Journal of Software Engineering and Applications, **2012, 5, 200-203**
doi:10.4236/jsea.2012.512b038 Published Online December 2012 (http://www.SciRP.org/journal/jsea)

Image Classification using Statistical Learning Methods

Jassem Mtimet, Hamid Amiri

Signal, Image and Technology of Information Laboratory, National Engineering School of Tunis, Tunis El Manar University, BP 37, Le Belvdre 1002, Tunis, Tunisia.
Email: mtimat.jassem@yahoo.fr, hamidlamiri@gmail.com

Received 2012

ABSTRACT

In general, digital images can be classified into photographs, textual and mixed documents. This taxonomy is very useful in many applications, such as archiving task. However, there are no effective methods to perform this classification automatically. In this paper, we present a method for classifying and archiving document into the following semantic classes: photographs, textual and mixed documents. Our method is based on combining low-level image features, such as mean, Standard deviation, Skewness. Both the Decision Tree and Neuronal Network Classifiers are used for classification task.

Keywords: Image Classification; Decision Tree; Neuronal Network; Statistical Analysis

1. Introduction

Nowadays, a huge number of documents are available in electronic format, whether as photos, plans, letters or press releases. With the continuous increase of the amount of such information, many applications for organizing this flood of documents are emerging. Amongst them, automatic image archiving systems are necessary to classify and to store a large collection of documents autonomously, to simplify searching and retrieving individual documents.

Recently automatic semantic classification and archiving of images has become an important field of research, aiming to automatically classify images, i.e. classification of images into significant categories, such as outdoor/indoor, city/landscape and people/non-people scenes [1,2].

In order to classify images into two classes (indoor/outdoor, city/landscape, etc.) Vailaya et al. use a Bayesian framework and obtain an average accuracy of 94.1% [3].

In [4] Gorkani et al. suggest an image classification method based on the most dominant orientation in the image's texture. In fact, this feature allows differentiating two final classes of images: city and landscape. Thus, they achieve a classification accuracy of 92.8%.

Another approach was proposed by Prabhakar et al. in [5]. They used three low-level image descriptors (color, texture and edge information) to separate pictures and graphic images. Their algorithm reaches an accuracy rate of 96.6%.

In [6] Schettini et al. aim to classify images into four classes (photographs, graphics, text and mixed documents). Therefore, from every image, they extract six features which represent color descriptor, edge representation, texture features, wavelet coefficients and skin color pixels percentage.

This paper presents a system able to automatically classify and archiving documents into the following three categories: photos, textual documents and mixed documents.

In Section 2, theoretic background of our approach is explained. Then in section 3, the experience plan is described, including data sets, experimental results and evaluation criteria, while in Section 4, results are discussed and new perspectives are suggested.

2. Proposed System

The system we propose allows discriminating documents into photographs, textual and mixed documents. It is based on two main stages (**Figure 1**): i) The features extraction: These features are extracted automatically from images using specific programs. For every single image, the values of these features will be used as coefficients of a representative vector. ii) The classification and archiving module: This is obtained after training and validating a model used to discriminate and store documents.

2.1. Features Extraction

Features selection is the key step leading to the success or failure of the classification phase. Therefore, several

features are tested, looking to their relevance. In fact, features selection is an empiric process, though many approaches are suggested to weight their importance. In our system, images are classified based on six low-level featured, these features are considered as the coefficients of the image representative vector. They are calculated as follows:

- **Mean:** is the average color value in the image.

$$\mu_i = \frac{1}{N} \times \sum_{j=1}^{N} P_{ij} \quad (1)$$

Were i represent the color channel and Pij is the probability of occurrence of pixel with intensity j.

- **Standard deviation:** is the square root of the variance of the distribution

$$\sigma_i = \sqrt{\left(\frac{1}{N} \sum_{j=1}^{N} \left(P_{ij} - \mu_i \right)^2 \right)} \quad (2)$$

- **Skewness:** represents the measure of the degree of asymmetry in the distribution.

$$s_i = \sqrt[3]{\left(\frac{1}{N} \sum_{j=1}^{N} \left(P_{ij} - \mu_i \right)^3 \right)} \quad (3)$$

- **Entropy:** represent the disorder or the complexity of the image. A high value of entropy indicates a complex textures.

$$E_i = -\sum_{j=1}^{N} P_{ij} \log_2 \log_2 P_{ij} \quad (4)$$

- **Image dimension:** represents the length and width of the image.

2.2. Classification Stage

After the extraction of the representative vector for each image, every document is classified as a photo, text or a mixed one. Photo family included indoor, outdoor, scenes, landscape, people, logos, and maps. Text family includes scanned and computer-generated text in various fonts. Mixed documents are documents that contain text and photo region.

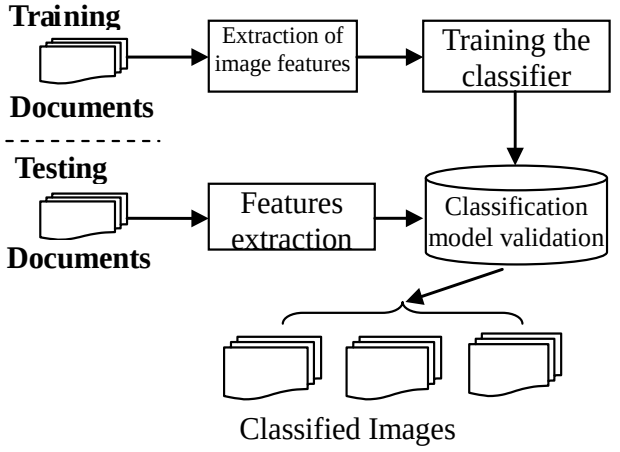

Figure 1. Implementation strategy.

Thus, two well known classifiers are used to classify our documents namely the Decision tree and the Neuronal Network [7,8].

➢ **The Decision Trees**

The Decision Tree Classifier is a set of hierarchical rules which are successively applied to the input data [9]. Those rules are thresholds used to split the data into two binary nodes. Each node is such that the descendant nodes contain more homogeneous data samples. Many features can be input into the Decision Tree to refine class description. A split is chosen because of its ability to render the nodes purer based on a purity measure and can be determined by any single feature [10].

In our paper we fitted the DT to the training data using the cross validation technique in order to select the best tree. Thus, we obtained two tree-based models (original, pruned) that were used in the classification task.

➢ **The Artificial Neuronal Network**

A neural network is a set of connected units (nodes, neurons). Each node has an input and output then it can be connects with other nodes. Each connection has a weight associated to it. The topology of the neural network, the training methodology and the connections between the different nodes define the type of the corresponding Neuronal Network [11-13]. In our case we used an RBF network. In which the input layer had 6 nodes that are equal to the number of features organized as vectors in the database. For the hidden layer, we chose 6 nodes while the output layer contains three nodes. By the end of this process, an input image is classified either as a photo, a pure text or a compound document.

3. Experimental Results

A data base of 291 documents was considered for both classification systems. From this set of documents 75% were used for training and 25% for testing the system performance. Thus, the training data set consists of 136 photo including indoor, outdoor, scenes, landscape images documents, 39 textual documents include scanned and computer-generated text in various font and 51 compound documents. Figure 2 shows some of the class images from the training data set.

In order to evaluate the accuracy of our approach, the following statistical coefficients are computed [14][15]:

- The recall rate= CCI/TI
- The precision rate= CCI/(TI+MI)
- F-measure= $\frac{\left(b^2+1\right) \cdot \text{Precision} \cdot \text{Recall}}{b^2 \cdot \left(\text{Precision} + \text{Recall}\right)}$. Here, b

equals 1.

CCI represents the number of Correctly Classified Images. MI is the number of Misclassified Images and TI is the number of Test Images for each class.

Figure 3 presents the results obtained by using the Decision Tree. We can see that only for textual documents the full Decision Tree achieve high F-measure value than the pruned one.

The results obtained using the neural network as classifier are presented in **Figure 4**. These results show that both classifiers achieve notable results in the classification of documents. The DT classifier outperforms the NN classifier in execution speed and Recall value (by 12%).

There are some cases of misclassification produced by the both classifiers. **Figure 5** shows examples of these images.

The main causes of misclassification on text are due to bad lighting conditions and to excessively noisy backgrounds that cause the final uniformity test to fail.

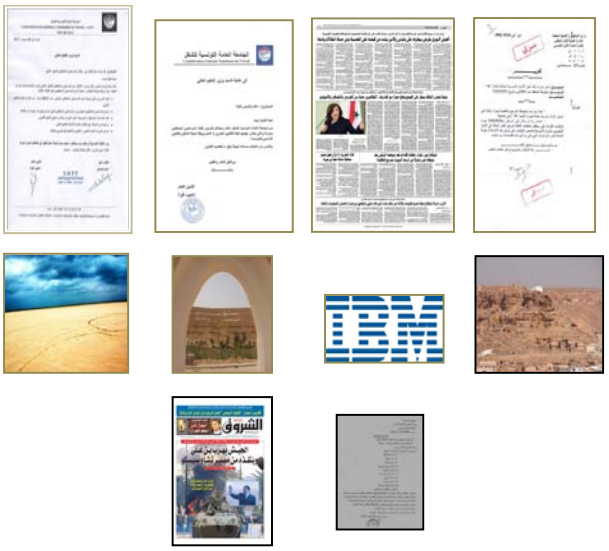

Figure 2. Examples of training data set images.

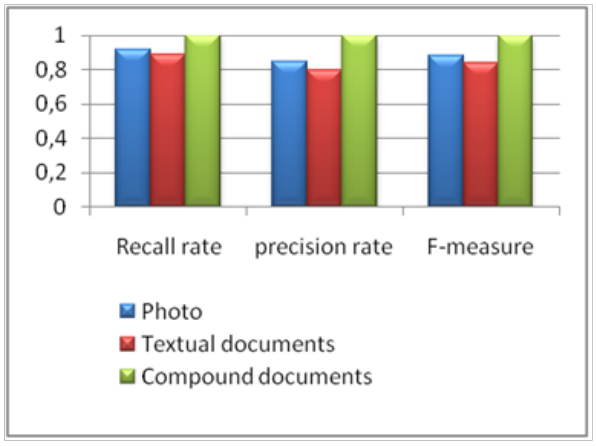

Figure 3. Classification results using DT.

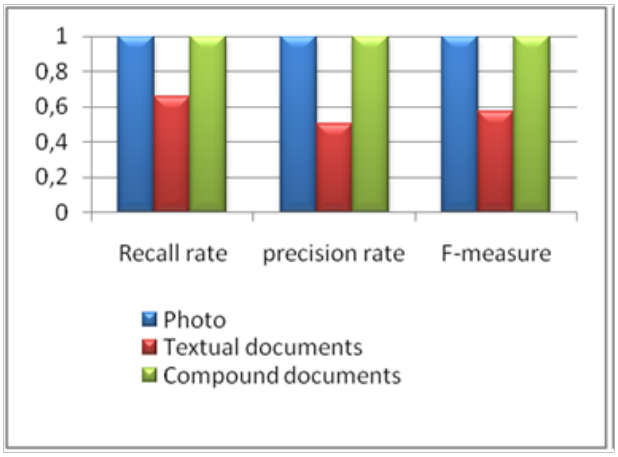

Figure 4. Classification results using NN.

Figure 5. Samples of misclassified images.

4. Conclusions

Automatic classification and archiving of images is an emerging research field in image processing. In this paper an algorithm for classifying photo, textual and mixed documents based on low-level image features was presented. Firstly, features are extracted from images to be assigned to a characteristic vector. Then, the Decision Tree and the neuronal Network classifiers are used to train and to validate a classification model using the extracted feature vectors. The obtained models allowed reaching an accuracy rate of 96% for discriminating a photo, a text and a mixed document.

Nevertheless, features relevance is weighted to select the most contributory ones, in order to increase classification and archiving performance. Moreover, we are currently studying other useful high-level feature to raise the accuracy and to build a new intelligent classifier.

REFERENCES

[1] Chih-Fong Tsai, On Classifying Digital Accounting Documents, The International Journal of Digital Accounting Research, Vol. 7, N. 13, pp. 53-71, 2007

[2] S.J. Simske, Low-resolution photo/drawing classification: metrics, method and archiving optimization, Proceedings IEEE ICIP, IEEE, Genoa, Italy, pp. 534-537, 2005.

[3] Vailaya, A., Figueiredo, M., A. Jain, and H. J. Zhang, Bayesian framework for hierarchical semantic classifica-

tion of vacation images, Proceedings of the IEEE International Conference on Multimedia Computing and Systems (ICMSC), pp. 518- 523, Florence, Italy, 1999.

[4] M. M. Gorkani and R. W. Picard, Texture orientation for sorting photos 'at a Glance', Proc. ICPR, pp. 459-464 Oct. 1994

[5] S. Prabhakar, H. Cheng, J.C. Handley, Z. Fan Y.W. Lin, Picture-graphics Color Image Classification, Proc. of ICIP, pp. 785-788, 2002.

[6] R. Schettini, C. Brambilla, G. Ciocca, Valsasna,M. De Ponti, A hierarchical classification strategy for digital documents, Pattern Recognition, vol 35, pp. 1759-1769, 2002.

[7] Olivier Bousquet, Stéphane Boucheron, and Gabor Lugosi, Introduction to Statistical Learning Theory, Advanced Lectures on Machine Learning, pp.169-207, 2003

[8] S. B. Kotsiantis, Supervised Machine Learning: A Review of Classification Techniques, Informatica journal, Volume 31, Number 3, pp. 249-268, 2007.

[9] Jay Gao, Decision Tree Image Analysis, Digital Analysis of Remotely Sensed Imagery book, The McGraw-Hill Companies, Inc. pp.351-388, 2009.

[10] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, New York: Chapman & Hall, 1984.

[11] G.P. Zhang, Neural Network for classification: A Survey, IEEE Transaction on Systems, Man and Cybernetics-Part C: applications and reviews, Vol.30, no. 4, pp. 451-462, 2000.

[12] Ajith Abraham, Artificial Neural Networks, Handbook of Measuring System Design, Peter Sydenham and Richard Thorn (Eds.), John Wiley and Sons Ltd., London, pp. 901-908, 2005.

[13] Hyontai Sug, Performance Comparison of RBF networks and MLPs for Classification, Proceedings of the 9th WSEAS International Conference on applied Informatics and Communications (AIC '09), pp.450-454, 2009.

[14] Lamiroy, Bart and Sun, Tao, Precision and Recall Without Ground Truth, In Ninth IAPR International Workshop on Graphics RECognition – GREC 2011, Seoul, Core, sep. 2011.

[15] John Makhoul and Francis Kubala and Richard Schwartz and Ralph Weischedel,Performance Measures For Information Extraction, In Proceedings of DARPA Broadcast News Workshop, pp. 249-252,1999.

