Fusing MPEG-7 visual descriptors for image classification

Evaggelos Spyrou[1], Hervé Le Borgne[2], Theofilos Mailis[1], Eddie Cooke[2], Yannis Avrithis[1], and Noel O'Connor[2]

[1] Image, Video and Multimedia Systems Laboratory, School of Electrical and Computer Engineering, National Technical University of Athens,9 Iroon Polytechniou Str, 157 73 Athens, Greece,
espyrou@image.ece.ntua.gr,
WWW home page: http://www.image.ece.ntua.gr/∼espyrou/
[2] Center for Digital Video Processing, Dublin City University, Collins Ave., D9, Ireland

Abstract. This paper proposes a number of content-based image classification techniques based on fusing various low-level MPEG-7 visual descriptors. The goal is to fuse several descriptors in order to improve the performance of several machine-learning classifiers. Fusion is necessary as descriptors would be otherwise incompatible and inappropriate to directly include e.g. in a Euclidean distance. Three approaches are described: A "merging" fusion combined with an SVM classifier, a back-propagation fusion combined with a K-Nearest Neighbor classifier and a Fuzzy-ART neurofuzzy network. In the latter case, fuzzy rules can be extracted in an effort to bridge the "semantic gap" between the low-level descriptors and the high-level semantics of an image. All networks were evaluated using content from the aceMedia Repository [3] and more specifically in a *beach/urban* scenes classification problem.

1 Introduction

Content-based image retrieval (CBIR) consists of locating an image or a set of images from a large multimedia database, in order to satisfy a user need. It is broadly accepted that such a task can not be performed by simply manually associating words to each image of a database, first because it would be a very tedious task with the exponential increasing quantity of digital images in all sort of databases (web, personal database from digital camera, professional databases and so on) and secondly because "images are beyond words", [1] that is to say their content can not be fully described by a list of words. It thus requires an extraction of visual information directly from the images. This is is usually called low-level features extraction.

[3] This work was supported by the EU project aceMedia "Integrating knowledge, semantics and content for user centered intelligent media services" (FP6-001765).Hervé Le Borgne and Noël O'Connor acknowledge Enterprise Ireland for its support through the Ulysse-funded project ReSEND FR/2005/56.

Unfortunately, it is an unsolved problem that requires bridging the gap between the target semantic classes, and available low-level visual descriptors. Regardless of whether classification is supervised or unsupervised, and regardless of the specific classification model employed, e.g., expectation maximization, vector quantization, k-means, [2], support vector machines (SVM) [3], or neural networks, it is commonly believed that in order to achieve robust global classification, i.e. without prior object detection or recognition, it is crucial to select an appropriate set of descriptors. Visual descriptors usually have to capture the particular properties of a specific domain and the distinctive characteristics of each image class. For instance, local color descriptors and global color histograms are used in indoor/outdoor classification [4] to detect e.g. vegetation (green) or sea (blue). Edge direction histograms are employed for city/landscape classification [5] since city images typically contain horizontal and vertical edges. Additional motion descriptors are also used for sports video shot classification [6].

In this paper we address the problem of fusing these low-level features for still images classifications. More specifically we focus on *early fusion* methods, when the combination of features is performed before or at the same time as the estimation of the distances between images. An alternative strategy is to perform matching on individual features and fuse the matching scores. More information on these *late fusion* methods can be found in [8].

In this work, fusion of several MPEG-7 descriptors is approached using three different machine learning techniques. More specifically, a Support Vector Machine is used with a "merging" descriptors' fusion, a Back-Propagation Feed-Forward neural network is trained to estimate the distance between two images based on their low-level descriptors and using its results, a K-Nearest Neighbor Classifier is applied. Finally in order to extract fuzzy rules and bridge low-level features with the semantics of images, a Falcon-ART Neurofuzzy Network is used. All three networks are trained using the same training set and all three classification strategies are evaluated using the same testing set.

Section 2 gives a brief description of the scope of the MPEG-7 standard and presents the three low-level MPEG-7 descriptors used in this work. Section 3 presents the three different techniques that aim at image classification using these descriptors.Section 4 describes the procedure followed to train the machine learning systems we applied along with the classification results and finally conclusions are drawn in section 5 and plans for future work are presented.

2 Feature extraction

In order to provide standardized descriptions of audio-visual (AV) content, MPEG-7 standard [9] specifies a set of descriptors, each defining the syntax and the semantics of an elementary visual low-level feature *e.g.*, color, shape. In this work, the problem of image classification is based on the use of three MPEG-7 visual descriptors. Their extraction is performed using the aceToolbox, developed within aceMedia[4]. This toolbox is based on the architecture of the MPEG-7

[4] http://www.acemedia.org

eXperimentation Model and uses a subset of these low-level visual descriptors of color and texture in order to identify and categorize images. A brief overview of each descriptor is presented below while more details can be found in [10]

Color Layout Descriptor (CLD) is a compact and resolution-invariant MPEG-7 visual descriptor defined in the YCbCr color space and designed to capture the spatial distribution of color in an image or an arbitrary-shaped region. The feature extraction process consists of four stages. The input image is partitioned into $8 \times 8 = 64$ blocks, followed by the detection of the representative (average) color. A DCT transformation is applied on the resulting image (size 8×8). The resulting coefficients are zig-zag-scanned and only 6 coefficients for luminance and 3 for each chrominance are kept, leading to a 12-dimensional vector. Finally, the remaining coefficients are nonlinearly quantized.

Scalable Color Descriptor(SCD) is a Haar-transform based encoding scheme that measures color distribution over an entire image. The color space used is the HSV, quantized uniformly to 256 bins.To sufficiently reduce the large size of this representation, the histograms are encoded using a Haar transform allowing also allowing scalable coding.

Edge Histogram Descriptor (EHD) captures the spatial distribution of edges, in a similar way to the CLD and provides a useful description of the edges, even when the underlying texture is not homogeneous. Four directions of edges ($0\,^{\circ}$, $45\,^{\circ}$, $90\,^{\circ}$, $135\,^{\circ}$) are detected in addition to non-directional ones. The input image is divided in 16 non-overlapping blocks and a block-based extraction scheme is applied to extract the five types of edges and calculate their relative populations, resulting in a 80-dimensional vector.

3 Image Classification based on MPEG-7 Visual Descriptors

In order to handle all the above descriptors at the same time for tasks like similarity/distance vector formalization or training of classifiers, it is necessary to fuse the individual, incompatible elements of the descriptors.

When the task is to compare two images based on a single MPEG-7 visual descriptor, several distance functions, MPEG-7 standardized or not, can be used. In our case, however, the comparison should consider all three low-level descriptors presented in section (2) with different weights on each. The problem is not trivial, as there is not a unique way to compute this distance and apart from this, the sufficiency of the standardized distance functions cannot be guaranteed for each given database.

3 methods are considered for this purpose, combined with appropriate classification techniques.They all share in common the same MPEG-7 Visual Descriptors described above. *Merging fusion* combines the three descriptors using a Support Vector Machine for the classification, *Back-propagation fusion* produces a “matrix of distances” among all images to be used with a K-Nearest Neighbor Classifier. Finally, a *Fuzzy-ART neurofuzzy network* is used not only for classification but also to produce semantic fuzzy rules.

3.1 Merging fusion/SVM classifier

A naive fusion strategy was implemented in order to be compared to others. It simply consists of merging all the descriptions into a unique vector and is called *merging fusion*. This strategy requires all features to have more or less the same numerical values to avoid scale effects. An alternative is to re-scale the data using principal component analysis for instance. Re-scaling is not necessary in our case since the mpeg-7 descriptor we use are already scaled to integer values of equivalent magnitude.

3.2 KNN classification using Back-Propagation Fusion

The second method is based on a back-propagation feed-forward neural network with a single hidden layer. The network's input consists of the low-level descriptions of the two images whose distance needs to be evaluated and its output is the normalized estimation of their distance, based on all available descriptors. A training set is constructed by carefully selecting some representative images from the available database in a way that all the possible different kind of scenes that belong to each category would be presented to the network. The network is trained under the assumption that the distance of two images belonging in the same class is minimum while in all the other cases is maximum. Thus, when two images with unknown distances are presented to the network, the network responds with an estimation of their distance. By presenting all the images used for the training, a matrix is created. Each element of the matrix is the estimated distance between the image that corresponds to its row and the one that corresponds to its column. This matrix is then used by a K-Nearest Neighbor (KNN) classifier that assigns to an image the same label as the majority of its K nearest neighbors.

A problem that occurs is that the distance between descriptors belonging to the same image is estimated rather as a very small number than zero which would obviously be the desired response. However, these small non-zero distances can be replaced by zero values as the distance an image to itself is a priori known to be zero. Apart from this, the response of the network depends on the row that the descriptors are presented into an input vector of the network. It is apparent that even for a well-trained network, the output would be slightly different and the "matrix of distances" would not respect the basic property of distances needed by the KNN classifier. To overcome this problem, an approach is to use only the distances of the upper triangular matrix or only those of the lower. This way, the distance $\hat{d}(i,j)$ between two images can be determined as

$$\hat{d}(i,j) = \hat{d}(j,i) = d(max(i,j), min(i,j))$$

Another way to symmetrize the matrix is the replacement of a distance by the average of the two corresponding outputs. More specifically, for a random element $d(i,j)$, the estimated distance $\hat{d}(i,j)$ is calculated as:

$$\hat{d}(i,j) = \hat{d}(j,i) = \frac{1}{2} \cdot [(d(i,j) + d(j,i)]$$

After these two modifications, the produced matrix respects the two fundamental properties demanded by the KNN classifier.

Moreover, another approach to efficiently fuse the different visual descriptors uses precalculated distance matrices for individual visual descriptors and tries to assign weights on each one, to produce a weighted sum. Three matrices are calculated containing the $L2$ (Euclidean) distances of all pairs of images, corresponding to CLD, SCD and EHD. This time, the input of the neural network consists of the three distances and the output is the normalized fused distance which can be considered their normalized and weighted. The resulting "matrix of distances" is used again as the input of a KNN classifier. This method results to a symmetric "matrix of distances", however, its diagonal is again set to zero in order to satisfy the demands of the KNN classifier.

3.3 Classification using a Falcon-ART Neurofuzzy Network

Image classification using a neural network or a Support Vector Machine fails to provide semantic interpretation of the underlying mechanism that actually realizes the classification. In order to extract semantic information, a Neurofuzzy Network can be applied. To achieve this, we used the Falcon-ART network [11] which is based on the Fuzzy-ART clustering algorithm presented in[12]. A short description of the Falcon-ART follows.

The training of the network is done in two phases, the "structure learning phase", where the Fuzzy-ART algorithm is used to create the structure of the network, and the "parameters learning stage", where the parameters of the network are improved according to the back-propagation algorithm. The Fuzzy-ART algorithm creates hyperboxes (a hyperbox is the extension of a box in a n-dimensional space) in the input space. Each hyperbox corresponds to a cluster and grows towards the direction of the presented training samples that actually belong to the category it represents. New hyperboxes may be created in accordance with the vigilance criterion and after the appliance of the Fuzzy-ART algorithm, the input space is clustered and a 5-layer network is created from the input hyperboxes and the connections among them. The parameters and the connections between the nodes of each layer arise from the Fuzzy-ART algorithm and the resulting five-layer network is trained according to the back-propagation algorithm.

A "merged" descriptor that contains the low-level information of all three MPEG-7 visual descriptors we used is given as the input of the network. After the training phase, the trained network's response should be the class that the input belongs. However, the main advantage of the neurofuzzy network is that it allows the extraction of semantic fuzzy rules. Hence, the way that the low-level features of the image determine the class to which it belongs becomes more obvious and can be described in natural language. Let $D = [d_1, d_2, \ldots, d_n]$ a low-level descriptor. In terms of this descriptor, a fuzzy rule could be stated as:

> IF d_1 is *low* AND d_2 is *medium* AND …AND d_n is *high*, THEN the image belongs to class 1

4 Experimental results

The image database used for the experiments is part of the aceMedia Content Repository [5]. More specifically it is part of the Personal Content Services database and consists of 767 high quality images divided in two classes *beach* and *urban*. 60 images (40 from *beach* and 20 from *urban*)were used to train the neural network and the other 707 images (406 from *beach* and 301 from *urban*) were used to test the efficiency of the different classification approaches. All possible combinations of the three visual descriptors were considered in all three approaches and the results are presented in table 1.

SVM Classifier using Merging Fusion: The merged vectors were directly used as input of a SVM classifier [13] with a polynomial kernel of degree one (*i.e* a linear kernel). Results with polynomial kernels of higher degree (up to 5) give similar results. 40 images from the *beach* category and 20 from the *urban*were selected as the representative examples of the given classes for the considered database, then used as training dataset for the SVM classifier. The remaining 707 images of the database were used for testing.

Classification results on table **??** show that the performances increase when several low-level features are merged. While individual features lead to classification results from 79.5% to 83.6%, the merging of two of them improve the classification results from 86.9% to 88.7%, and reaches 89% with the merging of the three.

Back-Propagation Fusion of Merged Descriptors: 40 images from the *beach* category and 20 from the *urban*were carefully selected as the most representative of the given database. The distance between two images was determined manually and was set to 0 for images belonging to the same category and to 1 for images belonging to different categories. In order to improve the training, from each pair of images, two training samples have been created by reversing the order of the images as the desired response of the network should remain the same, independently of the order. Thus, 2800 training samples were created in order to train the back propagation network. The remaining 707 images were used to evaluate the back-propagation fusion. By presenting all the possible pairs to the network, a "matrix of distances" was created. The symmetric "matrices of distances" (as described in part 3) were used with the KNN classifier.

Falcon-ART Neurofuzzy Network: The same 60 images as in back-propagation fusion were used as the training set of the Fuzzy-ART classifier and the Falcon-Art neurofuzzy network. The network was trained with their "merged descriptors" presented randomly at the network. The success rate was 95.8% on the training set and 87.7% on the test set, with the Fuzzy-ART algorithm creating 8 hyperboxes (rules) and the Falcon-ART neurofuzzy network being trained for 275 epochs. In order to have a more close to human perception description of the rules the Falcon-Art algorithm created, each dimension of an image descriptor was divided into three equal parts each one corresponding to

[5] http://driveacemedia.alinari.it/

Classification	EH	CL	SC	EH+CL	EH+SC	CL+SC	EH+CL+SC
Merging/linear SVM	79.5%	82.3%	83.6%	87.1%	88.7%	86.9%	89.0%
Back-Prop.L2 dist./KNN	-%	-%	-%	88.97%	89.25%	88.54%	93.49%
Back-Prop.KNN.	81.9%	87.13%	85.86%	67.04%	90.1%	91.37%	86.28%
Falcon-ART	81.4%	84.7%	83.67%	82.4%	83.6%	86.3%	87.7%

Table 1. Classification rate using several approaches on different MPEG-7 descriptors: edge histogram (EH), color layout (CL) and scalable color (SC)

part of image	edge type	Rule 1	Rule 2	Rule 3	Rule 4	Rule 5
upper	0°	M	L	M-L	M-L	L
	45°	M	L	M-L	M	M
	90°	M	L	M	M	M
	135°	H	M	M	M	M
	nondir.°	M	M	M	M	M
center	0°	M	L	M	M	M
	45°	M	M-L	H	M	H
	90°	M	M	M	M	H
	135°	H	M	M-L	H	H
	nondir.°	H	M	M	H	M
lower	0°	M	L	L	M	L
	45°	M	M	H	H	H
	90°	H	M	M	H	H
	135°	M	M-L	M	H	M
	nondir.°	M	M	M-L	H	M
class		*urban*	*beach*	*urban*	*beach*	*beach*

Table 2. Fuzzy Rules created by the Falcon-ART, trained with the EH descriptor

low, *medium*, *high* values and each hyperbox created by the Falcon-ART has led to a rule that uses *low*, *medium* and *high* values.

The extraction of the fuzzy rules is performed in a way that the rules are simple and comprehensive. We present an example of such a rule, when classification considers only the EHD descriptor as stated in section 2. The subimages are grouped to those describing the upper, middle and lower, parts of the image and a qualitative value (*low*, *medium* or *high*) is estimated for each type of edges. Thus, a fuzzy rule can be stated as:

> IF the number of 0° edges on the *upper* part of the image is *low* AND the number of 45° edges on the *upper* part of the image is *medium* AND ...AND the number of non-directional edges on the *lower* part of the image is *high*, THEN the image belongs to class *Beach*

In the case of the EHD descriptor, the Falcon-ART has created 5 fuzzy rules which are presented in detail in table 2.

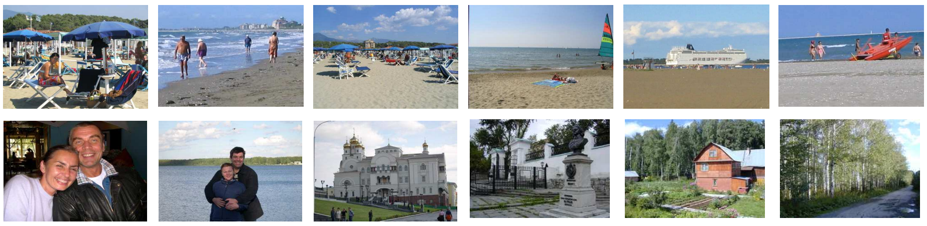

Fig. 1. Representative Images - First Row:Beach Images, Second Row: Urban Images

5 Conclusion and future works

All three tested methods were applied successfully to the problem of image classification using three MPEG-7 descriptors. Back-propagation fusion showed the best results followed by the merging fusion using the support vector machine. However fusion using the Falcon-ART was useful as it provided a linguistic description of the underlying classification mechanism. Future work will aim to use more MPEG-7 descriptors and more classes. Additionally, these classification strategies may be extended in matching the segments of an image with predefined object models with possible applications in image segmentation.

References

1. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE t. PAMI **22** (2000) 1349–1380
2. R.O. Duda, P.E. Hart, D.S.: Pattern Classification. John Wiley and Sons (2001)
3. Vapnik, V.: Statistical Learning Theory. John Wiley and Sons (1998)
4. Szummer, M., Picard, R.: Indoor-outdoor image classification. In: IEEE international workshop on content-based access of images and video databases,. (1998) Bombay, India.
5. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City images vs. landscapes. Pattern Recognition **31** (1998) 1921–1936
6. D.H. Wang, Q. Tian, S.G.W.K.S.: News sports video shot classification with sports play field and motion features. ICIP04 (2004) 2247–2250
7. Mc Donald, K., Smeaton, A.: A comparison of score, rank and probability-based fusion methods for video shot retrieval. In: CIVR 2005 - International Conference on Image and Video Retrieval. (20-22 July 2005) Singapore.
8. Chang, S.F., Sikora, T., Puri, A.: Overview of the mpeg-7 standard. IEEE trans. on Circuits and Systems for Video Technology **11** (2001) 688–695
9. Manjunath, B., Ohm, J.R., Vasudevan, V.V., Yamada, A.: Color and texture descriptors. IEEE trans. on Circuits and Systems for Video Technology **11** (2001) 703–715
10. Lin, C.T., Lee, C.S.G.: Neural-network-based fuzzy logic control and decision system. IEEE trans. Comput. **40** (1991) 1320–1336
11. Carpenter, G., Grossberg, S., Rosen, D.: A neural network realization of fuzzy art. technical report CAS/CNS-91-021 (1991)
12. Vapnik, V.: The Nature of Statistical Learning Theory. NY:Springer-Verlag (1995)