Content-based Indoor/Outdoor Video Classification System for a Mobile Platform

Mitko Veta, Tomislav Kartalov, and Zoran Ivanovski

Abstract—Organization of video databases is becoming difficult task as the amount of video content increases. Video classification based on the content of videos can significantly increase the speed of tasks such as browsing and searching for a particular video in a database. In this paper, a content-based videos classification system for the classes indoor and outdoor is presented. The system is intended to be used on a mobile platform with modest resources. The algorithm makes use of the temporal redundancy in videos, which allows using an uncomplicated classification model while still achieving reasonable accuracy. The training and evaluation was done on a video database of 443 videos downloaded from a video sharing service. A total accuracy of 87.36% was achieved.

Keywords—Indoor/outdoor, video classification, image classification

I. INTRODUCTION

SINCE video recording devices are becoming more and more available to the average consumer, the amount of generated video content is increasing exponentially. Furthermore, with the popularization of video sharing services and increased access to broadband Internet, a great number of videos are available at each moment. Taking into account the volume of the data, tasks such as organization or searching through the available content can be very time consuming and tiresome. Therefore such tasks need to be done automatically, preferable using solely the content of the videos to accomplish this. One way to organize content is by grouping it into semantically meaningful categories. These categories later can be used as queries to ease the browsing and navigation trough the database. Closely related problem to video database organization is organization of image databases and photo albums.

The problem that arises with automatic video or image database organization is how to bridge the semantic gap between low level concepts (such as color, texture, shape etc.) and more advanced semantic concepts (like indoor, outdoor, natural, artificial, etc). This is usually done by classifying images based on low-level features into one or more classes, each representing a higher level semantic concept [1].

This paper presents a content-based indoor/outdoor video classification system. The advantage of grouping videos into these two categories is obvious. If one wants to search a large

video collection and find all the videos from a birthday party for example, looking only trough the indoor videos would reduce the needed time. Since today's mobile devices are capable of storing and accessing large quantities of video data, but still posses low computational power, a restriction is imposed that the classification system must be fast enough to work in real-time on such a device (ex. mobile phone). Additional requirement is that the system should work with videos recorded with a variety of recording devices and compressed with different encoding techniques. In order to be stored on a mobile platform or transmitted over a mobile network the videos are usually compressed or trans-coded at very low bit-rates and this must also be taken into account.

In the development of the classification system an assumption was made that all the videos must belong to only one of the two classes. The ground truth assignment is done based on the dominant content of the video. This means that if a video which is mainly indoor has some outdoor elements it will still be considered as indoor. In comparison to image classification, in video classification the temporal redundancy can be used to improve the classification accuracy or use uncomplicated classification model but still achieve reasonable accuracy. This concept was used in the video classification system presented in this paper to keep the computational cost low.

The paper is organized as follows: in Section II a general overview of the related work in the field of indoor/outdoor classification is given. Section III describes the video summarization and feature extraction techniques that are used in the classification system and Section IV describes the used classification technique. Results are presented in Section V and conclusions and future work in Section VI.

II. RELATED WORK

A great deal of work has been done on the subject of content-based indoor/outdoor classification from low-level features, mainly concerning still images. Szummer and Picard in [2] use a two-stage approach for indoor/outdoor image classification, classifying sub-blocks of the image with k-NN as the first stage. The final decision for each image is based on the decisions for the sub-blocks using a majority rule. With a combination of Ohta color space histogram and MSAR for texture representation as features, classification accuracy of 90.3% is achieved. The accuracy is measured on a collection of over 1300 consumer images. Similar to this, in [3] LST color histogram and wavelet texture features are used for classification of image sub-blocks using SVM with RBF

kernel. In addition, semantic detectors for sky, clouds and grass are used to improve the classification accuracy. The results from the sub-blocks classification and the outputs of the semantic detectors are combined using a Bayesian network. The reported performance is 90.7% accuracy on a set of 1200 consumer images. The proportion of straight edges in the image is used as a feature in [4]. The authors claim that indoor images contain bigger proportion of straight edges than outdoor images. The final decision for the class of the image is made based on a simple rule applied to the proportion of straight edges contained in sub-blocks of the image. Using a multi-resolution scheme to improve the performance, a classification accuracy of 90.71% is reported on a set of 872 images.

Schettini et al. [5] have developed an image classification system which distinguishes between the classes indoor, outdoor and close-ups. The feature vector is consisted of color, texture, edge distribution and composition features. As a classifier decision forests of trees built according to the CART methodology are used. A rejection criterion to automatically reject ambiguous images is implemented and the reported classification accuracy is around 92% with 10% rejection. Similar approach to this was used in [6], where indoor/outdoor classification is used to improve color constancy.

Indoor/outdoor image classification is proposed as a first step towards hierarchical image classification scheme in [7]. In this paper, the image is tessellated in 10x10 sub-blocks for which first and second color moments in the LUV color space are computed. The feature vector for the entire image is produced with concatenation of the feature vectors of all sub-blocks. The reported accuracy is 90.5% with k-NN classifier using a codebook of 30 samples.

All of the previously mentioned techniques are concerned with image classification only. An attempt for video indoor/outdoor classification is made in [8] where video shots are classified based on a set of extracted key frames. The features used for each frame are mean, variance and number of peaks of the histogram in RGB color space. A feed forward neural network with back-propagation learning is used as a classifier. The decision for the entire shot is based on the decisions for the individual frames. No classification accuracy is reported.

In the here presented paper the video classification problem will be addressed, taking into account the computational complexity and feasibility of the system on a mobile platform.

III. VIDEO REPRESENTATION

A. Video Summarization

To reduce the amount of data and save computational time, each video is represented by a set of key frames. A simple key frame extraction procedure, similar to the one in [8], is used.

The key frame extraction procedure is the following: start at t_s from the beginning of the video and extract a frame every T seconds. By skipping the first t_s seconds from the video,

ambiguous frames, which very often appear at the beginning of the video, are avoided. The parameter T, should be chosen considering the following two antagonistic criteria: T should be big enough to avoid frames with similar content, but small enough to capture as much different content as possible. This simple approach has proved to be sufficiently effective, as shown later in the Results section.

B. Feature Extraction

Each key frame is divided into sub-blocks and feature vectors are calculated for each sub-block separately. The sub-blocks are produced by tessellating the frame with a 4x4 grid. Color and texture, which are used as features here, can be considered uniform in small parts of a video frame and thus they are better represented with a local descriptor [4].

One of the imposed conditions in the definition of this problem – the limited computational power and memory, requires a feature vector that is of very low dimensionality and inexpensive to compute. Two different sets of features, both of size 1x12, were considered and evaluated. The first is color-only feature vector - histogram in the YCbCr color space. Each channel is allocated 4 bins of the histogram. For the luminance the bins are uniformly distributed over the entire channel range [16-234]. Since extreme values (very low or very high) for the color channels are rare, the histogram is computed only for the sub-interval [90 165] for these channels in order to exploit the allocated 4 bins more efficiently. The choice of the YCbCr color space is straightforward – an assumption is made that the video decoder produces the frames in this color space, thus no extra computational power is needed for color space conversion.

The second set uses both color and texture features. The texture features are produced by first filtering the sub-block with the set of directional gradient filters given with (1). To lower the computational cost each sub-block is downscaled by two before filtering. In addition, the downscaling helps to lower the influence of some of the compression artifacts present in the frames. The downscaling is done with decimation only, and without low-pass filtering. Because of the high compression of the videos, low-pass filtering is not needed since high frequencies are rarely present.

The first two filters given with (1) produce the horizontal and vertical gradient of the image. The other two filters capture the gradients in the two diagonal directions. Filter coefficients are chosen so the outputs can be efficiently computed – multiplication with 0.5 can be replaced with binary shift operation. Also note that elements in the second row and column are the same in both filters, so the operations corresponding to those coefficients are done only once. Let $f_n(i, j)$ be the filter output of the n-th filter for the pixel at location (i,j), and M_d and N_d are the width and height of the downscaled sub-block in pixels. The texture feature is formed with the mean and approximation of deviation of the absolute value of the output of each filter, according to (2) and (3).

TABLE I
CLASSIFICATION ACCURACIES PER SUB-BLOCKS IN %

Features	$k=1$	$k=3$	$k=5$	$k=7$	$k=6$	$k=11$	$k=13$
Set 1	64.88	66.43	67.77	67.32	67.90	68.23	**68.35**
Set 2	63.41	65.43	67.47	**68.45**	67.65	68.29	67.53
Set 1, top 8	66.95	67.62	67.93	68.29	**68.84**	**68.84**	68.05
Set 2, top 8	67.93	69.51	70.91	**70.98**	70.30	70.61	70.61
Set 1, bottom 8	65.43	67.50	68.72	69.15	70.18	69.94	**70.55**
Set 2, bottom 8	62.62	65.30	66.04	65.79	65.37	66.52	**67.01**

The maximum accuracy for each set is given in bold

$$F_1 = \begin{bmatrix} 1 & -1 \end{bmatrix} \quad F_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$$F_3 = \begin{bmatrix} 0 & -0.5 & 0.5 \\ -0.5 & 1 & -0.5 \\ 0.5 & -0.5 & 0 \end{bmatrix}$$
(1)

$$F_4 = \begin{bmatrix} 0.5 & -0.5 & 0 \\ -0.5 & 1 & -0.5 \\ 0 & -0.5 & 0.5 \end{bmatrix}$$

$$m_n = \frac{1}{M_d N_d} \sum_{i,j} |f_n(i,j)|$$
(2)

$$d_n = \frac{1}{M_d N_d} \sum_{i,j} \left| |f_n(i,j)| - m_n \right|$$
(3)

This results in 8 values which form the texture part of the feature vector. The means and deviations are normalized, each separately, with the maximum value.

For the color part of the feature vector, a notion stated in [3] is exploited. The authors claim that the S component of this color space represents the daylight to tungsten illuminant variation. It is obvious that this would be very beneficial for the indoor/outdoor classification task. Our experiments have shown that this component is indeed very discriminative for the two classes. It can be seen from the equation for RGB to LST color space conversion that the S channel is proportional to the difference between R and B channels. If the difference between the Cb and Cr channels in the YCbCr space is considered, it can be derived that it is also proportional to the R-B difference (equations (4) and (5)).

$$S = k_s(R - B)$$
(4)

$$C_r - C_b = 0.587R + 0.0770G - 0.51B \approx k(R - B)$$
(5)

Because of this, the color feature is set to be a 4 bin histogram of the difference between the channels Cr and Cb. Here also, extreme values are rarely possible, so the histogram is computed only for the sub-interval [-75 75]. The texture and color features are concatenated so that the final length for this feature vector is also 12.

The descriptive performance of the features was evaluated using k-NN classification on a set of sub-block features extracted from 205 frames. These frames are a sub-set of the set of frames that were used for training of the system, as described in the next section. Method similar to leave-one-out cross-validation was used, where for classification of each sub-block, sub-blocks from the same frame are excluded. Euclidean distance was used as a distance metric. The lower and upper 8 sub-blocks were evaluated separately suspecting that the classification accuracies might be higher or lower when only one group is considered. The results are summarized in Table I.

When considering all the sub-blocks both set of features perform similar. However, the second set of features performs better in the case where only the upper 8 sub-blocks are considered. The better classification accuracy for the top part of the image can be easily explained by the presence of sky in outdoor scenes which can be easily captured by the color feature. Also, the upper part of the image is more affected by the light source in the scene, which can be discriminative for both classes. The first set of features produces better accuracy, by a smaller margin though, for the lower sub-blocks. These features can capture the presence of grass in the lower part of the image better. The accuracies achieved here for only one group of sub-blocks are lower but comparable to the per sub-block accuracies reported in [2] and [3] where more complex features are used.

In our classification system the second set of features and only the upper 8 sub blocks will be used. Using only the upper 8 sub-blocks in the classification system is of double benefit: the amount of operations is reduced by half while better classification accuracy is achieved. The second set was chosen because performs slightly better and achieves its maximum performance for lower k compared to the first set of features used on the lower sub-blocks. Also, our experiments showed that these features produce better accuracy on the video database.

IV. CLASSIFICATION

In our classification system each sub-block from a video to be decided upon is classified separately. Then, based on those results, a decision is made for the entire video. As a classifier at sub-block level k-NN with city-block distance is used because it is simple and yet effective. Experiments with support vector machines with RBF kernel were also made using the LIBSVM library [9]. They produced similar results as k-NN, however at much higher cost, making them

impractical for real-time implementation on a mobile platform.

One of the drawbacks of the k-NN classifier is that it does not build a model for the training data, but instead stores all the samples in the training set which requires a lot of memory. To reduce the training data, a vector quantization technique is used. First, features are computed for each sub-block from the frames in the training set. Then, these features are clustered into C clusters using the k-means algorithm with the city-block distance and preliminary clustering on 10% of the samples to choose the initial set of centroids. Each cluster is assigned a label according to the dominant presence of feature vectors from one of the classes. If more than P percent of the samples in the cluster belong to only one class (indoor or outdoor) the cluster is labeled according to the dominant class. If not, the cluster is labeled as the newly introduced class designated as "unknown". The cluster centroids with their appropriate labels are used as the new training set (codebook) for the k-NN classifier.

When classification for an unknown video is to be made, each sub-block extracted from that video is classified as indoor, outdoor or unknown using the k-NN classifier. The classification for the entire video is done according to (6), where B_i is the number of sub-blocks (from all key frames from the video) classified as indoor and B_o as outdoor. T_h is an empirically set threshold. The sub-blocks classified as unknown are discarded. If the inequality (6) holds the video is classified as indoor, else it is classified as outdoor.

$$\frac{B_i}{B_i + B_o} \geq T_h \qquad (6)$$

Discussion about the parameters and the classification accuracy is presented in the next section.

V. EXPERIMENTAL RESULTS

All the videos used for training and evaluation were collected from You Tube®. While collecting, special care was taken that the videos are user generated, without use of extensive editing and computer-generated effects. The video quality varies from low to medium. The typical duration of the videos is few minutes. A total of 443 videos were collected – 217 indoor and 226 outdoor.

The ground truth for the class of each video was assigned by the authors. As previously discussed, the assignment was done based on the dominant content (indoor or outdoor) present in the video. In most of the videos only one content is present but there are also some mixed videos. An example of this would be indoor scene with outdoor elements visible trough the window, or a video where the subject that is recording walks into a house. Based on the way the videos were collected, a safe assumption can be made that the diversity of the recording devices is great. The content of the videos is also very diverse, ranging from typical urban scenes to mountain views for the outdoor class, and from living

TABLE II
CONFUSION MATRIX

Ground truth	Classification result		
	Indoor	Outdoor	Accuracy
Indoor	184	32	85.25%
Outdoor	24	202	89.38%

rooms to malls for the indoor class. The videos were extensively compressed so they can be distributed more easily over the Internet. Safe assumption can be made that some of the videos were previously compressed and then trans-coded when uploaded to You Tube. Because of these reasons, almost all of the videos suffer from compression artifacts which make the indoor/outdoor classification task more difficult.

From the video database a training set consisting of 1115 frames was made out of which 536 are indoor and 576 outdoor. These frames were chosen to be typical representatives of their class. Damaged, blurred and overly dark frames were avoided as they can introduce confusion.

The algorithm was tested on all the videos in the database using a leave-one-out approach: before classifying a video from the database, frames from that video present in the training set are removed, and vector quantization is performed on the remaining feature vectors as described in the previous section. Then, sub-blocks from the video are classified using that particular quantization and a decision for the video is made according to (6). This procedure was repeated for each video in the database.

The parameters t_s and T for the key frame extraction were set at 5 s both. For the vector quantization, the number of clusters C was set to 200. For the cluster labeling, the P parameter was set at 70%. The T_h parameter was set to be 0.4 meaning that in order a video to be classified as indoor, only 40% of its blocks, who are not classified as "unknown", need to be classified as indoor. For the k-NN classification the parameter k was set to 1, which means that an unknown sub-block is classified according to its single nearest neighbor in the training set. These parameters were empirically set according to a subset of the videos and then tested on the remaining of the database. Table II gives the confusion matrix for the classification of all the videos in the database using this set of parameters.

Since the random initial values of the vector quantization can produce different classification results, the same experiment was repeated five times. The achieved accuracy in each experiment was very similar to the one in Table II.

To recognize the strong and weak points of the classification system, the video database was divided into sub-classes and classification accuracies were calculated for each sub class separately. For the outdoor class, the division was made based on the presence of sky in the video and weather the content is urban or natural scene. For the indoor class the division was made based on the light source in the video. The per sub-class accuracies are presented in Table III. It can be concluded from the table that the weak points of the classification system are outdoor videos recorded in an urban

TABLE III
CLASSIFICATION ACCURACIES FOR DIFFERENT SUB-CLASSES

Outdoor			Indoor		
urban with no/little sky	21 out of 31	67,74%	indoor with artificial lighting	120 out of 131	91,60%
urban with sky	77 out of 81	95.06%	indoor with natural lighting	22 out of 29	75,86%
natural with sky	69 out of 70	99.57%	indoor with mixed lighting	41 out of 52	78,85%
natural with no/little sky	32 out of 37	86,49%	miscellaneous	2 out of 5	40.00%
miscellaneous	3 out of 7	42,86%			
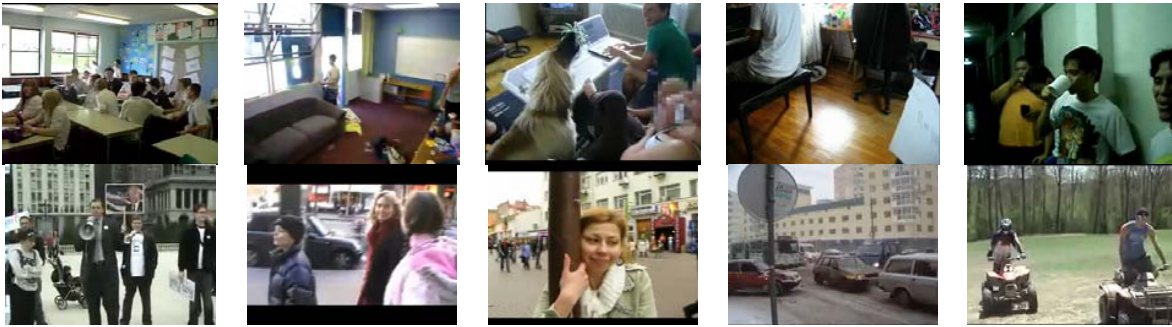
a) Frames from falsely classified videos
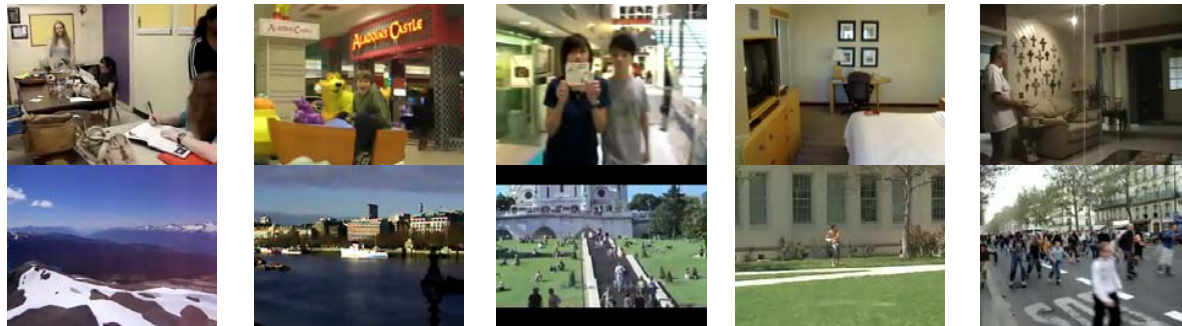
b) Frames from correctly classified videos

Fig. 1 Examples of frames from the database

environment without visible sky and indoor videos with natural or mixed light source. Some examples of frames from falsely and correctly classified videos are presented in Fig. 1.

VI. CONCLUSION

A computationally efficient indoor/outdoor video classification system was presented. The algorithm makes use of the temporal redundancy of information in videos. Since the decision is made for a complete video sequence, the accuracy of the single sub-block classification could be lower. This allows lowering of the computational complexity of the feature vector calculation and of the classification, while the total performance of the system is kept high. The system was evaluated on real-world videos downloaded from You Tube and achieved total accuracy of 87.36% on a database of 443 videos. The weak points were pin-pointed by examining the accuracies per sub-classes with respect to presence of sky and natural/artificial content for the outdoor class, and light source

for the indoor class. Our future work will address these issues more closely. Also, the algorithm will be tested on a database of video sequences from a single source, for example – sequences recorded with the camera of a single model of a mobile phone.

REFERENCES

[1] J.Stauder, J.Sirot, H. Le Borgne, E. Cooke, N.E.O'Connor "Relating visual and semantic image descriptors", Proceedings of European Workshop for the Integration of Knowledge, Semantic and Digital Media Technologies, EWIMT 2004, London, UK, November 25-26, 2004

[2] M. Szummer and R.W. Picard, "Indoor-outdoor image classification,"*1998 IEEE International Workshop on Content-Based Access of Image and Video Database, 1998. Proceedings.*, 1998, pp. 42-51.

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

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

[5] R. Schettini, C. Brambilla, C. Cusano, and G. Ciocca, "Automatic classification of digital photographs based on decision forests."[5] A. Payne and S. Singh, "Indoor vs. outdoor scene classification in digital photographs,"*Pattern Recognition*, vol. 38, 2005, pp. 1533-1545.

[6] S. Bianco, G. Ciocca, C. Cusano, and R. Schettini, "Improving Color Constancy Using Indoor–Outdoor Image Classification,"*Image Processing, IEEE Transactions on*, vol. 17, 2008, pp. 2381-2392.

[7] A. Vailaya, M.A.T. Figueiredo, A.K. Jain, H.J. Zhang, A. Technol, and P. Alto, "Image classification for content-based indexing,"*IEEE Transactions on Image Processing*, vol. 10, 2001, pp. 117-130.

[8] A. Miene, T. Hermes, G. Ioannidis, R. Fathi, and O. Herzog, "Automatic shot boundary detection and classification of indoor and outdoor scenes," *NIST SPECIAL PUBLICATION SP*, 2003, pp. 615-620.

[9] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm