AN ENHANCED SPATIOTEMPORAL VISUAL ATTENTION MODEL FOR SPORTS VIDEO ANALYSIS

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

ABSTRACT

Inspired by the human visual system, visual attention (VA) models seem to provide solutions to problems of semantic image understanding by selecting only a small but representative fraction of visual input to process. Having proposed a spatiotemporal VA model for video processing in the past, we propose considerable enhancements in this paper, including the use of steerable filters for 3D orientation estimation, and of PCA for fusion of features for the construction of saliency volumes. We further employ segmentation and feature extraction on salient regions to provide video classification using an SVM classifier. Finally, we provide results on sports video classification and comment on the usefulness of spatiotemporal VA for such purposes.

1. INTRODUCTION

Considerable research has been carried out into the mechanisms of attention in the human optical system. One of the main tasks of *selective visual attention* (VA) models is to select a small fraction of important information in visual input, in a way similar to humans. *Saliency*-based attention has been computationally modeled in the last decade by Itti and Koch, [1], 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 [9], visual context analysis [18], or image retrieval [19].

Having previously suggested a spatiotemporal VA model , [10], [11],, extending existing approaches and enabling processing of video apart from still images, we now investigate the application of this model to video classification. Image and video classification is a valuable tool towards other applications like object detection and recognition, visual content description, semantic metadata generation, indexing and retrieval. Unfortunately, it is an unsolved problem that requires bridging the gap between the target *semantic classes*, and available low-level *visual descriptors*. Regardless of whether classification is supervised or unsupervised, and regardless of the specific classification model employed, e.g., expectation maximization, vector quantization, k-means, [2], support vector machines (SVM) [3], or neural networks, 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 descriptors.

Visual descriptors 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 [7] to detect e.g. vegetation (green) or sea (blue). Edge direction histograms are employed for city/landscape classification [4] since city images typically contain horizontal and vertical edges. Additional motion descriptors are also used for sports video shot classification [5], [6].

Even in specific domains and appropriately selected descriptors, classification usually fails in several cases like 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 to differentiate efficiently between different classes. This would also decrease the dependency on descriptor selection or feature formulation. Our claim, and at the same time the inspiration for this work, is that the output of our model, namely the spatiotemporally salient regions, are representative of the input video and therefore can be exploited to provide a more intuitive approach for classification purposes, especially in the absence of *a priori* knowledge or object recognition.

Our spatiotemporal VA model [10] is an extension of Itti et al.'s scheme 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 framework, we treat a video

sequence as a volume with time being the third dimension. This volume is decomposed into a set of distinct *feature volumes* such as luminance, red, green, blue, yellow hues and various spatiotemporal orientations. A normalization operator is responsible for further enhancing salient regions of each volume, while the salient volume is obtained by simply averaging the enhanced ones.

In this paper, we present an enhanced version of our spatiotemporal VA model aiming to overcome specific drawbacks of the existing one. In particular, we develop a more robust, in terms of consistent results, method for generating the 2D and 3D orientation volumes, employing steerable filters. We also propose a new normalization / fusion operator that is based on Principal Component Analysis (PCA) [2], to substitute the traditional normalization/averaging process. Further, we develop segmentation and feature extraction to generate a description of a video volume based on a set of most salient spatiotemporal regions. Subsequently, an SVM classifier uses this description to classify entire video shots, and this framework is applied to the classification of sports video.

What is most important in the novel steerable filter approach for orientation estimation, is direct relation to motion analysis. In particular, one popular set of models for the extraction and analysis of motion is based on spatiotemporal energy mechanisms [14], [12]. In these models, the squared outputs of a set of oriented spatiotemporal subband filters are combined to produce local measures of motion energy. Thus, multiple motion analysis is addressed from the standpoint of orientation analysis using steerable filters [14]., [20]. In this framework, locating and analyzing interesting events in a sequence by considering the actual spatiotemporal evolution across a large number of frames can be done without the need for, e.g., a computationally expensive optical flow estimation assuming spatial coherency (e.g. image gradients).

Section 2 provides an overview of our enhanced spatiotemporal VA model. Section 3 describes the subsequent segmentation and feature selection process, while section 4 presents the SVM employed for classification. Results on VA-based classification are given in section 5 and conclusions are drawn in section 6.

2. ENHANCED SPATIOTEMPORAL VISUAL ATTENTION

In this section we briefly describe our earlier work on spatiotemporal visual attention, [10], [11] that extends the spatial visual attention model of Itti *et al.* [1], and comment on the proposed extensions.

2.1 Spatiotemporal VA Scheme

Initially a video volume for each separate color channel (RGB) and intensity ($I = (R+G+B)/3$) by stacking each frame on of top of the other is generated. Fig. 1 illustrates the process on a simple sequence. Two trucks are moving towards opposite directions and a box in the middle occludes one of them. The initial volume $\underline{I}$ looks like the one depicted in Fig. 1b. For visualization purposes we show a transparent view of *I*. Notice the spatiotemporal evolution of the objects that is already clear for this simple example. Afterwards, all 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 perform decomposition of the video volume at a number of different spatiotemporal scales using Gaussian pyramids. The required low-pass filtering and subsampling is obtained by 3D Gaussian low-pass filters and vertical/horizontal reduction by consecutive powers of two. The final result is a hierarchy of video volumes that represent the input sequence in decreasing spatiotemporal scales. Afterwards, feature volumes for each feature of interest, including intensity, color and 2D/3D orientation are computed [11] and decomposed into multiple scales. Every volume simultaneously represents the spatial distribution and temporal evolution of the encoded feature. The pyramidal decomposition allows the model to represent smaller and larger "events" in separate subdivisions of the channels.

A center-surround operation, denoted as $\ominus$, which is suitable for detecting locations that locally stand out from their surroundings, is implemented in the model as the difference between fine and coarse scales for a given feature. For example, if $\mathbf{I}(\sigma)$ is the intensity volume at scale σ, *c* is a set of coarse scales and *s* a set of finer scales the result is obtained by:

$$\mathbf{I}(c,s) = |\mathbf{I}(c) \ominus \mathbf{I}(s)| \quad (1)$$

Afterwards, all intermediate results are processed through across-scale addition, $\oplus$, which consists of reduction of each volume to a predefined scale σ' and point-by-point addition. The final feature volumes (*conspicuity volumes)*, namely $\bar{I}$ for intensity, $\overline{C}$ for color, $\overline{O}$ for 2D orientation and $\overline{O}_{3D}$ for 3D orientation are obtained by:

$$\bar{I} = \bigoplus_{c=2}^{\sigma_i} \bigoplus_{s=2}^{\sigma_i} N(\mathbf{I}(c,s)) \quad (2)$$

$$\overline{C} = \bigoplus_{c=2}^{\sigma_i} \bigoplus_{s=2}^{\sigma_i} \left[N(\mathbf{RG}(c,s)) + N(\mathbf{BY}(c,s)) \right] \quad (3)$$

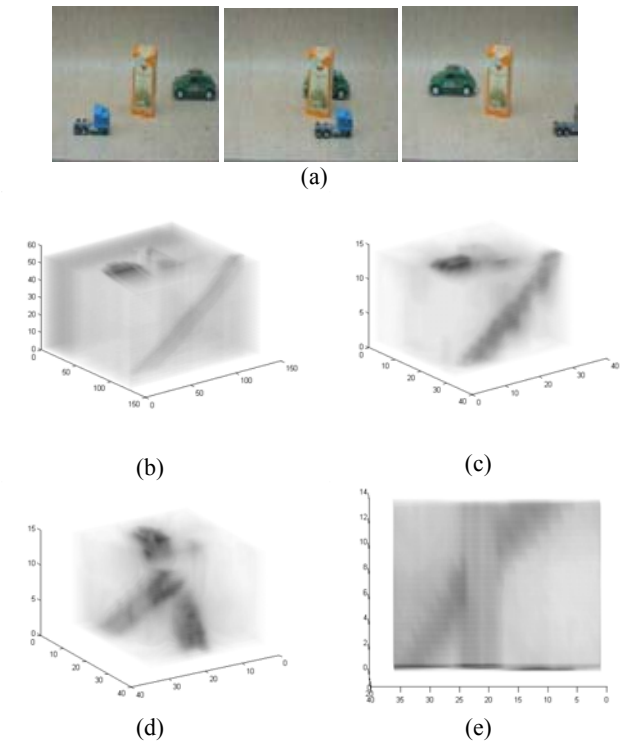

Figure 1 (a) three neighboring frames of a simple sequence; (b) initial video volume; (c)-(e) Saliency volume observed from different angles. The volumes are negative and transparent versions of the original ones (for visualization purposes)

$$\overline{O} = \sum_{\theta \in A} N\left(\bigoplus_{c=2}^{\sigma_i} \bigoplus_{s=2}^{\sigma_i} N(\mathbf{O}(c,s)) \right) \quad (4)$$

$$\overline{O}_{3D} = \sum_{\theta \in A^{3D}} N\left(\bigoplus_{c=2}^{\sigma_i} \bigoplus_{s=2}^{\sigma_i} N(\mathbf{O_{3D}}(c,s)) \right) \quad (5)$$

where N is a normalization morphological operator [11] and **RG**, **BY** stand for red-green and blue-yellow color combinations obtained by

$$\mathbf{RG}(c,s) = |(\mathrm{R}(c) - \mathrm{G}(c)) \ominus (\mathrm{G}(s) - \mathrm{R}(s))| \quad (6)$$

$$\mathbf{BY}(c,s) = |(\mathrm{B}(c) - \mathrm{Y}(c)) \ominus (\mathrm{Y}(s) - \mathrm{B}(s))| \quad (7)$$

The motivation for the creation of the separate channels and their individual normalization is the hypothesis that similar features compete strongly for saliency, while different modalities contribute independently to the saliency volume. 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 1c-e show transparent and negative versions of the saliency volume - obtained form the simple truck sequence- observed from different angles. The spatiotemporal tracks of the cars and the box have become salient. Fig. 2 illustrates all intermediate steps of the proposed model. The following subsections present the proposed changes to our previous spatiotemporal VA model [11].

2.2 Orientation

In an attempt to imitate as close as possible the receptive field sensitivity profile of orientation selective neurons in human primary visual cortex, Itti et al. used Gabor filters to obtain local orientation information. One main concern of using Gabor filters is the computational complexity. The filter should have narrow passband in the frequency domain in order to have fine orientation resolution. Hence, according to the uncertainty principle the filter should have large scale in the spatial domain. Nevertheless, fine resolution is not one of our main concerns, since the VA model exploits gross orientation information. Fine resolution is not a requirement for the VA model to capture the dominant orientations.

The actual drawback is the positive skewness in the filter responses of the Gabor wavelets [21], which becomes more important when using the 3D version of the filter. A 3D Gabor filter operating on a video volume is related to motion information. When only one motion is present in the video sequence, the maximum (in the orientation sphere) is still well localized in spite of the skewness. But if we have multiple motions, the overlapping of different filter responses, especially the overlapping of the skewness will disturb the locations of maximal values [20].

In [11] we used morphological tools for obtaining the 2d and 3d orientation volumes in an attempt to reduce the computational complexity without compromising on the results' quality. Orientation information was obtained from I using morphological processing with oriented structuring elements of the corresponding Gaussian pyramids. Although the results were satisfactory we had

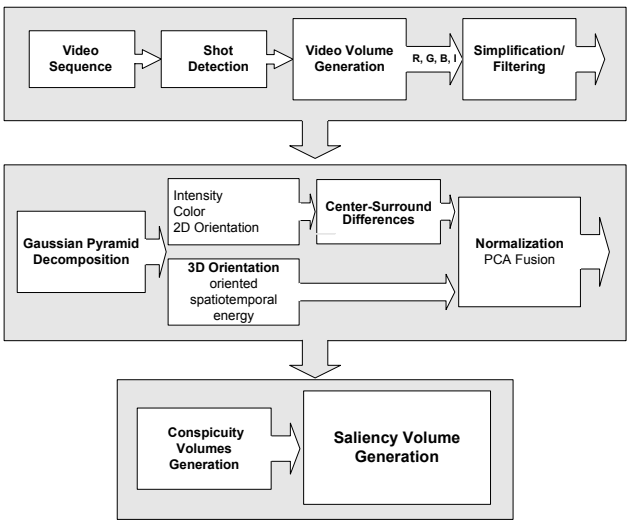

Figure 2 Spatiotemporal VA scheme

to face the difficulties in selecting the exact shape and size of the structuring elements. The proposed spatiotemporal VA model incorporates steerable filters in the orientation

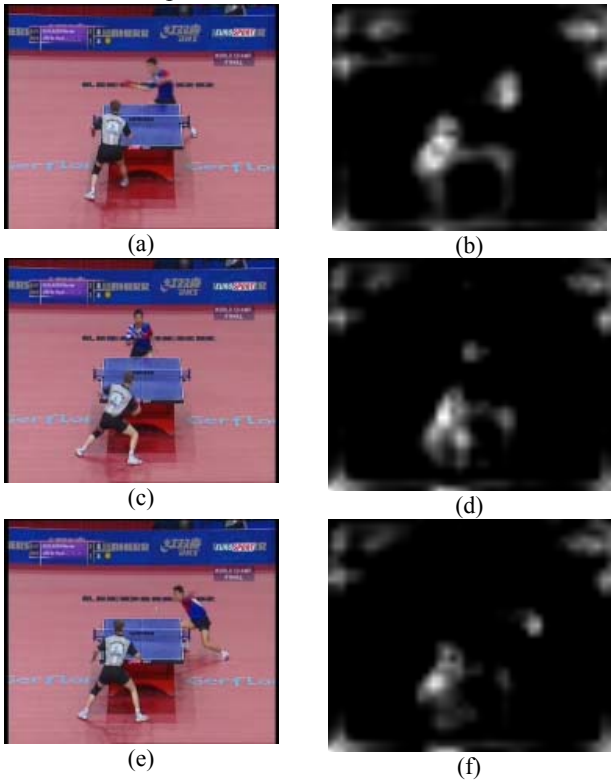

Figure 2 row-wise: original frame and the corresponding 3D orientation map.

module. Steerable filters describe a class of filters in which a filter of arbitrary orientation is synthesized as a linear combination of a set of "basis filters" [14]. The spatial orientation volume is obtained by measuring the orientation strength along a particular direction θ by the squared output of quadrature pair of bandpass filters steered to the specific angle. This is called the *oriented energy* [14] $E(\theta)$:

$$E_n(\theta) = [G_n^\theta]^2 + [H_n^\theta]^2 \quad (7)$$

where G_n^θ, H_n^θ are the n^{th} derivative of a e.g. Gaussian steered at angle θ and its Hilbert transform respectively.

Freeman *et al.* extend the filter to the spatiotemporal space and parameterize the orientation of the 3D filter kernel by the direction cosines between the axis defining the direction and the principal axes denoted as α, β, γ. In the case of axial symmetric steerable filters the functions are assumed to have an axis of rotational symmetry [13]. For example the first derivative of a 3D Gaussian function $G(x,y,z)$, whose orientation is represented with three directional angles (α, β, γ) between the axis through the filter lobes and x, y, z axis, respectively is defined as

$$G^{\alpha,\beta,\gamma}(x,y,z) = \cos(\alpha)G_x(x,y,z) + \cos(\beta)G_y(x,y,z) + \cos(\gamma)G_z(x,y,z) \quad (8)$$

where G_x, G_y, G_z are the three basis filters realized as rotated copies of G along x, y, z axis, respectively.

In the proposed orientation module we use Gaussian axial symmetric steerable filters described and implemented by [13]. Notice that the 3D orientation module-shown as an independent block in Fig. 2-functioning as a representation of inherent video motion does not pass through the center-surround differences module. Since our intention is to enhance consistent motions throughout the sequence, we by-pass this step in order to avoid penalizing long smooth motions that may be of interest in sports analysis.

Fig. 3 shows results on neighboring (10 frames apart) three frames of the same video shot, where the players are moving in various directions. Players' movements, with the one in the foreground moving longer, are distinguished from the rest of the regions. Notice the activity at the graphics region in the upper right and left corners (logos). Since the logo is present throughout the clip, its upright 3D orientation value is high

2.3 PCA normalization/fusion

There is an intrinsic difficulty in combining all the volumes resulting from the various feature extraction stages of the model. When no knowledge about the scene exists, there is no way to bias the system towards specific (salient) features. The spatiotemporal feature volumes represent *a priori* not comparable modalities, with different dynamic ranges and meaning. Due to the lack of top-down supervision (knowledge), there is a need for a fusion scheme that will enhance high activation areas and suppress others. Such a scheme will enhance the most salient subvolumes so as to prohibit non-salient regions from drastically affecting the result. There are two different types of fusion schemes found in the VA model: (a) Intra-fusion concerning fusion among volumes of the same feature (e.g. combination of volumes enhancing different orientations) and (b) inter-fusion concerning fusion among volumes of different modalities (e.g. combination of intensity, color and orientation). In our previous work, [10], [11], we applied a morphological transform N for intra-fusion and a simplified averaging operator for obtaining the final saliency volume (inter-fusion). Both operators were based on Itti *et al.*'s rationale. Although, the results were satisfactory we often obtained vague results due to the objectivity of the normalization operator's parameters and the blurring coming out of the final averaging operator.

Principal component analysis is a coordinate transformation typically associated with multivariate statistics. PCA finds orthogonal linear combinations of a set of features that maximize the variation contained

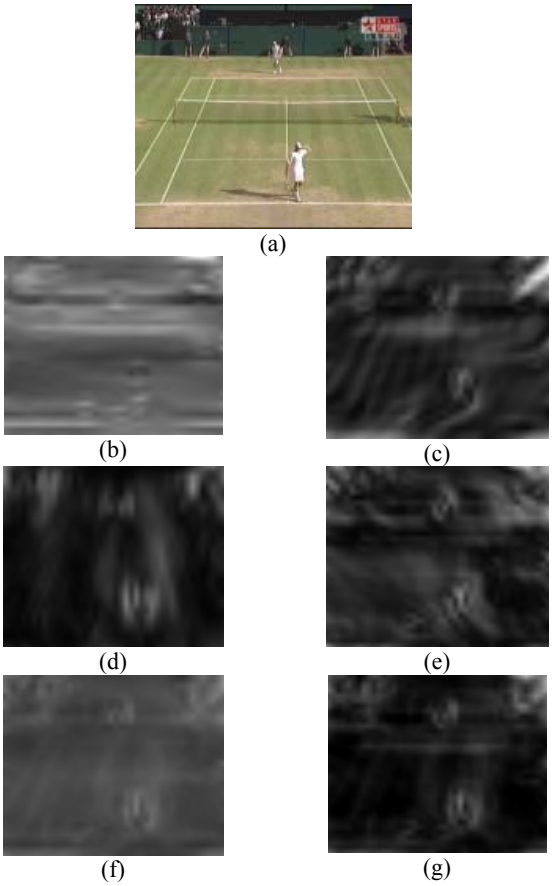

Figure 3 (a) Original frame; (c)-(e) Orientation maps at 0^o, 45^o, 90^o, 135^o respectively; (f) fusion with N and average; (g) fusion with PCA

within them, thereby displaying most of the original variation in a smaller number of dimensions.

Hence, it transforms a multidimensional space to one of an equivalent (or less) number of dimensions by creating a new series of images (components) in which the axes of the new coordinate systems point in the direction of decreasing variance. In the transformed space the first dimension contains the most variability in the data, the second the second most, and so on.

Although PCA is a common technique in the field of multi-band (e.g. spectral) imaging for contrast enhancement, visualization and compression purposes [16], [17], several researchers have used it for video content analysis and feature extraction [15]. The main strategy is to condense local spatial information and to preserve the temporal information by keeping all such reduced spatial information for all frames [15]. We use PCA for both normalization of similar content volumes and fusion of the conspicuous ones. We extract the feature volumes, reorder them into row vectors and stack them into a matrix. Hence, the matrix has one row for each input volume. PCA is applied to this matrix and we keep only the first principal component (PC1), which includes those spatiotemporal regions that contribute more to the variance of the input data. The normalized volume is obtained by rearranging PC1 vector into a 3D matrix (volume).

For example, when PCA is used as a normalization operator (N_{PCA}), we avoid the simplistic averaging one and Eq. 4 becomes

$$\overline{O} = N_{PCA}\left(\bigoplus_{c=2}^{\sigma_i}\bigoplus_{s=2}^{\sigma_i} N_{PCA}(\mathbf{O}(c,s))\right)$$

All other equations describing the model in section 2 are accordingly changed. Using N_{PCA}, rather than an operator based on local maxima detection [1], [10], [11], we ensure that we keep the data that are responsible for the most variability in the input volumes.

Figs. 4, 5 show the performance of the PCA operator. Fig. 4 b-e show the four 2D orientation bands (0^o, 45^o, 90^o, 135^o) obtained with steerable filters for a specific frame. Fig. 4f contains the result of applying the normalization operator N and fusion by averaging, while Fig. 4g is the result of applying PCA for both normalization (N_{PCA}) and fusion. Notice the difference in contrast between the last two images. PCA-based

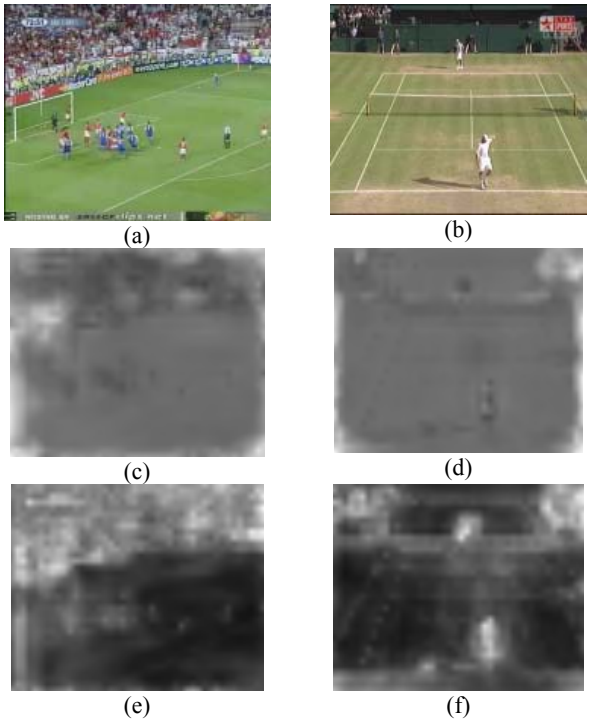

Figure 4 columnwise: original frame, average-based saliency, PCA-based saliency

enhancement/fusion performs better and enhances (retains) areas with salient activity like the players, parts of the audience and graphics. Same remarks are obtained

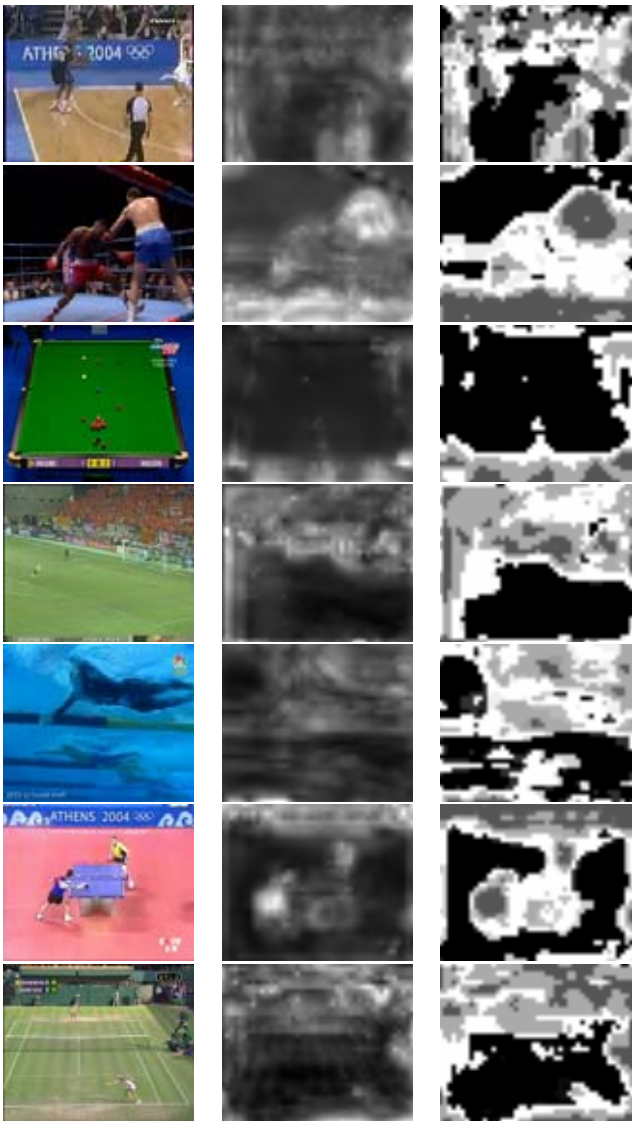

Figure 5. Indicative results on salient region extraction and segmentation for sports sequences.

from the results illustrated in Fig. 5. The original frame, the saliency obtained by averaging all

conspicuity volumes and the corresponding results using PCA are shown in a column-wise order.

3. SEGMENTATION & 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. Since our main goal is the classification of sport videos, a gross segmentation of the saliency volume should be adequate because separated objects (e.g. players, ball, goal post, tennis net etc.) are not of primary interest. 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 are extracted. Afterwards, we order the clusters in increasing order of saliency, discard the less salient one and label the rest.

Usually, the less salient regions are related to areas that are consistently present throughout the clip like the play-field or parts of the audience. Nevertheless, it is common sense that e.g. the color of the play-field may provide strong discrimination among several sport classes (e.g. basketball, soccer, swimming, table-tennis). In order to avoid misunderstandings, we emphasize that the gross segmentation achieved by the previous method does not exclude thoroughly the play-field. Large parts of it are preserved and provide the desired discrimination among the classes mentioned before.

Our claim is simple and is related to the core idea of visual attention. We expect that the less salient regions are not representative of the spatiotemporal video evolution and therefore features extracted from them 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.

4. SVM CLASIFFIER

An SVM [3] 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:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i \qquad (9)$$
$$\text{s.t. } y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i, \ \xi_i \geq 0$$

where the training data x_i are mapped to a higher dimensional space by function ϕ 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 a radial basis function (RBF) 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, which is selected for the classifiers.

5. VA-BASED CLASSIFICATION

5.1. Experimental Setup

To demonstrate the potential of the proposed scheme we select a number of videos from seven different sports. *Soccer (SO), swimming (SW), basketball (BA), boxing (BO)*, *snooker (SN), tennis (TE)* and *table-tennis (TB)* are the seven predefined classes of shots we use for conducting our experiments. The abbreviations in parentheses are used tin Tables 1, 2 that present the classification results. Most of the clips are from the Athens Olympic Games 2004. 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 the proposed algorithm. The saliency volume is clustered as described in section 3. The less salient region is discarded and features are extracted from each of the remaining salient ones. Fig. 6 shows indicative frames of each class and the obtained saliency masks corresponding to the three most salient regions. The third column shows the segmentation of the saliency mask. The black region is the less salient one, while the three gray levels represent the regions from which the features are extracted.

5.2. 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 20 snooker video shots are misclassified into soccer and 5 into basketball shots. Table 1 summarizes the results obtained by extracting features on the whole video volume without selecting salient regions, while Table 2 shows results using the spatiotemporal salient region selection with the proposed enhancements. The overall classification error on the test data for the multiclass problem in the case of no region selection is 26.37%. There is an improvement achieved through VA selection and the error falls to 15.38%.

Although the error improvement is not tremendous, 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. Pairs of classes, like soccer-snooker or basketball-table tennis, have similar global characteristics due to the similar color of the playfield and the Athens 2004 advertisements (blue-white). That's why the statistics for those classes in Table 1 are not satisfactory However, Table 2 shows improved results. In order to emphasize on this remark we attempted a binary classification using only the specific pairs of 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

Table 1. Confusion matrix of test data after classification without saliency (***overall testing error*: 26.37%**).

	SN	SW	BA	BO	SO	TE	TB
SN	35	0	5	0	20	0	0
SW	0	50	0	0	0	0	0
BA	0	0	70	5	5	10	10
BO	0	0	5	35	0	0	0
SO	10	0	5	0	60	5	0
TE	0	0	20	0	5	50	0
TB	0	0	15	0	0	0	35
Prec	**0,778**	**1,000**	**0,583**	**0,875**	**0,667**	**0,769**	**0,778**
Rec	**0,583**	**1,000**	**0,700**	**0,875**	**0,750**	**0,667**	**0,700**

for saliency-based learning was much lower, as reported in Table 3.

Table 2. Confusion matrix of test data after classification with saliency (***overall testing error*: 15.38%**).

	SN	SW	BA	BO	SO	TE	TB
SN	50	0	5	0	5	0	0
SW	0	50	0	0	0	0	0
BA	5	0	75	10	0	10	0
BO	0	0	0	35	5	0	0
SO	0	0	5	0	70	5	0
TE	0	0	5	0	0	70	0
TB	0	0	10	0	5	0	35
Prec	**0,909**	**1,000**	**0,750**	**0,778**	**0,824**	**0,824**	**1,000**
Rec	**0,833**	**1,000**	**0,750**	**0,875**	**0,875**	**0,933**	**0,700**

Table 3. Error rates of binary classifications

Method	**No Selection**	**Salient Selection**
Snooker vs. Soccer	21.43%	3.57%
Basketball vs. Table-tennis	20.00%	8.57%

6. CONCLUSIONS

In this paper we propose an enhanced version of our spatiotemporal VA scheme and experiment on its potential to improve the performance of an SVM classifier in learning and classifying video clips of seven sport classes. 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. The improvement of classification accuracy is not impressive,, but we believe that minor enhancements on the implemented model should boost further the classifier performance. For example, we noticed that graphics related to specific sports (e.g. score tables) become salient because of their repeating appearance, but are quite similar in terms of color/texture and thus do not provide enough discrimination among sports. Occasionally, the same holds for parts of the audience. Probably a filter on high texture activity (audience is highly textured) could be applied in order to get rid of those areas. Generally, ignoring such problematic regions or treating them in a different way than other ones should improve the statistics. Additionally, exploring more advanced features that fit well with the application in hand is in our future plans.

7. REFERENCES

[1] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans. on Patt. Analysis and Mach. Intell., vol. 20, no. 11, pp. 1254-1259, 1998

[2] R.O. Duda, P.E. Hart, D.G. Stork, "Pattern Classification", John Wiley and Sons, New York 2001.

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

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

[5] 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

[6] 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.

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

[8] H. Okamoto, Y. Yasugi, N. Babaguchi, T. Kitahashi, "Video clustering using spatio-temporal image with fixed length", ICME'02, pp. 2002-2008, 2002

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

[10] 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

[11] Rapantzikos K., Avrithis Y., Kollias S., "Handling uncertainty in video analysis with spatiotemporal visual attention", FUZZ-IEEE'05, accepted for publication.

[12] David J. Heeger. Optical flow using spatiotemporal filters. International Journal ofComputer Vision, pages 279{302, 1988.

[13] K.G. Derpanis, J.M. Gryn, "Three-dimensional nth derivative of Gaussian separable steerable filters", Technical report CS-2004-05, York University. Nov. 2004.

[14] W.T. Freeman, E.H. Adelson, "The Design and Use of Steerable Filters", IEEE Transactions on Pattern Analysis and Machine Intelligence., 1991.

[15] E. Sahouria, A. Zakhor, "Content analysis of video using principal components," IEEE Trans. on Circuits and Systems for Video Technology, Vol. 9, No. 8, Dec 1999.

[16] B.J. Campbell., "Introduction to Remote Sensing", second edition, The Guilford Press, New York, NY, 1996.

[17] S.J. Typ, A. Konsolakis, D.I. Diersen, R.C. Olsen, Principal-components-based display strategy for spectral imagery, IEEE Trans. On Geosci. And Remote Sens., vol. 41, no. 3, Mar 2003.

[18] A. Torralba, "Contextual Priming for Object Detection", Intern. Journal on Comp. Vis., vol. 53, no. 2, pp. 169-191, Jul 2003.

[19] A. Bamidele, F.W.M. Stentiford, J. Morphett, "An Attention Based Approach to Content Based Image Retrieval", BT. Advanced Res. Technology Journal on Intell. Spaces (Pervasive Computing), vol 22, no 3, Jul 2004.

[20] W.Yu, G. Sommer, K. Daniilidis, "Using skew Gabor filter in source signal separation and local spectral multi-orientation analysis", Image &* Vis. Computing, No. 4, 1 April 2005, pp. 377-392

[21] N.M. Grzywacz, A.L. Yuille, "A model for the estimate of local image velocity by cells in the visual cortex", Proc. Royal Society of London, B 239:129-161, 1990