SCENE CATEGORIZATION USING LOW-LEVEL VISUAL FEATURES

Ioannis Pratikakis, Basilios Gatos and Stelios C.A. Thomopoulos
Institute of Informatics and Telecommunications, National Centre for Scientific Research "Demokritos", PO. Box 60228, 15310 Aghia Paraskevi, Athens, GREECE

Keywords: Image categorization, visual features, Support Vector Machines (SVM) classifier.

Abstract: In this paper, we have built two binary classifiers for indoor/outdoor and city/landscape categories, respectively. The proposed classifiers consist of robust visual feature extraction that feeds a support vector classification. In the case of indoor/outdoor classification, we combine color and texture information using the first three moments of RGB color space components and the low order statistics of the energy wavelet coefficients from a two-level wavelet pyramid. In the case of city/landscape classification, we combine the first three moments of L*a*b color space components and structural information (line segment orientation). Experimental results show that a high classification accuracy is achieved.

1 INTRODUCTION

The growing proliferation of digital images due to advances in computer technologies and the advent of World Wide Web (WWW) makes imperative the need for robust methods for automatically analyzing, cataloguing, and searching for digital imagery. The major bottleneck for the automatic image categorization has been the gap between low level features and high level semantic concepts. Therefore, the obvious effort toward improving automatic semantics annotation is to focus on methodologies that will enable a reduction or even, in the best case, bridging of the aforementioned gap. In this work, we present a methodology that is directed towards reducing the semantic gap by permitting scene categorization using visual features.

In the literature, several authors have conducted research for the categorization of either indoor/outdoor or city/landscape images. The use of a Bayesian network for integrating knowledge from low-level and mid-level features for indoor/outdoor image classification, is proposed in (Luo, J., and Savakis, A., 2001). In (Stauder *et al., 2004*), there is an attempt to classify images into indoor/outdoor and city/landscape using a set of visual descriptors. In the case of indoor/outdoor classification they use the global color histogram in RGB color space along with a texture descriptor using a 16-tap QMF filter. In the case of city/landscape classification they use a contour descriptor that is a histogram of contour directions. In both cases, they use an SVM classifier. In (Szummer, M., and Picard, R., 1998), they combined features as histograms in Ohta color space, multiresolution, simultaneous autoregressive model parameters and coefficients of a shift invariant DCT, to classify indoor/outdoor images using a K-NN classifier.

In this paper, we have built two binary classifiers for indoor/outdoor and city/landscape categories, respectively. The proposed approach consists of robust fused visual feature extraction that feeds a support vector classification. In the case of indoor/outdoor classification, we combine color and texture information using the first three moments of RGB color space components and the low order statistics of the wavelet coefficients energy from the produced wavelet pyramid. In the case of city/landscape classification, we combine color and structural information using the first three moments of L*a*b color space components and the line segment orientation histogram. Our novelty is based upon building low dimensional combined visual features that achieve among the highest classification accuracies compared to the current state-of-the-art for the classification of indoor vs. outdoor images and city vs. landscape images. Experimental results show the performance of the proposed approach. This paper is organized as follows: Section 2 details the proposed visual feature extraction. Section 3 discusses the classification aspects. Section 4 is dedicated to the experimental

results that demonstrate the performance of the proposed methodology and finally in Section 5 conclusions are drawn.

2 FEATURE EXTRACTION

2.1 Indoor / Outdoor visual feature extraction

In the case of indoor/outdoor classification, robust features are extracted using a combination of color and texture visual information.

Color is an important cue for image categorization. Using a particular color space, the corresponding feature can carry all its specific characteristics. For this reason, we have used three different color spaces. Apart from RGB color space, we made experiments with L*a*b* and LST color space. The L*a*b color space is approximately perceptually uniform; thus, distances in this space are meaningful (Wyszecki, G. and Stiles, W., 1982). The LST color space is introduced in (Serrano *et al.*, 2004) and is defined by:

$$\begin{aligned} L &= \frac{k}{\sqrt{3}}(R+G+B), \\ S &= \frac{k}{\sqrt{2}}(R-B), \\ T &= \frac{k}{\sqrt{6}}(R-2G+B) \end{aligned} \quad (1)$$

where L is the luminance channel, S and T are the chrominance channels and k = 255/max{R,G,B}. The S,T chrominance components support light source intensity invariance. This is an important point since the spectral characteristics of the particular categories we are dealing with (indoor, outdoor, and city, landscape, as well) can vary considerably.

Let x_{ij} denotes the pixel value. Each image pixel has a three dimensional color vector $C(x_j) = [C_j^1, C_j^2, C_j^3]$ in the selected color space. We compute the three first color moments that are denoted as follows:

First moment (mean) :

$$E_i = \frac{1}{N}\sum_{j=1}^{N} x_{ij} \quad (2)$$

Second central moment (variance) :

$$\sigma_i = \left(\frac{1}{N}\sum_{j=1}^{N}(x_{ij} - E_i)^2\right)^{1/2} \quad (3)$$

Third central moment (skewness) :

$$s_i = \left(\frac{1}{N}\sum_{j=1}^{N}(x_{ij} - E_i)^3\right)^{1/3} \quad (4)$$

where $i \in \{R,G,B\}$ or $i \in \{L,*a,*b\}$ or $i \in \{L,S,T\}$, x_{ij} denotes the pixel value and N denotes the image size.

Considering the above analysis, the color feature vector consists of 9 coefficients.

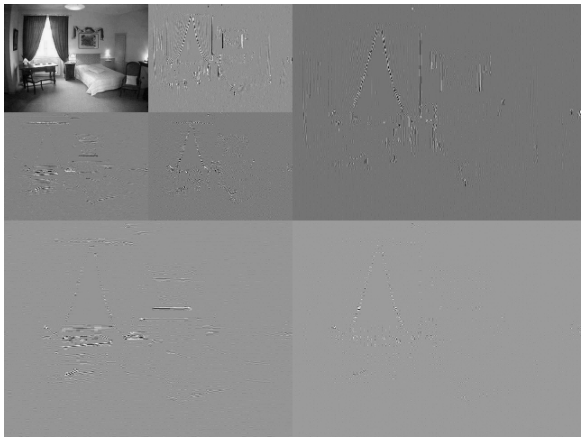

Figure 1: A two-level wavelet pyramid structure of the “indoor” image at Fig. 4a.

Apart from the color, visual information that corresponds to texture is used. For this purpose, we apply the wavelet transform using the Daubechies 7-9 biorthogonal filters (Daubechies, I., 1988) to the Luminance component of the color image. An example of the produced two-level wavelet pyramid structure is shown at Figure 1. Although, Haar is the simplest wavelet function (small spatial support) that consequently affects the execution time, we have not opted for it since there exist localisation drawbacks due to its non-overlapping wavelets at a given scale (Sebe, N., and Lew, M.S., 2003). In multiresolution wavelet analysis, we have the creation of sub-bands due to the application of the combination of the low-pass and the high-pass counterpart of the above mentioned filters. The final number of sub-bands is related to the number of resolution levels that an image is considered for analysis. In the proposed approach, the texture feature vector is made of the mean and the variance of the energy to each produced sub-band. The motivation for using these features is their reflection of texture properties that

has proven effective for discerning texture (Unser, M., 1995). The texture feature vector consists of 14 coefficients (7 for the mean and 7 for the variance) which are produced due to seven sub-bands that are created for two (2) resolution levels. In total, the feature vector used for indoor/outdoor classification consists of 23 coefficients.

2.2 City/Landscape visual feature extraction

In the case of city/landscape classification, robust features are extracted using a combination of color and structural information expressed by the line segment orientation.

The color is considered in the same manner as in the case of indoor/outdoor feature extraction. We obtain a vector of 9 coefficients that have been computed using the Equations 2-4. Together with color, we use a line segment descriptor. The underlying idea is to distinguish between long horizontal and vertical contours that dominate in city images and short length contours having other directions than either horizontal or vertical that can be found in landscape images. A similar contour descriptor has been proposed in (Stauder *et al.,* 2004) leading to the extraction of a 12-bin histogram while in (Vailaya *et al.,* 2001) the edge direction distribution has been proposed for the discrimination between city and non-city images.

To construct the line segment descriptor which is a histogram of line segment directions, we follow the next steps. First, we apply an edge detection using the Canny edge detector (Canny, J., 1986). The produced edges are thinned and thereafter we try to transform the edge representation into a line segment representation. For this, we apply a non-parametric curve segmentation into straight lines as it is explained in (Rosin, P.,L., and West, G.A.W., 1995). The direction of each straight line is calculated and categorized as being either horizontal, vertical or diagonal. Furthermore, the line segment length is taken into account in order to be labelled as either short or long segment. A segment will be considered as a long one if it is greater than 10% of the minimum dimension (either width or height). Finally, a histogram with six (6) bins is computed. A schematic representation of the different required steps is shown at Figure 5. In total, the feature vector used for city/landscape classification consists of 15 coefficients.

3 CLASSIFICATION - FEATURE FUSION

In the particular binary classification problem (indoor vs. outdoor and city vs. landscape) the classification step was performed using two well-known classification algorithms, K-NN (Theodoridis, S., and Koutroumbas, K., 1997) and Support Vector Machines (SVM) (Cortes C., and Vapnik, V., 1995)(Vapnik, V., 1998)(Chang, C.C., and Lin, C.-J.).

Formally, the support vector machines (SVM) require the solution of an optimisation problem, given a training set of instance-label pairs (x_i, y_i), i=1,…,m, where $x_i \in R^n$ and $y_i \in \{1,-1\}^m$. The optimisation problem is defined as follows :

$$\begin{aligned} &\min_{\omega,b,\xi} && \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{m}\xi_i \\ &subject\ \ to && y_i(\omega^T\phi(x_i)+b) \geq 1-\xi_i \\ & && \xi_i \geq 0 \end{aligned} \qquad (5)$$

According to this, training vectors x_i are mapped into a higher dimensional space by the function ϕ. Then, SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space. For this search, there are a few parameters that play a critical role at the classification performance. Firstly, the parameter C at Eq. 5, that applies a penalty at the error term. Secondly, the so-called kernel function denoted as : $K(x_i, x_j) \equiv \phi(x_i)^T\phi(x_j)$.

One of the main aspects in classification is the interaction between the features and the available classifiers. Mainly, there are two trends in this interaction. Either different features are combined into a final feature vector as the input to the classifier (Lim, H-H., and Jin, J.S., 2005), (Stauder *et al.,* 2004), or feature vectors associated with different modalities are fed into independent pattern classifiers whose classification outputs are then combined (Serrano *et al.*, 2004), (Szummer, M., and Picard, R., 1998), (Payne, A., and Singh, S., 2005). These basic trends have shown both advantages and disadvantages. A disadvantage of the latter trend is that the training of multiple classifiers on individual features may not be viable at all, as single feature does not provide sufficient discriminative power, resulting in many poor classifiers for fusion.

In our approach, we follow the former trend, where the classifier's input feature vector consists of a concatenation of each feature that is considered for the corresponding classification problem (indoor vs. outdoor, or city vs. landscape). A detailed discussion about these features has already been given at Section 2.

4 EXPERIMENTAL RESULTS

For our experiments, we have considered a generic database of about 1600 color images that have been collected from various sources like the web, the MPEG-7 VCE-2 dataset (MPEG-7), the Corel photo galleries (Corel) and the Microsoft Research Cambridge Object Recognition Image Database 1.0 (MRC). All images have a 24-bit color resolution and there exist a great variety in their size. We have split the available database in training dataset and testing dataset. The exact distribution in size of those datasets can be seen at Table 1.

Regarding the classification step, K-NN was used with ($k=5$). On the other hand, SVM was used in conjunction with the Radial Basis Function (RBF) kernel, a popular, general-purpose yet powerful kernel, denoted as:

$$K(x_i, x_j) \equiv \exp(-\gamma \left\| x_i - x_j \right\|^2) \quad (6)$$

Furthermore, a grid search was performed in order to find the optimal values for both the variance parameter (γ) of the RBF kernel and the cost parameter (C) of SVM (see Eq. 5).

Quantification of our experiments is shown at Table 2. First, we can observe that support vector classification (SVM) outperforms K-NN classification for both indoor/outdoor and city/landscape classification problems. Second, the proposed visual features combined with a support vector classification (SVM) produces high classification accuracy that is, 95% for indoor/outdoor and 89.97% for city/landscape images. The achieved performance is among the highest in the state-of-the-art. To test the influence of the color space, we have carried out experiments with different color spaces. Specifically, we have used RGB, LST and L*a*b color space. We found out that the combined feature set performs better in the case of RGB color space for indoor/outdoor classification, while the L*a*b color space supports a superior performance for city/landscape classification. The detailed comparison in terms of produced classification accuracy can be seen at Table 2.

Furthermore, for the sake of clarity, we do not only present examples of correctly classified images in Figure 4, but also examples of misclassified images are shown in Figures 2, 3. Even after a brief examination of these examples we may easily understand the difficulty to avoid misclassification. For example, in indoor images that have been classified as outdoor there is much green color that may correspond to outdoor as well as vivid colors that is not the case of indoor images. Also, in city images like the city of Amsterdam that we get a picture which includes a river and trees, these characteristics are the dominant characteristics of landscape images, advocating the involved misclassification. In these cases a contextual knowledge is imperative to be considered for having a correct classification.

5 CONCLUSIONS

In this work, we have presented two binary classifiers for indoor/outdoor and city/landscape categories. The basic component of the proposed scheme for robust feature extraction is expressed as: (i) in the case of indoor/outdoor classification, a combination of color and texture information using the first three moments of RGB color space components and the low order statistics of the wavelet coefficients energy from the produced wavelet pyramid; (ii) in the case of city/landscape classification, a combination of color and structural information using the first three moments of L*a*b color space component along with the line segment orientation histogram. The proposed features along with a support vector classification produce high classification accuracy.

We aim to apply the proposed methodology on larger datasets. We opt on working towards methods that do not calculate global image features. In this case, meaningful regions could be identified that correspond to objects in an image and consequently feature computation can be applied over the corresponding regions.

REFERENCES

Canny, J., 1986. A Computational Approach to Edge Detection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-8, No. 6, pp. 679-698.

Chang, C.C., and Lin, C.-J., *LIBSVM: a library for support vector machines*, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Corel Corp. http://www.corel.com

Cortes C., and Vapnik, V., 1995. Support-vector network, *Machine Learning*, vol. 20, pp. 273-297.

Daubechies, I., 1988. Orthonormal bases of compactly supported wavelets, *Communications on Pure and Applied Mathematics*, vol. XLI, no. 41, pp. 909-996.

Microsoft Research Cambridge (MRC) Object Recognition Image Database 1.0, http://research.microsoft.com/research/downloads/

MPEG-7 VCE-2 Dataset, http://mpeg.nist.gov

Lim, J-H., and Jin, J.S., 2005. Combining intra-image and inter-class semantics for consumer image retrieval, *Pattern Recognition*, vol. 38, pp. 847-864.

Luo, J., and Savakis, A., 2001. Indoor vs. Outdoor classification of consumer photographs using low-level and semantic features, *In Proceedings of International Conference of Image Processing (ICIP)*, vol. II, pp. 745-748.

Payne, A., and Singh, S., 2005, Indoor vs. outdoor scene classification in digital photographs, *Pattern Recognition*, vol. 38, pp. 1533-1545.

Rosin, P.,L., and West, G.A.W., 1995. Nonparametric segmentation of curves into various representations, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, No. 12, pp. 1140-1153.

Sebe, N., and Lew, M.S., 2003, Comparing salient point detectors, *Pattern Recognition Letters*, vol. 24, pp. 89-96.

Stauder, J., Sirot, J., Le Borgne, H., Cooke E., and O'Connor, N.E., 2004. Relating visual and semantic image descriptors, *In Proc. of European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*,pp. 221-228, London, UK, November 25-26.

Serrano, N., Savakis, A., and Luo, J., 2004. Improved scene classification using efficient low-level features and semantic cues, *Pattern Recognition*, vol. 37, pp. 1773-1784.

Szummer, M., and Picard, R., 1998. Indoor-outdoor image classification, *In Proc. of Int. Workshop on Content-based access of image and video databases*, pp. 42-51.

Theodoridis, S., and Koutroumbas, K., 1997. *Pattern Recognition*, Academic Press.

Unser, M., 1995. Texture Classification and Segmentation Using Wavelet Frames, *IEEE Trans. Image Processing*, vol. 4, no. 11, pp. 1549-1560.

Vailaya, A., Figueiredo, M.A.T., Jain, A.K, and Zhang, H-J., 2001, Image classification for content-based indexing, *IEEE Transactions on Image Processing*, vol. 10, No. 1, pp. 117-130.

Vapnik, V., 1998. *Statistical Learning Theory*, Wiley, New York.

Wyszecki, G. and Stiles, W., 1982. Color Science: Concepts and Methods, Quantitative Data and Formulae, second ed. Wiley.

Table 1: Training / Testing Dataset size

Database	Training Dataset size	Testing Dataset size
Indoor	170	173
Outdoor	210	210
City	209	209
Landscape	210	208

Table 2: Classification accuracy (**%**)

COLOR SPACE	CLASSIFIERS	CATEGORIES	
		Indoor / Outdoor	City / Landscape
RGB	K-NN	91,98	81,18
	SVM	**95** (*C*=170, γ=0.4)	86,19 (*C*=180, γ=0.007)
LST	K-NN	87,92	84,25
	SVM	94,73 (*C*=170, γ=0.007)	84,65 (*C*=180, γ=0.008)
L*a*b	K-NN	90,16	78,99
	SVM	94,21 (*C*=170, γ=0.4)	**89,97** (*C*=180, γ=0.005)

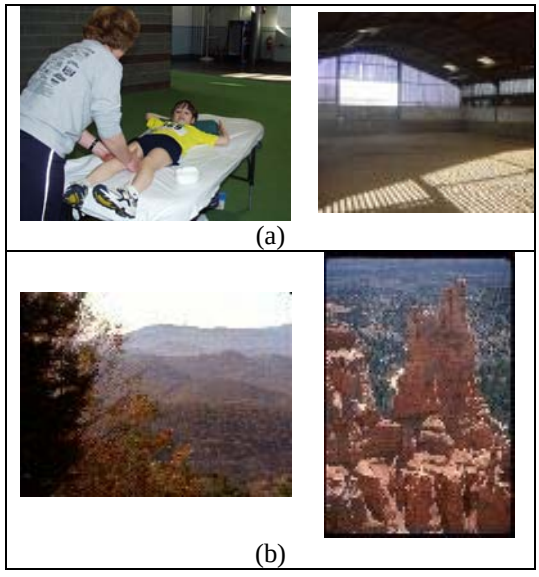

Figure 2: Examples of misclassified images : (a) Indoor images classified as outdoor; (b) Outdoor images classified as indoor.

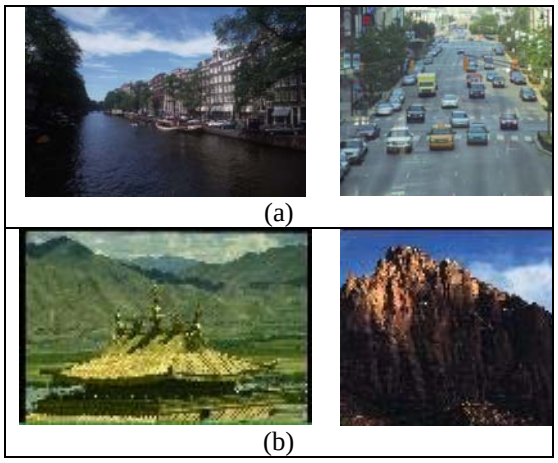

Figure 3: Examples of misclassified images : (a) City images classified as Landscape; (b) Landscape images classified as City.

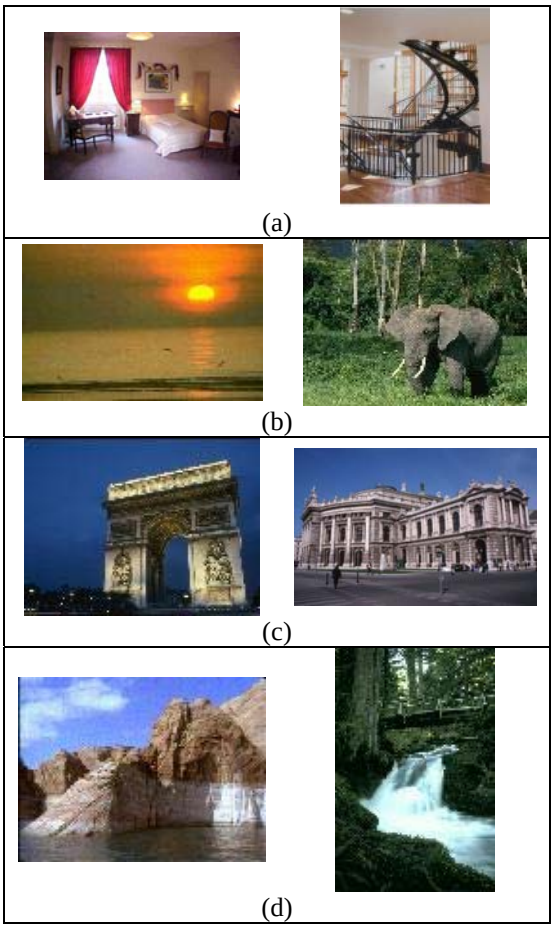

Figure 4: Examples of correctly classified images : (a) Indoor images; (b) Outdoor images; (c) City images; (d) Landscape images.

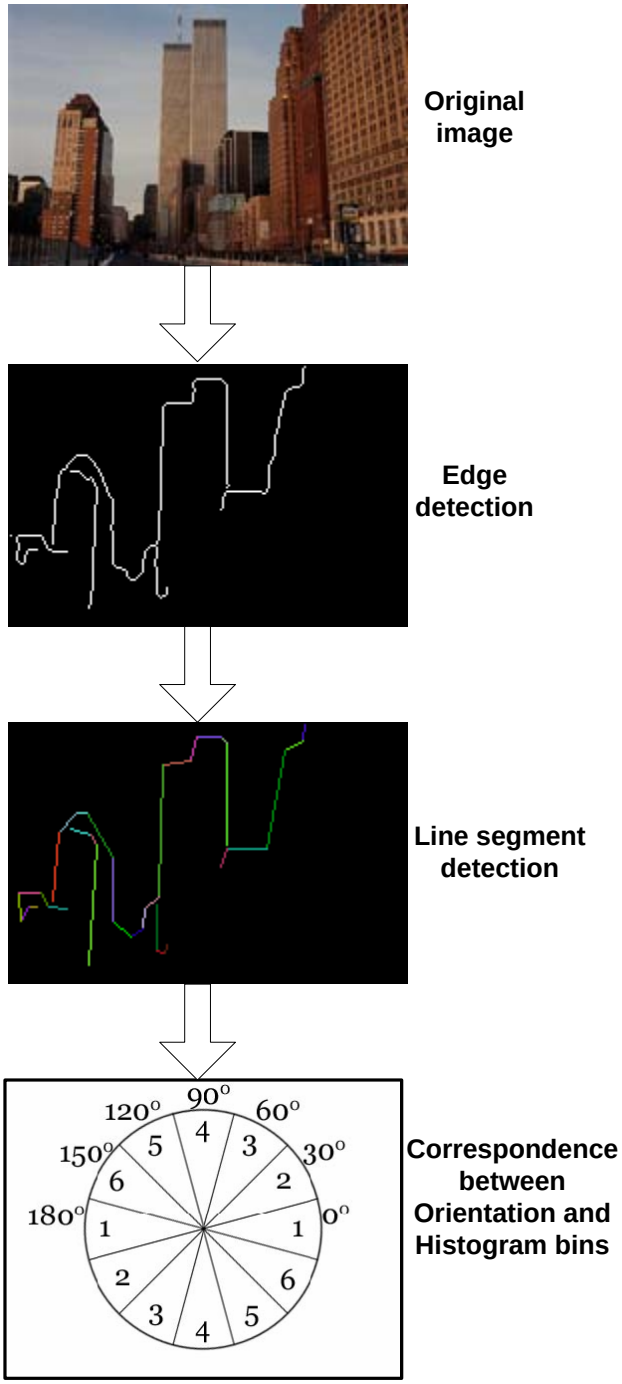

Figure 5: Visual block representation of the required steps for the computation of the proposed contour descriptor.