

Combining MPEG-7 Based Visual Experts For Reaching Semantics

Medeni Soysal^{1,2} and A. Aydin Alatan^{1,2}

¹ Department of Electrical and Electronics Engineering, M.E.T.U.,

² TÜBİTAK BİLTEN,

Balgat, 06531, Ankara, Turkey

medeni.soysal@bilten, [alatan@eee}.metu.edu.tr](mailto:alatan@eee.metu.edu.tr)

Abstract. Semantic classification of images using low-level features is a challenging problem. Combining experts with different classifier structures, trained by MPEG-7 low-level color and texture descriptors is examined as a solution alternative. For combining different classifiers and features, two advanced decision mechanisms are proposed, one of which enjoys a significant classification performance improvement. Simulations are conducted on 8 different visual semantic classes, resulting in accuracy improvements between 3.5-6.5%, when they are compared with the best performance of single classifier systems.

1 Introduction

Large collections of digital multimedia data are used in various areas today [1]. This is an inevitable result of the technological advances that make it easy to create, store and exchange digital multimedia content. Most of this content is indexed by manual annotation of data during the process of input. Although manual annotation is inevitable for some cases, replacing it with automatic annotation whenever possible, lifts a great burden.

MPEG-7 standard comes up with many features, supporting both manual and automatic annotation alternatives. In this standard, although many detailed media descriptions for manual annotation exist, automatic annotation is encouraged by many audio-visual low-level descriptors. These low-level descriptions (features) can be extracted automatically from the data by using many state-of-the-art algorithms. In this context, the most challenging problem is to find some relations between these low-level features and high-level semantic descriptions, desired by typical users. Focusing on visual descriptors, some typical examples of such high-level visual semantic descriptions can be indoor, sea, sky, crowd, etc.

Classification using low-level descriptions is widespread in many different areas as well as image classification. However, utilization of standardized features like those in MPEG-7 and combining them are relatively new ideas in the area of image classification. The work presented reaches one step beyond the previous approaches, and performs supervised classification of still images into semantic classes by utilizing multiple standard-based features and various classifier structures concurrently in two different settings.

These settings, namely advanced decision mechanisms that are proposed in this paper are compared against common single classifier-single descriptor setting and some other techniques in terms of classification performances. In this way, interesting and important relations are revealed.

The paper is organized as follows. In Section 2, low-level image features that are used in classification are introduced. Classifier types used and modifications on them are explained in Section 3. Various well-known basic methods to combine the experts, which use the features explained in Section 2 and have one of the classifier structures in Section 3, are discussed in Section 4. Two advanced decision mechanisms are developed in Section 5, as an alternative to the classical methods. These proposed decision mechanisms are compared with the best performance of single experts experimentally in Section 6. Section 7 summarizes the main results of the work and offers concluding remarks.

2 Low-Level Image Features

Successful image classification requires a good selection among low-level representations (i.e. features). In this research, color and texture descriptors of MPEG-7 [2] are utilized. A total of 4 descriptors are used, while two of them (color layout and color structure) are color-based, the other two (edge histogram and homogeneous texture) are texture descriptors.

MPEG-7 Color Layout descriptor is obtained by applying DCT transformation on the 2-D array of local representative colors in YCbCr space. Local representative colors are determined by dividing the image into 64 blocks and averaging 3 channels on these blocks. After DCT transformation, a nonlinear quantization is applied and first few coefficients are taken. In these experiments, only 6 coefficients for luminance and 3 coefficients for each chrominance are used, respectively [2].

MPEG-7 Color Structure descriptor specifies both color content (like color histogram) and the structure of this content by the help of a structure element [2]. This descriptor can distinguish between two images in which a given color is present in identical amounts, whereas the structure of the groups of pixels is different.

Spatial distribution of edges in an image is found out to be a useful texture feature for image classification [2]. The edge histogram descriptor in MPEG-7 represents local edge distribution in an image by dividing the image into 4x4 sub-images and generating a histogram from the edges present in each block. Edges in the image are categorized into five types, namely, vertical, horizontal, 45° diagonal, 135° diagonal and non-directional edges. In the end, a histogram with 16x5=80 bins is obtained, corresponding to a feature vector with 80 dimensions.

MPEG-7 Homogeneous Texture descriptor characterizes the region texture by mean energy and energy deviation from a set of frequency channels. The channels are modeled by Gabor functions and the 2-D frequency plane is portioned into 30 channels. In order to construct the descriptor, the mean and the standard deviation of the image in pixel domain is calculated and combined into a feature vector with the mean and energy deviation computed in each of the 30 frequency channels. As a result, a feature vector of 62 dimensions is extracted from each image [2].

3 Classifiers

In this research, 4 classifiers are utilized, which are Support Vector Machine [8], Nearest Mean, Bayesian Plug-In and K-nearest neighbors [4]. Binary classification is performed by experts obtained via training these classifiers with in-class and informative out-class samples. These classifiers are selected due to their distinct natures of modeling a distribution. For distance-based classifiers (i.e. Nearest Mean and K-Nearest Neighbor) special distance metrics compliant with the nature of the MPEG-7 descriptors are utilized. Since the outputs of the classifiers are to be used in combination, modifications are achieved on some of them to convert uncalibrated distance values to the calibrated probability values in the range [0,1]. All of these modifications are explained in detail along with the structure of the classifiers in the following subsections.

3.1. Support Vector Machine (SVM)

SVM performs classification between two classes by finding a decision surface via certain samples of the training set. SVM approach is different from most classifiers in a way that it handles the risk concept. Although other classical classifiers try to classify training set with minimal errors and therefore reduce the empirical risk, SVM can sacrifice from training set performance for being successful on yet-to-be-seen samples and therefore reduces structural risk [8]. Briefly, one can say that SVM constructs a decision surface between samples of two classes, maximizing the margin between them. In this case, a SVM with second-degree polynomial kernel is utilized. SVM classifies any test data by calculating the distance of samples from the decision surface with its sign signifying which side of the surface they reside.

On the other hand, in order to combine the classifier outputs, each classifier should produce calibrated posterior probability values. In order to obtain such an output, a simple logistic link function method, proposed by Wahba [5] is utilized as below.

$$P(\text{in - class} | x) = \frac{1}{1 + e^{-f(x)}} \quad (1)$$

In this formula, $f(x)$ is the output of SVM, which is the distance of the input vector from the decision surface.

3.2. Nearest Mean Classifier

Nearest mean classifier calculates the centers of in-class and out-class training samples and then assigns the upcoming samples to the closest center. This classifier again, gives two distance values as output and should be modified to produce a posterior probability value. A common method used for K-NN classifiers is utilized in this case [6]. According to this method, distance values are mapped to posterior probabilities by the formula,

$$P(w_i | x) = \frac{1}{d_{m_i}} / \sum_{j=1}^2 \frac{1}{d_{m_j}} \quad (2)$$

where d_{m_i} and d_{m_j} are distances from the i^{th} and j^{th} class means, respectively. In addition, a second measure recomputes the probability values below a given certainty threshold by using the formula [6]:

$$P(w_i | x) = \frac{N_i}{N} \quad (3)$$

where N_i is the number of in-class training samples whose distance to the mean is greater than x , and N is the total number of in-class samples. In this way, a more effective nearest mean classifier can be obtained.

3.3. Bayesian Gaussian Plug-In Classifier

This classifier fits multivariate normal densities to the distribution of the training data. Two class conditional densities representing in-class and out-class training data are obtained as a result of this process [4]. Bayesian decision rule is then utilized to find the probability of the input to be a member of the semantic class.

3.4. K-Nearest Neighbor Classifiers (K-NN)

K-NN classifiers are especially successful while capturing important boundary details that are too complex for all of the previously mentioned classifiers. Due to this property, they can model sparse and scattered distributions with a relatively high accuracy.

Generally, the output of these classifiers are converted to probability, except for K=1 case, with the following formula:

$$P(w_i | x) = K_i / K \quad (4)$$

where K_i shows the number of nearest neighbors from class- i and K is the total number of nearest neighbors, taken into consideration. This computation, although quite simple, underestimates an important point about the location of the test sample relative to in-class and out-class training samples. Therefore, instead of the above method, a more complex estimation is utilized in this research:

$$P(w_i | x) = \sum_{y_j} \frac{1}{d(x, y_j)} / \sum_{i=1}^k \frac{1}{d(x, y_i)} \quad (5)$$

where y_j shows in-class nearest neighbors of the input and y_i represent all k -nearest neighbors of the input.

Although, this estimation provides a more reliable probability output, it is observed that applying another measure to the test samples with probabilities obtained by (5) below a threshold also improves the result. This measure utilizes the relative positions of training data among each other [6]. This metric is the sum of the distances of each in-class training sample to its k in-class nearest neighbors:

$$g(x) = \sum_{i=1}^k d(x, y_i) \quad y_i: i^{\text{th}} \text{ in-class nearest neighbor} \quad (6)$$

After this value is computed for each training sample and input test sample, the final value is obtained by,

$$P(\text{in-class} | x) = 1 - (N_i / N) \quad (7)$$

where N_i is the number of in-class training samples with $g(x)$ value smaller than the input test sample and N is the number of all n -class training samples. In this way, a significant improvement is achieved in 3-NN, 5-NN, 7-NN and 9-NN classifier results.

For 1-NN case, since the conversion techniques explained here are not applicable, the probability estimation technique employed in the case of nearest mean classifier is applied.

4 Expert Combination Strategies

Combining *experts*, which are defined as the instances of classifiers with distinct natures working on distinct feature spaces, has been a popular research topic for years. Latest studies have provided mature and satisfying methods. In this research, six popular techniques, details of which are available in literature are adopted [3]. In all of these cases, a priori probabilities are assumed as 0.5 and the decision is made by the following formula:

$$P(\text{in-class} | X) = \frac{P_1}{P_1 + P_2} \quad (8)$$

Here, P_1 is the combined output of experts about the likelihood of the sample X belonging to the semantic class while P_2 is the likelihood for X not belonging to the semantic class. Decision is made according to the Bayes' rule; if the likelihood is above 0.5, the sample is assigned as in-class, else out-class. P_1 and P_2 are obtained by using combination rules, namely, *product rule*, *sum rule*, *max rule*, *min rule*, *median rule* and *majority vote* [3]. In product rule, R experts are combined as follows,

$$P_1 = \prod_{i=1}^R P_i(\text{in-class} | X) \quad P_2 = \prod_{i=1}^R P_i(\text{out-class} | X) \quad (9)$$

Similarly, sum rule calculates the above probabilities as,

$$P_1 = \sum_{i=1}^R P_i(\text{in-class} | X) \quad P_2 = \sum_{i=1}^R P_i(\text{out-class} | X) \quad (10)$$

Others, which are derivations of these two rules, perform the same calculation as follows:

$$\text{Max Rule} \quad P_1 = \max_{i=1}^R P_i(\text{in-class} | X) \quad P_2 = \max_{i=1}^R P_i(\text{out-class} | X) \quad (11)$$

$$\text{Min Rule} \quad P_1 = \min_{i=1}^R P_i(\text{in-class} | X) \quad P_2 = \min_{i=1}^R P_i(\text{out-class} | X) \quad (12)$$

$$\text{Median Rule} \quad P_1 = \text{med}_{i=1}^R P_i(\text{in-class} | X) \quad P_2 = \text{med}_{i=1}^R P_i(\text{out-class} | X) \quad (13)$$

Lastly, majority vote (MV) counts the number of experts with in-class probabilities higher than 0.5 and assigns to P_1 . P_2 is the number of voting experts minus P_1 .

$$\text{Majority Vote} \quad P_1 = \frac{N_1}{N_1 + N_2} \quad P_2 = \frac{N_2}{N_1 + N_2} \quad \begin{array}{l} N_1 : \# \text{ experts voting in-class} \\ N_2 : \# \text{ experts voting out-class} \end{array} \quad (14)$$

5 Advanced Decision Mechanisms

In order to improve the classification performance, which is achieved by expert combination strategies, two different advanced mechanisms are implemented. These mechanisms, namely Multiple Feature Direct Combination (MFDC) and Multiple Feature Cascaded Combination (MFCC), use the output of single experts in two different ways. They are applied only to semantic classes, for which more than one of the low-level features are required. In these mechanisms, only five types of experts are involved, leaving out 3-NN, 7-NN and 9-NN type experts, to prevent the dominance of K-NN. These experts are based on SVM, Nearest Mean, Bayesian Gaussian Plug-in, 1-NN and 5-NN classifiers.

MFDC mechanism combines output of sole experts, which are trained by all low-level features, in a single step. For instance, $3 \times 5 = 15$ experts will be combined for a class that is represented by three different low-level features. In MFCC case, Single Feature Combination (SFC) outputs are utilized. SFC combines experts trained by the same low-level feature and gives a single result. Next, MFCC uses the output of each SFC to generate a resultant in-class probability. These two mechanisms are illustrated in Figure 1.

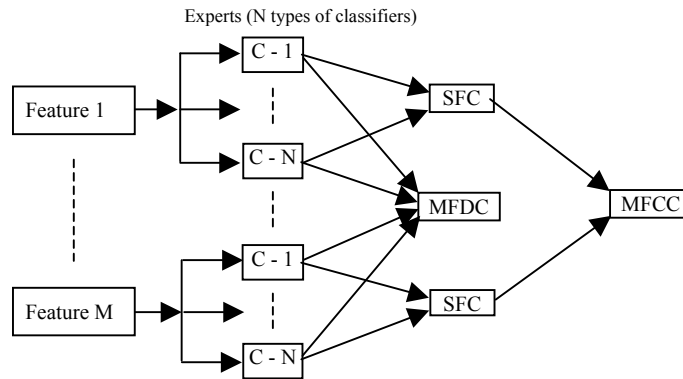


Fig. 1. Single Feature Combination(SFC) and Advanced Decision Mechanisms, Multiple Feature Direct Combination(MFDC), Multiple Feature Cascaded Combination(MFCC)

6 Implementation Issues

A total of 1600 images, collected from various sources and having different resolutions, are used for training and test phases. A total of eight semantic classes are classified. For each class, 100 in-class and 100 out-class samples are used. Boosting [4] is used to prevent dependence of results on images. Five tests are performed by taking 20 distinct samples from each of in-class and out-class data sets, and training the experts by remaining classified data consisting of 80 in-class and 80 out-class training samples. Results are evaluated by considering the average of these five tests.

The semantic classes selected for these tests have the common property of being convenient to be inferred from low-level visual features extracted from the entire image. This means that, the characteristics of these classes are usually significant in the entire image and therefore the need for segmentation is mostly avoided.

Eight classes that are subjects of the tests are *football*, *indoor (outdoor)*, *crowd*, *sunset-sunrise*, *sky*, *forest*, *sea* and *cityscape*. For each class, MPEG-7 color and texture descriptors that proved to capture the characteristics best in the pre-experiments, are utilized. The corresponding features to classifying these classes are tabulated in Table 1.

Table 1. Semantic classes and related features

Semantic Class	Low-Level Features
Football	Color Layout
Indoor	Edge Histogram
Crowd	Homogeneous Texture
Sunset-Sunrise	Color Layout, Color Structure, Edge Histogram
Sky	Color Layout, Color Structure, Homogeneous Texture
Forest	Color Structure, Edge Histogram, Homogeneous Texture
Sea	Color Layout, Homogeneous Texture
Cityscape	Color Structure, Edge Histogram, Homogeneous Texture

Table 2. Performances of SFC v.s. single experts

		Max Single	Single Feature Comb. (SFC)					
			Prd	Sum	Max	Min	Med	MV
Football	Accuracy	91.0	87.5	89.5	87.5	87.5	90.0	91.0
	Precision	91.6	98.8	98.8	97.3	97.3	98.8	92.7
	Recall	91.0	76.0	80.0	77.0	77.0	81.0	89.0
Indoor	Accuracy	83.0	84.0	83.0	83.5	83.5	81.0	84.0
	Precision	81.1	91.3	91.0	88.3	88.3	90.8	90.4
	Recall	87.0	75.0	73.0	77.0	77.0	69.0	76.0
Crowd	Accuracy	79.5	75.5	81.0	77.0	77.0	79.0	78.5
	Precision	83.5	72.5	81.8	73.3	73.3	81.6	79.6
	Recall	73.0	84.0	80.0	87.0	87.0	75.0	77.0

7 Experimental Results

Combination of experts has been tested on eight semantic classes. For the first three of these classes (*football*, *indoor* and *crowd*), only one representative low-level feature is used and therefore only Single Feature Combination (SFC) is available. Other five classes (*sunset-sunrise*, *sky*, *forest*, *sea* and *cityscape*) are represented by multiple features and therefore advanced decision mechanisms (MFDC and MFCC) are also applicable. Performances of the techniques on these two sets of classes are presented separately in different tables, Table 2 and Table 3, respectively. In order to provide a good basis of comparison, for each class, the result of an “optimal combination formula” which is obtained by combining experts with the best results, is also included. Obviously, such a case is not practical, since it should be determined case-by-case basis for each class.

In this section, although the accuracy results are used for comparison, precision and recall results are also included in the tables. This is because of the fact that they convey information about different properties of the techniques, which is hidden in accuracy.

Table 3. Performances of single experts, SFC, MFDC, and MFCC on different classes.

		Max Single	Max SFC	MFCC						MFDC						Optimal Comb. Formula	
				Prd	Sum	Max	Min	Med	MV	Prd	Sum	Max	Min	Med	MV		
Sunset Sunrise	Accuracy	92.5	92.0	92.5	90.0	92.0	92.0	90.0	90.0	92.5	91.0	84.5	91.0	91.0	93.5	Prd CSD-1NN CSD-5NN	
	Precision	90.9	88.8	93.5	91.2	92.6	92.6	91.2	91.2	93.5	93.1	82.2	91.4	90.6	92.5		
	Recall	95.0	97.0	92.0	89.0	92.0	92.0	89.0	89.0	92.0	89.0	91.0	91.0	92.0	92.0		95.0
Sky	Accuracy	93.0	92.5	96.0	96.5	94.0	95.0	96.5	96.5	96.0	97.0	83.0	88.5	95.5	96.0	Sum CSD-5VM CSD-1NN HTD-5NN	
	Precision	89.0	92.3	94.5	94.7	94.5	95.4	94.7	94.7	94.5	95.4	86.4	97.6	92.2	94.5		
	Recall	100.0	94.0	98.0	99.0	94.0	95.0	99.0	99.0	98.0	99.0	79.0	79.0	100.0	100.0		98.0
Forest	Accuracy	79.0	82.0	86.5	86.0	85.0	85.0	85.5	85.5	86.5	84.5	78.0	83.5	83.0	85.0	Max CSD-5NN EHD-1NN HTD-5VM	
	Precision	78.4	84.6	85.7	84.3	84.3	84.3	84.1	84.1	85.7	84.0	75.3	84.1	81.0	83.8		
	Recall	81.0	80.0	90.0	90.0	88.0	88.0	89.0	89.0	90.0	87.0	86.0	86.0	88.0	88.0		88.0
Sea	Accuracy	80.5	83.0	86.0	86.0	86.0	86.0	60.0	86.0	84.5	74.0	79.0	82.5	81.0	85.5	Prd CLD-Med HTD-Med	
	Precision	75.8	81.8	89.0	89.0	89.0	89.0	89.0	56.0	89.0	89.7	73.3	83.5	88.6	94.5		
	Recall	93.0	85.0	84.0	84.0	84.0	84.0	84.0	64.0	84.0	80.0	75.0	75.0	77.0	79.0		76.0
Cityscape	Accuracy	82.0	81.5	85.0	86.5	82.0	82.0	87.0	87.0	85.0	83.5	71.0	77.0	81.0	85.5	Med CSD-5VM EHD-Bayes HTD-Bayes	
	Precision	82.6	81.9	84.9	86.0	83.5	83.5	86.1	86.1	84.9	82.9	74.1	84.3	78.9	88.0		
	Recall	81.0	81.0	86.0	87.0	81.0	81.0	88.0	88.0	86.0	84.0	67.0	67.0	85.0	85.0		84.0

For the classes in Table 2, it is seen that SFC leads with at least one rule except for the *football* case. However, improvements are not significant and also performance depends on the choice of the best combination for each of the above classes. For *football*, the majority vote rule gives the same result (% 91) with the best expert, which is a 1-NN. *Indoor* class is classified slightly better than the best expert (% 83) by product and majority vote results (% 84). In *crowd* classification, sum rule reached 81% and beat 9-NN classifier, whose performance was 79.5%.

Significant improvements are observed in the cases, where the proposed advanced decision mechanisms are applicable. MFDC and MFCC outperform the best single expert and best SFC for nearly all classes. The only case in which advanced decision mechanisms do not yield better results than the best single expert is *sunset (sunrise)* classification.

MFDC though being successful against single experts, could not beat the “optimal combination formula” in most of the cases. However, the “optimal combination formula” gives inferior results against MFCC for the most cases. For instance, MFCC improves the performance of classifications, especially when its second stage combination rule is fixed to median, while SFCs in the previous stage are obtained by the product rule. This should be due to the fact that these two rules have properties, which compensate the weak representations of each other. Product rule, although known to have many favorable properties, is a “severe” rule, since a single expert can inhibit the positive decision of all the others by outputting a close to zero probability [3]. Median rule, however, can be viewed as a robust average of all experts and is therefore more resilient to this weakness belonging to the product rule. This leads us to the observation that combining the product rule and the median rule is an effective method of increasing the modeling performance. This observation on MFCC is also supported by a performance improvement of 3.5% for *sky*, 6.5% for *forest*, 5.5% for *sea* and 5% for *cityscape* classification, when it is compared against the best single classifier. MFCC also achieves a performance improvement of at least 1-2% over even the manually selected “optimal combination formula”.

Another important fact about the performances achieved in classification of these classes using advanced decision mechanisms is the increase in precision values they provide. In the application of classification of these methods to large databases with higher variation compared with data sets used in experiments, usually recall values are sustained, however precision values drop severely. The methods proposed in this text, therefore have also an effect of increasing robustness of classification.

In addition, although the averages of the test sets are displayed for each class, when the separate test set performances are analyzed, MFCC shows quite stable characteristics. The variance of its performance from one test set to another is less than all others. Typical classification results can also be observed at our ongoing MPEG-7 compliant multimedia management system site, Bi1VMS (<http://vms.bilten.metu.edu.tr/>).

8 Conclusion

Reaching semantic information from low-level features is a challenging problem. Most of the time, it is not enough to train a single type of classifier with a single low-level feature to define a semantic class. Either it is required to use multiple features to represent the class, or it is needed to combine different classifiers to fit a distribution to the members of the class in the selected feature space.

Advanced decision mechanisms are proposed in this paper, and among the two methods, especially, Multiple Feature Cascaded Combination (MFCC) achieves significant improvements, even in the cases where single experts have already had very high accuracies. The main reason for this improvement is the reliability and stability the combination gains, since experts that are good at modeling different parts of the class distribution are combined to complement each other. For MFCC, it is observed that classification performance significantly improves, when correct combination rules are selected at each stage. For instance, combining the product rule results of the first stage by using median rule is found out to be quite successful in all cases. This observation can be explained by the complementary nature of the rules.

References

1. Forsyth, D.A.: Benchmarks for Storage and Retrieval in Multimedia Databases. Proc. Of SPIE, Vol. 4676 SPIE Press, San Jose, California (2002) 240-247
2. Manjunath, B.S., Salembier, P., Sikora, T.: Introduction to MPEG-7. John Wiley&Sons Ltd. England (2002)
3. Kittler, J., Hataf, M., Duin, R.P.W., Matas, J.: On Combining Classifiers. IEEE Trans. PAMI Vol. 20. No. 3. Mar. 1998.(1998) 226-239
4. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. John Wiley&Sons Ltd. Canada (2001)
5. Platt, J.C.: Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In: Advances in Large Margin Classifiers. MIT Press. Cambridge. MA (1999)
6. Arlandis, J., Perez-Cortes, J.C., Cano, J.: Rejection Strategies and Confidence Measures for a k-NN Classifier in an OCR Task. IEEE (2002)
7. Tong, S., Chang, E.: Support Vector Machine Active Learning for Image Retrieval. Proc. ACM. Int. Conf. on Multimedia. New York (2001) 107-118
8. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag. New York (1995)