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, National Technical University of Athens, 9 Iroon Polytechniou Str, 157 73 Athens, Greece
espyrou@image.ece.ntua.gr
http://www.image.ece.ntua.gr/∼espyrou/
[2] Center for Digital Video Processing, Dublin City University, Collins Ave., Ireland

Abstract. This paper proposes three content-based image classification techniques based on fusing various low-level MPEG-7 visual descriptors. 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 KNN 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 repository of the aceMedia project[1] and more specifically in a *beach/urban* scene classification problem.

1 Introduction

Content-based image retrieval (CBIR) consists of locating an image or a set of images from a large multimedia database. Such a task can not be performed by simply manually associating words to each image , firstly 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. Thus an extraction of visual information directly from the images is required, and is usually called *low-level features extraction*.

Unfortunately, bridging the gap between the target semantic classes and the available low-level visual descriptors is an unsolved problem. Hence it is crucial to select an appropriate set of visual descriptors that 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

[1] 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.

W. Duch et al. (Eds.): ICANN 2005, LNCS 3697, pp. 847–852, 2005.

in indoor/outdoor classification [2] to detect e.g. vegetation (green) or sea (blue). Edge direction histograms are employed for city/landscape classification [3] since city images typically contain horizontal and vertical edges. Additionally, motion descriptors are also used for sports video shot classification [4].

Nonetheless, the second crucial problem is to combine the low-level descriptors in such a way that the results obtained with individual descriptors are improved. The combination of features is performed before or at the same time as the estimation of the distances between images (*early fusion*)or directly at the matching scores (*late fusion*) [5].

In this work, fusion of several MPEG-7 descriptors is approached using three different machine learning techniques. A SVM is used with a "merging" descriptors' fusion, a Back-Propagation neural network is trained to estimate the distance between two images based on their low-level descriptors and a KNN Classifier is applied to evaluate the results. Finally in order to extract fuzzy rules and bridge low-level features with the semantics of images, a Falcon-ART Neurofuzzy Network is used.

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 along with the classification results. Finally conclusions are drawn in section 5.

2 Feature Extraction

In order to provide standardized descriptions of audio-visual (AV) content, MPEG-7 standard [6] 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 which are extracted using the aceToolbox, developed within the aceMedia project[7][2] and is based on the architecture of the MPEG-7 eXperimentation Model [8]. A brief overview of each descriptor is presented below, while more details can be found in [9].

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.

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

Edge Histogram Descriptor. (EHD) captures the spatial distribution of edges. Four directions of edges (0 °, 45 °, 90 °, 135 °) are detected in addition

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

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.

3 Image Classification Based on MPEG-7 Visual Descriptors

Several distance functions, MPEG-7 standardized or not, can be used when a single descriptor is considered. However, in order to handle all the above descriptors at the same time for tasks like similarity/distance estimation, feature vector formalization or training of classifiers, it is necessary to fuse the individual, incompatible elements of the descriptors, with different weights on each.

Three methods are considered for this purpose, combined with appropriate classification techniques. *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 extract semantic fuzzy rules.

3.1 Merging Fusion/SVM Classifier

In the first fusion strategy, all three descriptors are merged into a unique vector, thus is called *merging fusion*. If $D_{SCD}, D_{CLD}, D_{EHD}$ are the three descriptors referenced before then the merged descriptor is equal to:

$$D_{merged} = [D_{SCD}|D_{CLD}|D_{EHD}]$$

All features must have more or less the same numerical values to avoid scale effects. In our case, the MPEG-7 descriptors are already scaled to integer values of equivalent magnitude. A Support Vector Machine [10] was used to evaluate this fusion.

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. Its input consists of the low-level descriptions of two images and its output is the normalized estimation of their distance. The network is trained under the assumption that the distance of two images belonging in the same class is 0, otherwise, it is 1. These distances are used as input of 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. However,it is a priori set to zero. Moreover, even for a well-trained network, the output would be slightly different depending on the row they are presented, thus the distance matrix would not respect the symmetry property needed by the KNN

classifier. To overcome this, we used only the distances either of the upper or of the lower triangular matrix, or replacing a distance by the average of the two corresponding outputs of the neural network.

Another approach to efficiently fuse the different visual descriptors uses pre-calculated distance matrices for individual visual descriptors assigning weights on each one, to produce a weighted sum. This time, the input of the network consists of the three distances and results to a distance matrix which is used again as the input of a KNN classifier.

3.3 Classification Using a Falcon-ART Neurofuzzy Network

Image classification using a neural network or a SVM fails to provide semantic interpretation of the underlying mechanism that 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].

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 input of the network is a merged descriptor according to the process of section 3.1. After training, the network's response is the class that the input belongs. 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.

In order to have a description close to human perception for the rules of the Falcon-Art algorithm, each dimension of an image descriptor was divided into three equal parts, each one corresponding to *low*, *medium*, *high* values; each hyperbox created by the Falcon-ART then leads to a rule that uses these values. We present an example of such a rule, when classification considers only the EHD descriptor. 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 *Beach*

4 Experimental Results

The image database used for the experiments is part of the aceMedia content repository [3] and more specifically of the Personal Content Services database. It consists of 767 high quality images divided in two classes *beach* and *urban*. All the results are presented in table 1. 40 images from the *beach* category and 20

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

Fig. 1. Representative Images - 1-3:Beach Images, 4-6: Urban Images

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

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

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*

from the *urban* were selected and used as training dataset. The remaining 707 (406 from *beach* and 301 from *urban*) images were used for evaluation.

SVM Classifier using Merging Fusion: The merged vectors were directly used as input of a SVM classifier 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. 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: The distance between two images was determined manually and was set to 0 for images belonging to

the same category and to 1 otherwise. The symmetric distance matrices were used with the KNN classifier, as described in section 3. Best performance was achieved using all the descriptors and the distances between the images. In this case the success rate was 93.49%. All the results are shown on table 1.

Falcon-ART Neurofuzzy Network: The same 60 images' merged descriptions were presented randomly at the Falcon-ART neurofuzzy network. In the case of the EHD descriptor, the Falcon-ART has created 5 fuzzy rules which are presented in detail in table 2. 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.

5 Conclusion and Future Works

All 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 SVM. The Falcon-ART provided a linguistic description of the underlying classification mechanism. Future work will aim to use more MPEG-7 descriptors. 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. Szummer, M., Picard, R.: Indoor-outdoor image classification. In: IEEE international workshop on content-based access of images and video databases,. (1998)
3. Vailaya, A., Jain, A., Zhang, H.J.: On image classification: City images vs. landscapes. Pattern Recognition **31** (1998) 1921–1936
4. D.H. Wang, Q. Tian, S.G., Sung, W.K.: News sports video shot classification with sports play field and motion features. ICIP04 (2004) 2247–2250
5. Mc Donald, K., Smeaton, A.: A comparison of score, rank and probability-based fusion methods for video shot retrieval. In: CIVR. (2005) Singapore.
6. 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
7. Kompatsiaris, I., Avrithis, Y., Hobson, P., Strinzis, M.: Integrating knowledge, semantics and content for user-centred intelligent media services: the acemedia project, Proc. of WIAMIS 04, Portugal, April 21-23, 2004. (2004)
8. MPEG-7: Visual experimentation model (xm) version 10.0. ISO/IEC/ JTC1/SC29/WG11, Doc. N4062 (2001)
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. Vapnik, V.: The Nature of Statistical Learning Theory. NY:Springer-Verlag (1995)
11. Lin, C.T., Lee, C.S.G.: Neural-network-based fuzzy logic control and decision system. IEEE trans. Comput. **40** (1991) 1320–1336