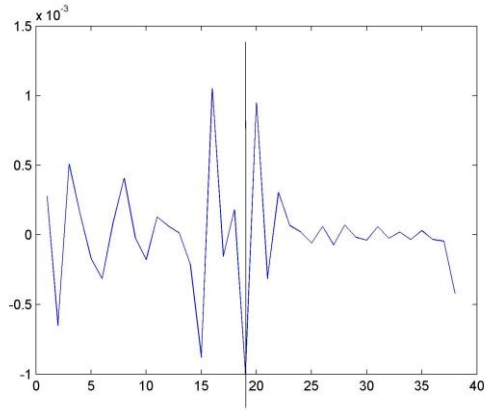

Figure 1: The Scree Test Acceleration Factor

We thus executed the following steps presented in the Algorithm 1.

Algorithm 1: The Scree Test Acceleration Factor

Input: a vector π_j size d

For $i = 1$ to d Sort the weights in descending order $\pi^{[j]}$.

Thus we obtain a new order $\pi^{[j]} = (\pi^{[j],1}, \pi^{[j],2}, \ldots, \pi^{[j],i}, \ldots, \pi^{[j],d})$; where i indicates the index order.

End for

For $j = 1$ to d (on the sorted vector)

Compute the first difference $df_i = \pi^{[j],i} - \pi^{[j],i+1}$ and we obtain the vector $\pi_{df1}^{[j]}$

End for

For $p = 1$ to d (on the $\pi_{df1}^{[j]}$ vector)

Compute the second difference (acceleration) $acc_i = df_i - df_{i+1}$ obtaining the vector $\pi_{df2}^{[j]}$

End for

For $l = 1$ to d (on the $\pi_{df2}^{[j]}$ vector)

Find the scree: $\max_i (abs(acc_i) + abs(acc_{i+1}))$

End for

OUTPUT:

Retain all the features displayed before the scree (we used the initial index values of features before sorting).

A. Complexity of the Scree Test procedure

The Scree Test acceleration procedure has four steps until finding the scree in the vector. We will analyze all these steps:

• Ascending sort: to made the sort of the weight vector we are using the Merge sort procedure which has an logarithmic complexity: $O(d \log d)$;

• First difference: the complexity for the first difference

df_i is the $O(d)$;

• Second difference: for the second difference the complexity is the same as previously: $O(d)$;

• Find the scree: to find the scree in a vector, the complexity will be $O(d)$.

As there is no nested loop, the total computational time for the Scree Test acceleration algorithm is the sum of the complexity of the four steps, and respectively it will be $O(d \log d + 3d)$.

III. CLASSIFICATION

As classification model we test the supervised Self-Organizing Maps and the Linear Discriminant Analysis, and we note that the use of the PCA improves the classification results. So, for all the dataset the LDA were used.

Given a dataset **X** size **n**x**m**, where **X = x₁, x₂, ..., xₙ** represents the set of object with m features.

Let, $B \in R^{nxp}$ be the transformation matrix that maps these features to p-dimensional features, i.e. $z_j \in R^p$ (j=1,...,m), and $z_j = B^T x_j$.

$$R_{LDA}(B) = \frac{|B^T \sum_b B|}{|B^T \sum_w B|}$$

where

$$\sum_w = \frac{1}{N} \sum_{k=1}^{c} \sum_{x_j \in D_k} (x_j - \mu_k)(x_j - \mu_k)^T$$

is the within-class covariance matrix, and

$$\sum_b = \sum_{k=1}^{c} P_k (\mu_k - \mu)(\mu_k - \mu)^T$$

is the between-class covariance matrix.

Indeed to use the initial dataset as input for the LDA method, we use the eigenvalues vectors issues from the PCA.

IV. FUSION

There are two types of combining classification (clustering) results: the fusion and the collaboration.

The goal of the fusion based techniques is to find a consensus for all the results using a fusion approach, as is the voting procedure. Contrarily, the collaborative classification is based on the changing the information during the learning process.

For this challenge, we tested the both types of methods, and we conclude that for these datasets, the better one is the fusion method.

As fusion method technique we use the voting principle.

V. PROPOSED METHOD

We introduce in this section the proposed methodology using the principal component analysis within the Cattel ScreeTest and Linear Discriminant Analysis for the classification. The method is used for all five datasets and the classification results (the labels vectors) are used to find the consensus by applying a voting technique.

Algorithm 2 : Proposed method

Input: images vectors vector $x_i \ldots x_n$

For $i = 1$ to n (for all datasets)

PCA with U 1100 eigenvectors on the train data

For j=1 to m (on the development dataset) :

Apply LDA on the U and obtain the model M;

End For

Plot the test data on the same features space using the U and obtaining Ut;

Affect the Ut to the model M;

End For

Output: Label of the test data;

We repeat the algorithm for all five datasets by computing the accuracy index for all of them.

At the end we use a fusion technique to fusion the classification results using a voting approach.

VI. EXPERIMENTAL RESULTS

Using the proposed method to all datasets, we obtain a classification accuracy index equals to 95.47 %, but we found that using only the HOG2 and HOG3 datasets, the accuracy index grow up to 96.53%.

Note that we classified the dataset using the LDA algorithm on the result of a PCA with 1100 eigenvectors.

A.Visualization

Even the topological learning methods doesn't improve the results for this challenge compared to the PCA and LDA, it allows the visualization of the classification results.

So, in this section we show an example of an extended SOM algorithm to classify and to browse a images dataset proposed by Rogovschi and Grozavu [9].

1)Images topological map browsing

The topological learning allows building a multi-level map which could be benefit to browse an images dataset by levels.

Firstly, we visualize the map with the best matching units (the most representative images) and then, we can choose the next level to visualize (or to skip some levels) until we are satisfied of the result. This process is doing in a 3D (hierarchical) visualization by displaying the maps with the corresponding captured images step by step like shown in the figure 1.

Our purpose is to automate the browsing task using not only the annotated text, but also the similar images founded during the unsupervised learning.

The idea is to present an images map to the user in order to detect not only the searched image, but also the similar images from the map (neighbored cells using the Euclidean distance). 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 the respective cluster.

The figure 1 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 no more correlated images to the last one.

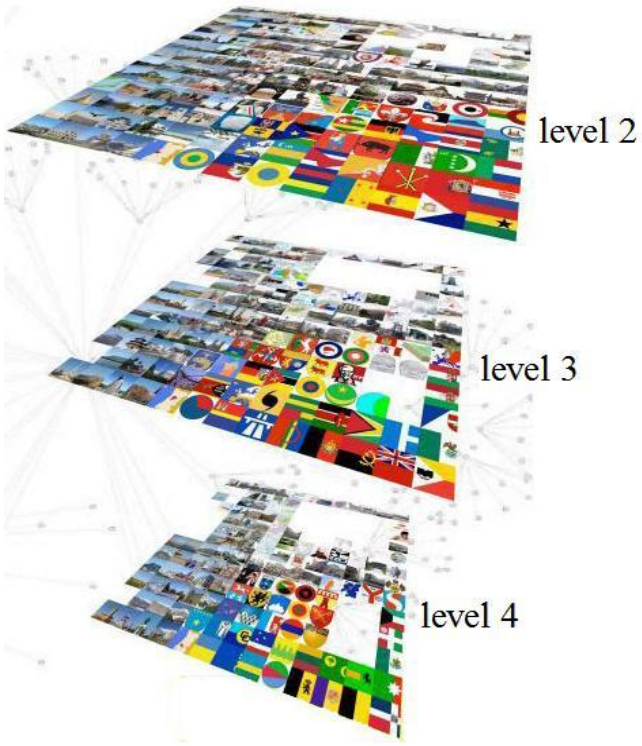

Figure 2. Images DataSet browsing using *lwo*-SOM technique.

VII. CONCLUSION

In this paper we adapted the supervised and unsupervised learning to deals with an images dataset. For the supervised learning we used the PCA and LDA algorithms coupled within a fusion approach. This new methodology was tested on the challenge of The German Traffic Sign Recognition Benchmark obtaining good results.

For the unsupervised learning, 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.

As future work, the fusion of both methods (to classify and to browse the images dataset) will be an interested challenge.

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.

[7] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In submitted to International Joint Conference on Neural Networks, 2011.

[8] R. Cattell. The scree test for the number of factors. Multivariate Behavioral Research, 1:245–276, 1966.

[9] Rogovschi N., GROZAVU N. (2010), « A content-based image retrieval system based on unsupervised topological learning», in Proc. ICMIA'10 : IEEE International Conference on Data Mining and Intelligent Information Technology Applications, November 30 - December 2, 2010, Seoul, Korea.