FACTA UNIVERSITATIS
Series: **Electronics and Energetics** Vol. 27, N^{o} 3, September 2014, pp. 399 - 410
DOI: 10.2298/FUEE1403399C

EFFECTIVE COMBINING OF COLOR AND TEXTURE DESCRIPTORS FOR INDOOR-OUTDOOR IMAGE CLASSIFICATION

Stevica Cvetković[1], Saša V. Nikolić[1], Slobodan Ilić[2]

[1]University of Niš, Faculty of Electronic Engineering, Niš, Serbia
[2]Technische Universitat München (TUM), Munich, Germany

Abstract. *Although many indoor-outdoor image classification methods have been proposed in the literature, most of them have omitted comparison with basic methods to justify the need for complex feature extraction and classification procedures. In this paper we propose a relatively simple but highly accurate method for indoor-outdoor image classification, based on combination of carefully engineered MPEG-7 color and texture descriptors. In order to determine the optimal combination of descriptors in terms of fast extraction, compact representation and high accuracy, we conducted comprehensive empirical tests over several color and texture descriptors. The descriptors combination was used for training and testing of a binary SVM classifier. We have shown that the proper descriptors preprocessing before SVM classification has significant impact on the final result. Comprehensive experimental evaluation shows that the proposed method outperforms several more complex indoor-outdoor image classification techniques on a couple of public datasets.*

Key words: *feature extraction, image classification, image color analysis, image edge detection, support vector machines.*

1. INTRODUCTION

Indoor-outdoor image classification is a problem that attracts considerable attention of scientific population involved in content based image retrieval [1]. It is a restricted case of the general image classification problem, which represents the basis for decision of further processing steps depending on the scene type. For instance, assumption that indoor and outdoor images are usually taken under different illumination conditions can be used for decision about forthcoming color correction approach [2]. Furthermore, indoor-outdoor classification can be exploited for many image processing applications such as image orientation detection [3], image retrieval [4], or robotics [5].

Received January 28, 2014; received in revised form March 28, 2014

Corresponding author: Stevica Cvetković
University of Niš, Faculty of Electronic Engineering
(e-mail: stevica.cvetkovic@elfak.ni.ac.rs)

Several approaches for indoor-outdoor image classification have been proposed in the literature so far. In the recent methods there is an evident trend of introduction of additional information about camera or scene [6], [7]. However, this kind of information which includes exposure time, object distance or flash fired info, is commonly unavailable to the system. There is also evident involvement of domain specific assumptions about indoor-outdoor scenes, such as presence of sky or grass in outdoor images [8], [9], or intentional favoring of specific image partitions [10]. In our work, the goal is to define an effective and efficient method for indoor-outdoor image classification based on standardized low-level descriptors and proven machine learning techniques. Another goal is to achieve sufficient generality of the method and to avoid introduction of domain specific knowledge about the scene. To this end we propose a carefully engineered procedure for the composition of MPEG-7 color and texture descriptors characterized by efficient extraction, compact representation and high discriminative power. We conducted comprehensive empirical tests to determine the optimal combination of descriptors for this purpose. After important preprocessing procedure, combined descriptors will be conducted to input of an optimally tuned Support Vector Machines (SVM) classifier. We have empirically shown a large impact of feature preprocessing before SVM on the accuracy of the system. Experimental evaluation will show that the proposed method outperforms some more complex state-of-the-art indoor-outdoor classification techniques on a couple of standard datasets.

To be successfully applied for the image classification task, an image descriptor (feature) should be highly discriminative and invariant against image content [11]. Such a descriptor should generate features with high variance and good distribution over category samples. In addition, it should be robust to different levels of image quality and resolution. There is a large collection of visual descriptors available in the literature with corresponding strengths and weaknesses [12], [13]. In order to provide standardized descriptors of image and video content, MPEG-7 standard defines three classes of still image visual descriptors: color, texture and shape descriptors [14], [15]. Each class of visual features characterizes only a certain aspect of image content, so the combination of features is necessarily employed to provide an appropriate description of image content. We performed exhaustive experiments on combining several MPEG-7 color and texture descriptors which were carefully chosen to meet the requirements of fast calculation, compact representation and high discriminative power.

Once features have been extracted, method for automatic image classification should be applied. Approaches for image classification can be roughly grouped in two categories: (a) Learning-based methods that are able to learn optimal parameters based on input training samples. These methods include SVM [6], [8], [10], [16]-[18], Neural Networks [19], [20], decision trees [2], Hidden Markov Models [36], etc. (b) Non-parametric methods that perform classification directly on the data, without learning the parameters. The most widely used non-parametric method is k-Nearest Neighbors (k-NN) which determines image class based on the class of its most similar images [4], [21], [22]. Although non-parametric methods require no learning steps and are able to naturally handle a large number of classes, they often suffer from high variation along the decision boundary caused by finite sampling in terms of bias-variance decomposition [23]. As a consequence, their accuracy could be inferior compared to learning-based methods [24]. In addition, processing time of non-parametric methods is considerably larger than the learning-based methods, which makes them inconvenient for large scale classification systems. For our purpose we propose to apply binary SVM image classifier on preprocessed feature vectors formed by combination of several visual descriptors. We applied classification over low-level features only, i.e. features that can be automatically extracted without any a priori knowledge of the image content.

Although our paper does not involve fundamentally new procedures, it has several main contributions: (1) it has empirically shown that a baseline method for indoor-outdoor image classification can reach satisfactorily good results without using complex image descriptors and sophisticated machine learning techniques, (2) gives extensive review on indoor-outdoor image classification topic that is lacking in the literature, and (3) provides comprehensive statistical analysis of descriptor combination with feature scaling methods in order to demonstrate importance of each component.

In the rest of the paper we first give overview of the previous research in the field of indoor-outdoor image classification. In Section 3, the reasons for choosing specific MPEG-7 descriptors are explained including a brief overview of feature extraction procedures as well as strategy for their combination. Then, details of SVM image classification method are presented, including description of feature preprocessing step and SVM parameters selection. Finally, testing methodology and results are presented and discussed.

2. PREVIOUS RESEARCH

The research on indoor-outdoor image classification can be traced back to the work of Szummer and Picard [21] who applied a two-stage classification approach on features that combine Ohta color space histogram and multi-resolution simultaneous autoregressive model (MSAR). At the first stage, they used k-NN to classify subblocks of the image, while the final decision was based on the majority rule. The accuracy of 90.3% is achieved on a set of over 1300 consumer images. In a similar approach, Serrano et al. [18], extracted LST color histogram and Wavelet texture features for classification of image sub-blocks. They used linear SVM classifiers to train color and texture features separately. Recognition rate was 90.2%, on a set of 1200 images. In [4], indoor-outdoor classification is proposed at the highest level of a hierarchical image classification method. Color moments in the LUV color space were computed for 10x10 image subblocks. Concatenation of the feature vectors of all subblocks produced final feature vector. Finally, k-NN classifiers have been evaluated on a database of 6931 vacation photographs achieving accuracy of 90.5%. Straight edges were used as a feature in a method proposed in [22]. The authors claim that the proportion of straight edges in indoor images is larger in comparison to outdoor images. The final classification of the image is based on a k-NN rule applied to the proportion of straight edges contained in sub-blocks of the image. In addition, a multi-resolution estimates are used to improve the results. Tests conducted on a set of 872 photographs, reported classification accuracy of 90.71%. Gupta et al. [19] use a fuzzy clustering method to initially segment an image into sub-regions. Segments are then described using simple color, texture, and shape features. The probabilistic neural network is finally applied for the classification that reported accuracy of 92.36% on a benchmark set of 902 images. Indoor-outdoor classification is used to improve automatic illuminant estimation in [2]. The feature vector consists of color, texture and edge information. Decision forests of classification and regression trees are used for classification. Testing was performed on a collection of 6785 images, downloaded from the web or acquired by digital cameras. They reported a classification accuracy of 93.1%.

Table 1 Chronological overview of the published methods for indoor-outdoor image classification

Authors	Year	Classifier	Accuracy (in %)	Number of images
Szummer & Picard [21]	1998.	kNN	90.30	1343
Vailaya et al. [4]	2001.	kNN	90.50	6931
Serrano et al. [18]	2004.	SVM	90.20	1200
Serrano et al. [8]	2004.	SVM+Bayesian	90.70	1200
Payne & Singh [22]	2005.	kNN	90.71	872
Boutell & Luo [6]	2005.	SVM+Bayesian	94.10	5120
Liu et al. [7]	2005.	LDA+Boosting	92.20	13000
Lu et al. [9]	2005.	GMM+LDA	93.80	1400
Gupta et al. [19]	2007.	Neural Network	92.36	902
Bianco et al. [2]	2008.	Decision Forest	93.10	6785
Kim et al. [10]	2010.	SVM	90.26	1276

All of the previously mentioned methods are concerned with low-level features only that are extracted directly from digital images with no impact of human perception. In addition, there have been proposed several methods based on high-level image information, i.e. semantic assumptions about the scene. Authors of [18] have extended their approach in [8] by introducing semantic detectors for grass and sky to improve the classification accuracy. The results from the SVM sub-block classification and semantic detectors were integrated using a Bayesian classifier. A classification accuracy of 90.7% was reported on a set of 1200 consumer images. Boutell and Luo [6] proposed a fusion of low-level image information with the camera metadata information provided in exchangeable image file format (EXIF), such as exposure time, flash fired and subject distance. First, they applied SVM to classify images by low-level for color and texture-features. Then, Bayesian network was used to classify low-level features integrated with EXIF metadata. On a benchmark set of 5120 images, the reported accuracy was 94.1%. In a similar approach [7], the combination of color moments and edge direction histogram was extracted as low-level features. To improve the classification accuracy, they utilized 14 EXIF features associated with images. Linear discriminant analysis (LDA) algorithm was utilized to implement linear combinations between all extracted features. Finally, the combined features are used with the original features in boosting classification algorithm. On a large set of about 13000 digital photographs, they achieved 92.2% accuracy. Authors of [9] first trained Gaussian Mixture Models (GMM) to describe the color-texture properties of image patches for 20 predefined materials (building, blue sky, bush, etc.). These models are then applied to a test image to produce 20 probability density response maps which are later used to train LDA classifiers for scene categories. A database of 1400 photos taken from 43 persons was used for testing. The indoor-outdoor classification rate was 93.8%. In [10], authors make an assumption that foreground objects (human bodies and faces), which often appear in the central part of the image, may negatively affect the system performance. They partitioned the image into five blocks and extracted edge and color orientation histogram (ECOH) for each block. Then, the features are weighted according to the block positions (central part is less weighted) and concatenated to generate the final feature vector. SVM classifier evaluated on 1276 images obtained the 90.26% classification rate.

3. Feature Extraction and Combination

Statistical analysis of global visual descriptors for image retrieval in [11], [12], has shown a large overlapping of the information they extract. Although MPEG-7 suggests a number of different descriptors, most of them are highly dependent on each other [11]. On the other side, some of them like Dominant Color Descriptor (DCD), are too computationally expensive for practical applications. To extract features with compact representation and low computational costs we considered the following five MPEG-7 descriptors: Scalable Color Descriptor (SCD), Color Structure Descriptor (CSD), Color Layout Descriptor (CLD), Homogeneous Texture Descriptor (HTD), and Edge Histogram Descriptor (EHD). In order to test their behavior and make a selection of the best descriptors for combining, we have conducted exhaustive experiments. We will show that a combination of only a few of them is sufficient for successful indoor-outdoor image classification. First, we give a brief overview of the selected descriptors and describe a method that we used for their combination. Further details about descriptor extraction procedures could be found in [14], [25], [26].

Color descriptors

Scalable Color Descriptor (SCD) measures color distribution over an entire image. It is a histogram in the HSV color space that is encoded using the Haar transform. The histogram is extracted in HSV space uniformly quantized to 16 levels of H, 4 levels of S and 4 levels of V, giving 256 bins in total. These values are truncated into an 11 bit integer representation and non-linearly mapped into a 4-bit representation. This representation gives a higher significance to smaller values with high probability. To reduce the size of this representation, the histogram values are encoded using Haar transform. Its representation is scalable in terms of a coefficient number varying from 16 up to 256. Our experiments have shown that using more than 64 coefficients does not necessarily lead to a significant accuracy improvement. Therefore, we used the following representation of the descriptor

$$\mathbf{f}^{SCD} = (f_1^{SCD},..., f_{64}^{SCD}) \tag{1}$$

Color Structure Descriptor (CSD) extends the image color histogram with information about local spatial structure of the color. It is based on the concept of Color Structure Histogram which counts the number of times a particular color is contained within the 8x8 window as the window scans over the image. MPEG-7 specific color space, denoted as HMMD [14], is used for the extraction. It is first non-uniformly quantized into N colors [27], determining the number of bins in the Color Structure Histogram. Then, the window scans over the entire image, and for each color which is present within the window, it increments a corresponding histogram bin. Finally, histogram values are normalized and nonlinearly quantized to 8 bits/bin. In our experiments we used CSD containing 64 bins:

$$\mathbf{f}^{CSD} = (f_1^{CSD},..., f_{64}^{CSD}) \tag{2}$$

Color Layout Descriptor (CLD) has been designed to efficiently represent spatial layout of colors inside an image. It is obtained by applying the Discrete Cosine Transformation (DCT) on local representative colors of 64 image blocks in YCbCr color space. The descriptor is characterized by compact representation, invariance to resolution changing and low computational complexity. The extraction process starts with image partitioning of each RGB color channel into 8x8=64 non-overlapping blocks to guarantee resolution invariance. Then, a single representative color is computed for each block by simple pixel averaging. In the next step, conversion to YCbCr color space is done and color channels are transformed by DCT to obtain three sets of 64 DCT coefficients. Finally, a zigzag scanned DCT coefficients are concatenated into a feature vector containing the most informative elements of each YCbCr color channel. Our rough experiment has shown that the feature vector with 22 elements represents a good choice:

$$\mathbf{f}^{CLD} = (f_{Y1}^{CLD},..., f_{Y10}^{CLD}, f_{Cb1}^{CLD},..., f_{Cb6}^{CLD}, f_{Cr1}^{CLD},..., f_{Cr6}^{CLD}) \tag{3}$$

Texture descriptors

Homogeneous Texture Descriptor (HTD) characterizes the region texture by the mean energy and the energy deviation from a set of 30 frequency channels. The descriptor is extracted by first partitioning the frequency space into 30 equidistant channels. The individual feature channels are filtered with a bank of 2-D Gabor functions, and the mean and standard deviation of the energy in each of the channels is calculated. The final form of the descriptor that consists of 62 coefficients is

$$\mathbf{f}^{HTD} = (f_{DC}^{HTD}, f_{SD}^{HTD}, e_1^{HTD},..., e_{30}^{HTD}, d_1^{HTD},..., d_{30}^{HTD}) \tag{4}$$

The first two components are the mean and standard deviation of the complete image, and e_i^{HTD} and d_i^{HTD} are mean energy and energy deviation of the corresponding *i*-th frequency channel, respectively.

Edge Histogram Descriptor (EHD) represents spatial distribution of 5 types of edge orientations inside local image partitions called sub-images. One local edge histogram is generated for each of 4×4=16 subimages, representing distribution of five edge orientations inside a subimage. To generate the local edge histogram, edges in the sub-image are extracted and classified into five categories depending on the orientation (vertical, horizontal, diagonal- , diagonal- , and non-directional). Since there are 16 subimages, final edge histogram will contain 16x5 = 80 bins formed by concatenation of the local histograms

$$\mathbf{f}^{EHD} = (f_1^{EHD},..., f_{80}^{EHD}) \tag{5}$$

Combination of descriptors

When using multiple visual features for image classification, crucial problem is how to combine them in order to measure image similarity. Generally, there are two approaches for feature combination (fusion, aggregation, composition, merging) [17], [28], [34]. The first one, named “early fusion” performs a combination of features before the estimation of the distances between images. In contrast, “late fusion” applies classification on each feature separately, after which it integrates these results into final decision.

An obvious disadvantage of late fusion approach is its computational expensiveness, as every feature requires separate classification stage. Another disadvantage is the potential loss of correlation in mixed feature space [17]. Since our goal is to develop computationally efficient and accurate method, we focused on the “early fusion” approach. Specifically, we create the final feature vector by concatenating the extracted feature vectors (3 color and 2 textures), where not all feature vectors are necessarily involved. Formally, the most extensive form of final feature vector is:

$$\mathbf{f}_i = (\mathbf{f}^{SCD}, \mathbf{f}^{CSD}, \mathbf{f}^{CLD}, \mathbf{f}^{HTD}, \mathbf{f}^{EHD}) \tag{6}$$

When considering the combination of only two features (e.g. CLD+EHD), we will use the final feature vector in the form $\mathbf{f}_i = (\mathbf{f}^{CLD}, \mathbf{f}^{EHD})$. As it will be shown in the experimental evaluation, not all the features have to be combined to achieve the best performance. Our experiments have shown that the combination of only a few of them is sufficient for fast and accurate indoor-outdoor image classification. Depending on the number of features chosen for image representation, the final feature vector will contain from 22 up to 292 elements. The feature vector formed in this way will serve as an input of SVM classifier described in the following section.

4. SVM Based Indoor-Outdoor Image Classification

SVM is one of the most popular machine learning methods for classification of multimedia content [6], [8], [10], [16]-[18]. It is a supervised machine learning technique that performs learning from examples in order to predict the values of previously unseen data. SVM can be formalized as an optimization problem which finds the best hyperplane for two or more groups of vectors by maximizing the size of the margin between groups. In order to get the best performance SVM, it is crucial to apply appropriate feature scaling and SVM parameters tuning [35]. The procedures that we performed are described in details below.

Although many existing SVM approaches apply some sort of feature scaling, the impact of this on SVM classification performance is still not sufficiently clear. Our intension is to empirically test significance of feature scaling procedures before SVM classification. In general, complex image features may contain a significantly different range of values since they are combined from several components (e.g. texture and color information). As a consequence, components with a higher variance will be dominant in determining distance between images. To avoid high influence of the feature component with a large variance, each element of the feature vector must be scaled using appropriate method.

Let us consider a collection of *m* indoor-outdoor images where each image is represented by its n dimensional feature vector formed by combination of several MPEG-7 descriptors. We examined two basic and efficient feature scaling methods: a) linear min-max scaling [11], and b) scaling to zero mean and unit variance (“z-score”) [29]. Linear min-max method can be mathematically represented as:

$$\mathbf{f}_i^{'}(j) = \frac{\mathbf{f}_i(j) - \min \mathbf{f}_i(j)}{\max \mathbf{f}_i(j) - \min \mathbf{f}_i(j)}, \quad i = 1,\ldots,m;\ j = 1,\ldots,n\,, \tag{7}$$

where $\mathbf{f}_i^{'}(j)$ represents element at position *j* in the scaled feature vector of the image *i*, min $\mathbf{f}_i(j)$ and max $\mathbf{f}_i(j)$ are minimal and maximal elements at position *j* among all training feature vectors. The resulting feature vector will be normalized to range [0, 1]. This approach has the advantage that the relative distributions (variances) of both rows and columns of the feature matrix are preserved.

Another approach to be considered is scaling to zero mean and unit variance ("z-score"). It is defined by:

$$\mathbf{f}_i^{'}(j) = \frac{\mathbf{f}_i(j) - mean\,\mathbf{f}_i(j)}{stdev\,\mathbf{f}_i(j)}, \quad i = 1,...,m;\ j = 1,...,n \tag{8}$$

where *mean* $\mathbf{f}_i(j)$ and *stdev* $\mathbf{f}_i(j)$ represents mean and standard deviation of elements at position *j* among all training feature vector.

The comparative evaluation of two scaling approaches will reveal their influence on SVM classification performance. As we will show later on, "z-score" significantly outperforms the first method, and hence was chosen as the optimal one. It will be shown that proper feature scaling may increase classification accuracy up to several percent.

Besides appropriate scaling of feature vectors, SVM requires to choose a kernel function with corresponding parameters. We considered a commonly used non-linear Gaussian RBF kernel with L2 norm, over the scaled feature vectors. For optimal selection of the SVM parameters pair we applied "n-fold cross validation" which separates the training dataset into *n* subsets and tests every subset using a SVM classifier trained on the remaining subsets. Systematic "grid-search" [13] was performed over various pairs of values to select the pair with the best accuracy. In order to limit the search complexity, parameter values for evaluation were sampled to form a grid of equidistant steps.

5. Experimental Evaluation

Image datasets

Currently, there is an evident lack of a comprehensive standard dataset for indoor-outdoor image classification testing. Most of the proposed methods in the literature use their custom image datasets. For the purpose of objectivity, we will test and compare our results only with relevant methods whose test datasets are publicly available. Thus, methods presented in [19] and [10] were used for comparison using datasets they provided.

The first image dataset is the IITM-SCID2 (Extended Scene Classification Image Database) introduced in [19]. It contains 902 indoor-outdoor images with a wide variation of scenes and resolutions in range from 80x80 up to 2048x1536 pixels. Out of this dataset, 193 indoor and 200 outdoor images are used for training, while 249 indoor and 260 outdoor images for testing. Compared to the second dataset, images of this dataset show a large variation in the scene content and resolution, which makes this dataset more suitable for testing of the real world performances. The second dataset, hereafter referred as COREL-INOUT, was provided by the authors in [10]. Its basis is the Wang's image database [30] extended with various images obtained from the web. It consists of a total of 1276 indoor-outdoor images of different scenes, all of the 256x256 pixels size. Specifically, 650 of the images were used for the training of classifiers among which 320 are indoor and 330 outdoor images. For the verification phase, other 626 images composed of 310 indoor and 316 outdoor images, are used. Examples of images from both datasets are shown in Fig. 1.

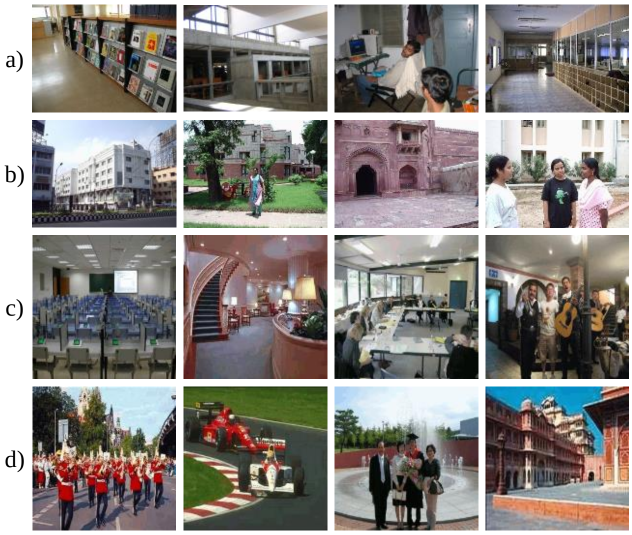

Fig. 1 Examples of test images: a) IITM-SCID2 Indoor, b) IITM-SCID2 Outdoor, c) COREL-INOUT Indoor, d) COREL-INOUT Outdoor

Test results

We have tested performance of indoor-outdoor image classification using various combinations of descriptors as well as two feature scaling approaches. The prototype system is implemented in MATLAB, where MPEG-7 features are extracted using C++ implementation [31]. For SVM classification we utilized MATLAB implementation of LibSVM library [32]. In the first experiment we have tested the impact of feature vectors scaling on SVM classification accuracy. Table 2 presents accuracy of SVM classification for different single descriptors, when feature vectors are scaled using two approaches: min-max and “z-score”.

Table 2 Impact of feature scaling method on the SVM classification accuracy (in %)

Descriptor	IITM-SCID2 dataset		COREL-INOUT dataset	
(dimension)	min-max	z-score	min-max	z-score
SCD (64)	81.73	84.68	70.73	70.93
CSD (64)	84.87	87.43	80.67	83.23
CLD (22)	78.98	82.71	80.83	82.59
HTD (62)	79.17	82.12	74.76	79.39
EHD (80)	83.10	87.03	83.70	83.87

We have performed a comprehensive experimental evaluation in order to get the combination of MPEG-7 features that gives the best classification performances. Table 3 presents SVM classification accuracy of the proposed method for different combinations of MPEG-7 color and texture descriptors. Note that each combined descriptor includes at least one color and one texture descriptor. The same tests are performed on IITM-SCID2 and COREL-INOUT image datasets.

Table 3 Accuracy of SVM classification using different combination of descriptors and "z-score" preprocessing (in %)

Descriptor combination	Dimens.	IITM-SCID2 dataset	COREL-INOUT dataset
SCD+HTD	126	88.61	81.95
SCD+EHD	144	91.36	87.38
CSD+HTD	126	88.02	86.74
CSD+EHD	144	90.37	88.66
CLD+HTD	84	88.02	87.70
CLD+EHD	102	91.55	91.53
SCD+CSD+HTD	190	89.59	84.82
SCD+CSD+EHD	208	91.55	88.66
SCD+CLD+HTD	148	92.34	88.66
SCD+CLD+EHD	166	93.71	91.53
CSD+CLD+HTD	148	91.16	89.30
CSD+CLD+EHD	166	92.93	91.05
SCD+HTD+EHD	206	91.36	89.30
CSD+HTD+EHD	206	92.73	91.05
CLD+HTD+EHD	164	92.34	92.01
SCD+CSD+CLD+HTD	212	91.94	88.82
SCD+CSD+CLD+EHD	230	93.32	91.05
SCD+CSD+HTD+EHD	270	92.34	89.78
SCD+CLD+HTD+EHD	228	93.32	92.17
CSD+CLD+HTD+EHD	228	93.71	92.49
SCD+CSD+CLD+HTD+EHD	292	93.71	92.01

It can be observed that the combination of four descriptors CSD+CLD+HTD+EHD gives the best overall results for both datasets, and therefore can be considered the optimal combination of MPEG-7 descriptors. It can also be noted that among combinations of two descriptors, CLD+EHD performs better than all others. When considering three descriptors combinations, SCD+CLD+EHD gives the best average accuracy. General observation is that the introduction of additional descriptor does not necessarily lead to performance improvement. If a request is to have a fast and sufficiently accurate descriptor, than CLD+EHD represents a reasonable choice, providing excellent costs/performance ratio.

Finally, Table 4 presents the results of our method using the most accurate MPEG-7 descriptors combination (CSD+CLD+HTD+EHD) with "z-score" scaling and SVM classification, compared to the results of methods [19] and [10] on IITM-SCID2 and COREL-INOUT datasets, respectively. The results presented in Table 4 show that the proposed method outperforms both compared methods. We have achieved 93.71% classification accuracy on IITM-SCID2 dataset, which is better than 92.36% reported in [19]. On the second dataset, a result of 92.49% is improvement of over 2% compared to [10]. Since the overall accuracy is over 92.49%, it may be concluded that the proposed method is very effective for the indoor-outdoor image classification. There should also be noted high quality of the results despite the relatively small size of the training datasets; knowing that SVM requires a rather large dataset of images to obtain good generalization capabilities.

Table 4 Accuracy comparison of different methods for indoor-outdoor image classification (in %)

Method	IITM-SCID2 dataset			COREL-INOUT dataset		
	Total	Indoor	Outdoor	Total	Indoor	Outdoor
Gupta et al. [19]	92.36	94.00	90.80	-	-	-
Kim et al. [10]	-	-	-	90.26	90.00	90.29
Our method	93.71	95.58	91.92	92.49	93.55	91.46

6. Conclusion

We have presented a relatively simple but highly accurate method for indoor-outdoor image classification based on combination of MPEG-7 features and SVM classification. Since we intended to create a computationally efficient method, we chose to apply the combination of low-level color and texture features in which all features contribute equally to the final result. We have empirically found that the combination of four MPEG-7 descriptors (CSD+CLD+HTD+EHD) scaled to zero mean and unit variance before input into SVM classifier, outperforms all others. Also, the combination of only two descriptors CLD+EHD is a good trade-off if we further intend to reduce computational costs while retaining the high level of accuracy. Experiments conducted on two public datasets achieved 93.71% and 92.49% accuracy, which is comparative to the top results previously published in the literature. Future research will be targeted towards using regions of interest (ROI) [33] for performance improving.

References

[1] R. Datta, D. Joshi, J. Li, J. Z. Wang, “Image retrieval: ideas, influences, and trends of the new age,” ACM Computing Surveys, vol. 40, no. 2, pp. 1-60, 2008.

[2] S. Bianco, G. Ciocca, C. Cusano, R. Schettini, “Improving color constancy using indoor-outdoor image classification,” IEEE Transactions on Image Processing, vol. 17, no. 12, pp. 2381-2392, 2008.

[3] L. Zhang, M. Li, H.-J. Zhang, “Boosting image orientation detection with indoor vs. outdoor classification,” Proceedings of WACV ’02, Washington, DC, USA, IEEE Computer Society, 2002, pp. 95-99.

[4] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, H.-J Zhang, “Image classification for content-based indexing,” IEEE Transactions on Image Processing, vol. 10, no. 1, pp. 117-130, 2001.

[5] J. Collier, A. Ramirez-Serrano, “Environment classification for indoor/outdoor robotic mapping,” Proceedings of Canadian Conference on Computer and Robot Vision CRV’09, Kelowna, British Columbia, Canada, 2009, pp. 276-283.

[6] M. Boutell, J. Luo, “Beyond pixels: exploiting camera metadata for photo classification,” Pattern Recognition, vol. 38, no. 6, pp. 935-946, 2005.

[7] X. Liu, L. Zhang, M. Li, H. Zhang, D. Wang, ”Boosting image classification with LDA-based feature combination for digital photograph management,” Pattern Recognition, vol. 38, pp. 887-901, 2005.

[8] N. Serrano, A. Savakis, J. Luo, “Improved scene classification using efficient low-level features and semantic cues,” Pattern Recognition, 37(9), pp. 1773-1784, 2004.

[9] L. Lu, K. Toyama, G. D. Hager, ”A two level approach for scene recognition,” Proceedings of CVPR’05, Washington, DC, IEEE Computer Society, 2005, pp. 688-695.

[10] W. Kim, J. Park, C. Kim, “A novel method for efficient indoor–outdoor image classification,” Journal of Signal Processing Systems, vol. 61, no. 3, pp. 251-258, 2010.

[11] H. Eidenberger, „Statistical analysis of content-based MPEG-7 descriptors for image retrieval,” Multimedia Systems, vol. 10, no. 2, pp. 84-97, 2004.

[12] T. Deselaers, D. Keysers, H. Ney, "Features for image retrieval: A quantitative comparison," Proceedings of DAGM SSPR'04, Tübingen, Germany, Springer, 2004, pp. 228-236.

[13] T. Deselaers, D. Keysers, H. Ney, "Features for image retrieval: an experimental comparison," Information Retrieval, vol. 11, pp. 77-107, 2008.

[14] B.S. Manjunath, P. Salembier, T. Sikora, Introduction to MPEG-7, San Francisco, CA, USA, Wiley, 2002.

[15] S. Chang, T. Sikora, A. Puri, "Overview of the MPEG-7 standard," IEEE Transactions on Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 688-695, 2001.

[16] S. N. Lindstaedt, R. Mörzinger, R. Sorschag, V. Pammer, G. Thallinger, "Automatic image annotation using visual content and folksonomies," Multimedia Tools and Applications, vol. 42, no. 1, pp. 97-113, 2009.

[17] C. G. M. Snoek, M. Worring, A. W. M. Smeulders, "Early versus late fusion in semantic video analysis," Proceedings of ACM Multimedia '05, New York, NY, USA, ACM, 2005, pp. 399-402.

[18] N. Serrano, A. Savakis, A. Luo, "A computationally efficient approach to indoor/outdoor scene classification," Proceedings of ICPR'02, Quebec City, Canada, 2002, pp. 146-149.

[19] L. Gupta, V. Pathangay, A. Patra, A. Dyana, S. Das, "Indoor versus outdoor scene classification using probabilistic neural network," EURASIP Journal on Advances in Signal Processing, pp. 1-11, 2007.

[20] S. Park, "Content-based image classification using a neural network," Pattern Recognition Letters, vol. 25, no. 3, pp. 287-300, 2004.

[21] M. Szummer, R. W. Picard, "Indoor-outdoor image classification," Proceedings of IWCBAIVD'98, IEEE Computer Society, 1998, pp. 42-51.

[22] A. Payne, S. Singh, "Indoor vs. outdoor scene classification in digital photographs," Pattern Recognition, vol. 38, no. 10, 2005, pp. 1533-1545, 2005.

[23] H. Zhang, A. C. Berg, M. Maire, J. Malik, "SVM-KNN: discriminative nearest neighbor classification for visual category recognition," in Proceedings of CVPR'06, New York, NY, USA, 2006, pp. 2126-2136.

[24] M. Varma and D. Ray, "Learning the discriminative power-invariance trade-off," Proceedings of ICCV'07, Rio de Janeiro, Brazil, 2007, pp. 1-8. Avilable: http://dx.doi.org/10.1109/ICCV.2007.4408875

[25] B. S. Manjunath, J. R. Ohm, V.V. Vinod, A. Yamada, "Color and texture descriptors," IEEE Trans. Circuits and Systems for Video Technology, vol. 11, no. 6, pp. 703-715, 2001.

[26] A. Yamada, M. Pickering, S. Jeannin, L. Cieplinski, J. R. Ohm, M. Kim, MPEG-7 Visual part of experimentation Model Version 10.0. ISO/IEC JTC1/SC29/WG11/N4063, 2001.

[27] R. Datta, J. Li, J. Z. Wang, "Content-based image retrieval: approaches and trends of the new age," Proceedings ACM SIGMM MIR '05, New York, NY, USA, ACM, 2005, pp. 253-262.

[28] S. Ayache, G. Quénot, J. Gensel, "Classifier fusion for SVM-based multimedia semantic indexing," Proceedings of ECIR'07, Berlin, Germany, Springer-Verlag, 2007, pp. 494-504.

[29] R. J. Larsen and M. L. Marx. An introduction to mathematical statistics and its applications, Pearson Prentice Hall, 2006.

[30] J. Z. Wang, J. Li, G. Wiederhold, "Simplicity: semantics-sensitive integrated matching for picture libraries," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 9, pp. 947-963, 2001.

[31] M. Bastan, H. Cam, U. Gudukbay, O. Ulusoy, "BilVideo-7: an MPEG-7- compatible video indexing and retrieval system," IEEE Multimedia, vol. 17, pp. 62-73, 2010.

[32] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, pp. 1-27, 2011.

[33] J. Lee, J. Nang, "Content-based image retrieval method using the relative location of multiple ROIs," Advances in Electrical and Computer Engineering, vol. 11, no. 3, pp. 85 – 90, 2011.

[34] M. Soysal and A.A. Alatan, "Combining MPEG-7 Based Visual Experts For Reaching Semantics," in Proc. of VLBV03, Madrid, 2003.

[35] D. Lu and D. Weng, "A survey of image classification methods and techniques for improving classification performance," International Journal of Remote Sensing, vol. 28, Issue 5, 2007, pp. 823-870.

[36] J. Li and J.Z. Wang, "Automatic Linguistic Indexing of Pictures by A Statistical Modeling Approach," IEEE Trans. on PAMI, vol. 25, No. 9, 2003, pp.1075-1088.