ON THE USE OF SPATIOTEMPORAL VISUAL ATTENTION FOR VIDEO CLASSIFICATION

Konstantinos Rapantzikos, Yannis Avrithis and Stefanos Kollias
School of Electrical & Computer Engineering
National Technical University of Athens
e-mail: {rap,iavr}@image.ntua.gr, stefanos@cs.ntua.gr

ABSTRACT

It is common sense among experts that visual attention plays an important role in perception, being necessary for obtaining salient information about the surroundings. It may be the "glue" that binds simple visual features into an object [1]. Having proposed a spatiotemporal model for visual attention in the past, we elaborate on this work and use it for video classification. Our claim is that simple visual features bound to spatiotemporal salient regions will better represent the video content. Hence, we expect that feature vectors extracted from these regions will enhance the performance of the classifier. We present statistics on sports sequences of five different categories that verify our claims.

I. INTRODUCTION

Image and video classification is an important, unsolved problem in multimedia content understanding that requires bridging the gap between the target semantic categories, or classes, and the low-level visual descriptors that can be automatically obtained. At the same time, it is a valuable tool towards other applications like object detection and recognition, visual content description, semantic metadata generation, indexing and retrieval.

In *unsupervised* classification, the classes are not known *a priori*, therefore the problem is typically handled through clustering, utilizing tools like expectation maximization (EM) [5], vector quantization [3] or k-means [4]. On the other hand, in *supervised* classification, the classes are known and utilized by a learning mechanism. The class models used can be based on the description of the classes themselves, typically employing again the above techniques, on the definition of limits between classes, employing support vector machines (SVM) [5], or on implicit modeling using neural networks.

In all cases, it is commonly believed that in order to achieve robust global classification, i.e. without prior object detection or recognition, it is crucial to select an appropriate set of *visual descriptors* that usually have to capture the particular properties of a specific domain and the distinctive characteristics of each image class. For instance, local color descriptors and global color histograms are used in indoor/outdoor classification [9] to detect e.g. vegetation (green) or sea (blue). Edge direction histograms are employed for city/landscape classification [6] since city images typically contain horizontal and vertical edges. Additional motion descriptors are also used for sports video shot classification [7, 8], while other alternatives are orientations, contours, texture models and DCT or wavelet coefficients [9].

Even in specific domains and appropriately selected descriptors, `classification usually fails in cases of close-up scenes (e.g., faces). If we could select the regions in an image or video that best describe its content, a classifier could be trained on such regions and learn efficiently to differentiate between different classes. This would also decrease the dependency on descriptor selection or feature formulation. In the absence of prior knowledge or object detection, [10] suggests a selection of video shots based on their homogeneity to cluster TV programs, while also employing a spatiotemporal volume representation that efficiently captures the dynamic nature of video sequences.

Selecting this small fraction of important information in a way similar to the human optical system is the main task of the *selective visual attention* (VA) process. *Saliency*-based attention has been computationally modeled in the last decade by Itti and Koch, [2], and seems to provide a reasonable first step towards the understanding of the visual input. Bottom-up attention, i.e., employing no *a priori* knowledge, has been employed as a pre-processing step towards more complex tasks like object recognition [11]. In our previous work [12, 13], we have extended Itti et al.'s scheme towards a *spatiotemporal VA* model that treats the temporal dimension of a video sequence as an intrinsic feature and provides a unifying framework to analyze the spatial and temporal video organization. In this work, we employ this model to video content classification, claiming that selective attention in the absence of knowledge, object recognition or domain-specific feature selection, can provide a powerful tool to train the classifier on spatiotemporal salient regions that better represent the video content.

II. SPATIOTEMPORAL VISUAL ATTENTION

In this section we briefly describe our earlier work [12, 13] towards extending the saliency-based visual attention of Itti *et al.* [2] to the spatiotemporal domain. Under the spatiotemporal framework, we treat the video sequence as a video volume with temporal evolution being the third dimension. The acquired frames form a video volume by stacking them one of top of the other. This volume is decomposed into a set of distinct "channels" such as luminance, red, green, blue, yellow hues and various orientations. The number and response properties of these filters have been chosen according to what is known of their neuronal equivalents in the early stages of visual processing in primates. Each of these *feature volumes* encodes a certain property of the video.

After obtaining the spatiotemporal data formation, the input volumes are morphologically filtered by a flat zone approach to avoid spurious details or noisy areas that might otherwise be erroneously attended by the proposed system. Following the structure of the static image-based approach of Itti *et al.*, we then generate feature volumes for each feature of interest, including intensity, color and 2D/3D orientation [13]. Every volume simultaneously represents the spatial distribution and temporal evolution of the encoded feature. A *normalization* operator is responsible for enhancing the most salient subvolumes inside them so as to prohibit non-important regions from drastically affecting the result.

The process described above is performed at a number of different spatiotemporal scales, to allow the model to represent smaller and larger "events" in separate subdivisions of these channels. This multiple scale representation is obtained through Gaussian pyramids. Center-surround operations, which are suitable for detecting locations that locally stand out from their surroundings, are implemented as differences between a fine and a coarse scale for a given feature. Finally, a linking stage fuses the separate volumes and produces a saliency volume that represents interesting events as enhanced (in terms of intensity) spatiotemporal regions. Fig. 1 illustrates all intermediate steps of the proposed model.

III. SUBVOLUME SELECTION & FEATURE EXTRACTION

The final saliency volume encodes the per voxel saliency of the original video. Obtaining a meaningful spatiotemporal segmentation of the saliency volume is not a simple and straightforward task. In this paper we focus our research on the usefulness of spatiotemporal attention for learning and classification. Hence, we adopt a simple segmentation technique that allows for non-hard thresholding and labeling of the various salient subvolumes. *K-means* is used to partition the final volume into regions of different saliency. Voxels are clustered in terms of their saliency value (intensity of each voxel) and a predefined number of clusters is extracted. After ordering the clusters in increasing order of saliency, we discard the less salient one and label the rest.

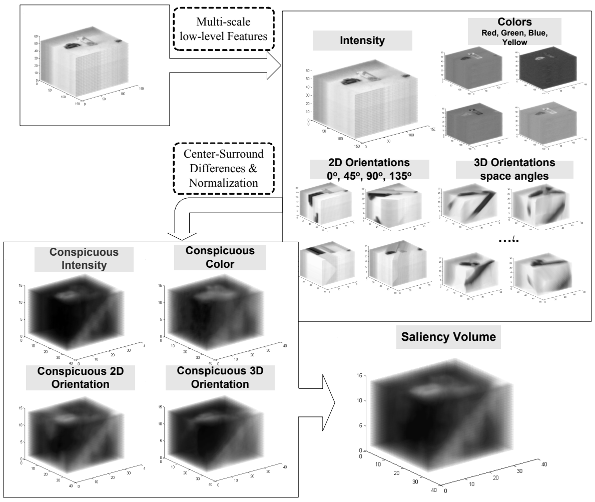

Figure 1. Spatiotemporal VA architecture. The feature extraction stage and the saliency volume generation are shown.

The reasoning on the previous procedure is related to the feature extraction process that follows. We expect that the less salient region is not representative of the video and therefore features extracted from such a region could confuse the classifier and increase the classification error.

In order to emphasize on the performance improvement achieved by the spatiotemporal saliency learning, we calculate the same simple features both on each separate (labeled) salient subvolume and the whole video volume. For this, we use color histograms to represent the color distribution among the RGB channels and a set of co-occurrence features for texture. Global color histograms are simple descriptors, fast to compute, and scale/rotation invariant; they also work on partial images. To keep the feature space low, we calculate color histograms by quantizing them in a small number of bins and obtain four texture measurements, namely entropy, inertia, energy and homogeneity from the co-occurrence matrix.

In order to formulate the above features into a single vector, we keep the three most salient regions, and, for each one, we encode the color histograms using 8 bins per color channel (i.e., 24 elements per region), and the texture features using each of the above measurements for 4 different region slices (i.e., 16 elements per region). The total size of each feature vector is thus 120.

IV. SVM CLASIFFIER

An SVM [5] performs pattern recognition for dichotomic classification problems (binary classification). It maximizes the distance between a hyperplane w and the closest samples to it, with the constraint that the samples from the two classes lie on separate classes of the hyperplane. These closest points are called support vectors. Given a training set of instance-label pairs $(x_i, y_i), i = 1,...,l$ where $x_i \in \Re^n$ and $y \in \{-1,1\}^l$, the SVMs require the solution of the following optimization problem:

$$\min_{w,b,x} \frac{1}{2} w^T w + C \sum_{i=1}^{l} x_i \qquad (1)$$

$$\text{s.t. } y_i \left(w^T f(x_i) + b \right) \geq 1 - x_i, \ x_i \geq 0$$

where the training data x_i are mapped to a higher dimensional space by function f and the second term of (1) is the penalty term with parameter C.

The multi-class classification problem is commonly solved by a decomposition to several binary problems for which the standard binary SVM can be used. The one-against-all decomposition is often applied. In this case the classification problem to k classes is countered by training k different classifiers, each one trained to distinguish the examples in a single class from the examples in all remaining classes. When it is desired to classify a new example, the k classifiers are run, and the classifier which outputs the largest (most positive) value is chosen.

In this work, we train the SVM classifiers using the linear kernel after appropriately selecting a model. For model selection we perform a "grid-search" on the regularization parameter $C = \{2^0, 2^1, 2^2, 2^3, 2^4\}$ using 5-fold cross-validation. After obtaining the parameter that yields the lowest testing error, we perform a refined search in a shorter range and obtain the final parameter value, C=5, which is selected for the classifiers.

V. LEARNING FROM SALIENCY

A. Experimental Setup

To demonstrate the potential of the proposed scheme we select a number of videos from five different sports. *Soccer, swimming, basketball, boxing* and *snooker* are the five predefined classes of shots we use for conducting our experiments. Each class includes far- and near-field views, close-ups on players and frames where all the playfield, players and audience are present. The length of the shots ranges from 6 to 7 seconds. All clips, each consisting of a single shot, are resized to have the same spatial dimensions and were manually annotated as belonging to either of the given classes. The spatiotemporal saliency volume was obtained using our algorithm that includes 3D orientation as a feature volume [13], as explained in section 2. The saliency volume is clustered as described in section 3. Non-salient regions are discarded and features are extracted from each remaining region. Fig. 2 shows indicative frames of each class and the obtained saliency masks corresponding to the three most salient regions. The third column shows the mask superimposed on the corresponding frame of the initial video. Representative results can be found at http://www.image.ntua.gr/~rap/vlbv05/VA/class

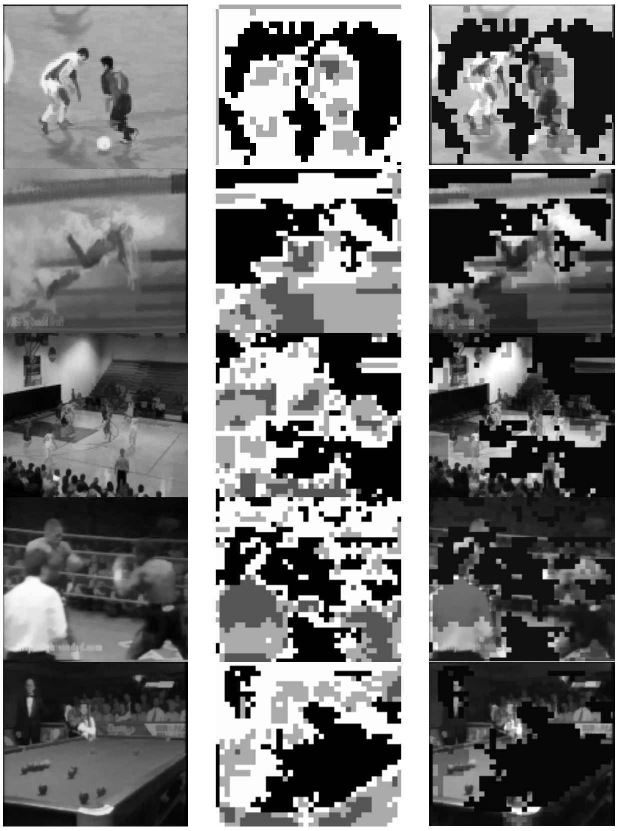

Figure 2. Indicative results on salient region extraction for soccer, swimming, basketball, boxing and snooker sequences.

B. Results

Results in the form of confusion matrices are given in Tables 1 and 2. Each row shows the classification of ground truth, with the last two being the precision and recall for each class. For example, the first row of Table 1 shows that 16 soccer video shots are misclassified into snooker shots in the case of classification without using spatiotemporal saliency region selection.

Table 1 shows that the classification of swimming, basketball and boxing are quite good. Almost no classification errors are reported for these classes. Nevertheless, the soccer class seems to be misclassified as snooker and basketball. The overall classification error on the test data for the multiclass problem in the case of no region selection is 25.58%. Table 2 shows the obtained results using spatiotemporal salient region selection. There is an improvement and the error falls to 16.28%.

Table 1. Confusion matrix of test data after classification on whole video (*testing error*: 25.58%).

	Soccer	Swim.	Basket.	Boxing	Snook.
Soccer	4	4	12	0	16
Swimming	0	36	0	0	0
Basketball	0	0	32	0	0
Boxing	0	0	4	36	0
Snooker	8	0	0	0	24
Precision	**0,333**	**0,900**	**0,667**	**1,000**	**0,600**
Recall	**0,111**	**1,000**	**0,889**	**1,000**	**0,667**

Table 2. Confusion matrix of test data after classification on salient regions (*testing error*: 16.28%).

	Soccer	Swim.	Basket.	Boxing	Snook.
Soccer	24	4	8	4	4
Swimming	0	36	0	0	0
Basketball	0	4	28	0	0
Boxing	4	0	8	24	0
Snooker	0	0	0	0	32
Precision	**0,857**	**0,818**	**0,636**	**0,857**	**0,889**
Recall	**0,545**	**0,818**	**0,636**	**0,545**	**0,727**

Although the error improvement is relatively small (more extended experiments are needed), there is an interesting result that supports our initial claim that the salient region selection may provide the feature extractor with regions that represent the video content more efficiently. Two of the classes, namely the soccer and snooker ones, have similar global characteristics due to the similar color of the playfield. The grass and the snooker table have similar green hues. This is why the soccer class is confused with the snooker one, as noticed above. However, Table 2 shows that the soccer videos are less misclassified as snooker.

Table 3. Soccer vs. Snooker classification

Method	**No Selection**	**Salient Selection**
Soccer vs. Snooker	29.41%	5.88%
Soccer vs. Basketball	23.53%	17.64%

In order to emphasize on this remark we attempted a binary classification using only these two classes for training and testing. The best classifier is selected as explained above. The results revealed the discrimination power of the proposed method. The overall testing error for saliency-based learning was much lower, as reported in Table 3. A similar experiment was conducted for the case of soccer vs. basketball, which seem to be misclassified due to the presence of audience regions that are similar in both classes. An improvement was achieved, possibly due to rejection of such regions in case they are not salient.

VI. CONCLUSIONS

In this paper we have studied and experimented on the potential of spatiotemporal saliency to enhance the performance of a SVM classifier in learning and classifying sports clips. The results are promising and show that the proposed region selection improves the classification accuracy, regardless of the simple features employed, which are independent of the specific domains tested. In the future, we plan to extend our experiments on a larger database, which will possibly include more categories of sports videos, and explore more robust features to further improve the classification performance.

Acknowledgments

This research work was supported by the European network of Excellence MUSCLE.

VII. REFERENCES

[1] Editorial, Visual Attention, Vision Research, vol. 44, pp. 1189-1191, 2004.

[2] L. Itti, C. Koch, E. Niebur, “A model of saliency-based visual attention for rapid scene analysis”, *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)* 1998, vol. 20, no. 11, pp. 1254-1259.

[3] R.M. Gray, R.A. Olshen: "Vector quantization and density estimation", Proceedings Sequences-97, Positano, Italy, June 1997.

[4] K. Fukunaga: “Introduction to statistical pattern recognition”, Second Edition, Academic Press, New York, 1990.

[5] V. Vapnik. Statistical Learning Theory. John Wiley & Sons, 1998.

[6] A. Vailaya, A. K. Jain and H.-J. Zhang: "On image classification: city images vs. landscapes", Pattern Recognition, Vol. 31, pp. 1921-1936.

[7] D.H. Wang, Q. Tian, S. Gao, W.-K. Sung, “News sports video shot classification with sports play field and motion features”, ICIP’04, pp. 2247-2250, 2004

[8] Q. Tang, J.-H. Lim, J.S. Jin, H. Sun, Q. Tian, “A generic mid-level representation for semantic video analysis”, ACM Conf. On Multimedia, pp. 33-44, 2003.

[9] M. Szummer, R. Picard: “Indoor-outdoor image classification", Proc. IEEE Workshop on Content-based Access to Image and Video Databases, Bombay, India, January 1998.

[10] H. Okamoto, Y. Yasugi, N. Babaguchi, T. Kitahashi, “Video clustering using spatio-temporal image with fixed length”, ICME’02, pp. 2002-2008, 2002

[11] U. Rutishauser, D. Walther, C. koch, P. Perona, “Is bottom-up attention useful for object recogntion?”, CVPR’04, pp. 37-44, Jul 2004

[12] Rapantzikos K., Tsapatsoulis N., Avrithis Y., “Spatiotemporal Visual Attention Architecture for Video Analysis Proc. of IEEE International Workshop On Multimedia Signal Processing (MMSP’04), Sienna, 2004

[13] Rapantzikos K., Avrithis Y., Kollias S., “Handling uncertainty in video analysis with spatiotemporal visual attention”, FUZZ-IEEE’05, Reno, Nevada, May 2005.