Signal Processing: *Image Communication* 24 (2009) 557–571

Contents lists available at ScienceDirect

Signal Processing: *Image Communication*

journal homepage: www.elsevier.com/locate/image

Spatiotemporal saliency for video classification

Konstantinos Rapantzikos [a,*], Nicolas Tsapatsoulis [b], Yannis Avrithis [a], Stefanos Kollias [a]

[a] *School of Electrical & Computer Engineering, National Technical University of Athens, Greece*
[b] *Department Communication and Internet Studies, Cyprus University of Technology, Cyprus*

ARTICLE INFO

Article history:
Received 13 October 2007
Received in revised form
4 March 2009
Accepted 5 March 2009

Keywords:
Spatiotemporal visual saliency
Video classification

ABSTRACT

Computer vision applications often need to process only a representative part of the visual input rather than the whole image/sequence. Considerable research has been carried out into salient region detection methods based either on models emulating human visual attention (VA) mechanisms or on computational approximations. Most of the proposed methods are bottom-up and their major goal is to filter out redundant visual information. In this paper, we propose and elaborate on a saliency detection model that treats a video sequence as a spatiotemporal volume and generates a local saliency measure for each visual unit (voxel). This computation involves an optimization process incorporating inter- and intra-feature competition at the voxel level. Perceptual decomposition of the input, spatiotemporal center-surround interactions and the integration of heterogeneous feature conspicuity values are described and an experimental framework for video classification is set up. This framework consists of a series of experiments that shows the effect of saliency in classification performance and let us draw conclusions on how well the detected salient regions represent the visual input. A comparison is attempted that shows the potential of the proposed method.

1. Introduction

Rapid increase of the amount of video data necessitates the development of efficient tools for representing visual input. One of the most important tasks of representation is selecting the regions that represent best the underlying scene and discarding the rest. Recent approaches focus on extracting important image/video parts using saliency-based operators, which are either based on models inspired by the Human Visual System (HVS) [21,30,31] or on models aiming to produce state-of-the-practice results [14,43,44,51]. Saliency is typically a local measure that states how much an object, a region or a pixel stands out relative to neighboring items. This measure has given rise to a large amount of work in image/frame-based analysis with interesting results in many applications. Nevertheless, the lack of exploitation of spatiotemporal (space–time) information in most of these methods renders them not quite appropriate for promoting efficient representation of video sequences, where inter- and not intra-frame relations are most important. The concept of saliency detectors operating in spatiotemporal neighborhoods has only recently begun to be used for spatiotemporal analysis with emerging applications to video classification [17,26], event detection [20,32,39,49] and activity recognition [20,33].

Most of the saliency estimation methods using bottom–up visual attention (VA) mechanisms follow the model of Koch and Ullman and hypothesize that various visual features feed into a unique saliency map [7] that encodes the importance of each minor visual unit. The latter work along with the seminal work of Treisman et al. [4] are the ancestors of these models, since they proposed an efficient solution to attentional selection based on local contrast measures on a variety of features (intensity, color, size, etc.). Itti et al. were among the first to provide a

* Corresponding author. Tel.: +30 6974748850; fax: +30 2107722492.
E-mail address: rap@image.ntua.gr (K. Rapantzikos).

0923-5965/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.image.2009.03.002

sophisticated computational model based on the previous approach [32]. Tsotsos et al. [22] proposed a different model for attentional selection that is still based on the spatial competition of features for saliency and is closely related to current biological evidence. Nevertheless, the centralized saliency map is not the only computational alternative for bottom–up visual attention. Desimone and Duncan argue that salience is not explicitly represented by a single map, but instead is implicitly coded in a distributed manner across various feature maps that compete in parallel for saliency [16,42]. Attentional selection is then performed on the basis of top–down enhancement of the feature maps relevant to a target of interest and extinction of those that are distracting, but without an explicit computation of salience. Such approaches are mainly based on experimental evidence of interaction/competition among the different visual pathways of the HVS [13].

Applications are numerous, since saliency is a quite subjective notion and may fit with many computer vision tasks with most of them related to spatial analysis of the visual input. The computational model of Itti et al. is currently one of the most commonly used spatial attention models with several applications in target detection [30], object recognition [46] and compression [29]. Rutishauer et al. [46] investigate empirically to what extent pure bottom–up attention can extract useful information about objects and how this information can be utilized to enable unsupervised learning of objects from unlabeled images. Torralba [2,3] integrates saliency (low-level cues driven focus-of-attention) with context information (task driven focus-of-attention) and introduces a simple framework for determining regions-of-interest within a scene. Stentiford uses VA-based features for demonstrating the achieved efficiency and robustness in an image retrieval application [14]. Although the method has been tested on small sets of patterns, the results are quite promising. Ma et al. propose and implement a saliency-based model as a feasible solution for video summarization, without fully semantic understanding of video content or complex heuristic rules [51].

Most of the above approaches process the input video sequence in a frame-by-frame basis and compensate for temporal incoherency using variants of temporal smoothing or calculating optical flow for neighboring frames. Real spatiotemporal processing should exploit the fact that many interesting events in a video sequence are characterized by strong variations of the data in both the spatial and temporal dimensions. Large-scale volume representation of a video sequence, with the temporal dimension being long, has not been used often in the literature. Indicatively, Ristivojević et al. have used the volumetric representation for three-dimensional (3D) segmentation, where the notion of "object tunnel" is used to describe the volume carved out by a moving object in this volume [36]. Okamoto et al. used a similar volumetric framework for video clustering, where video shots are selected based on their spatiotemporal texture homogeneity [17].

Nevertheless, this representation has certain similarities to the spatiotemporal representation used recently for salient point and event detection. These methods use a small spatiotemporal neighborhood for detecting/selecting points of interest in a sequence. Laptev et al. build on the idea of Harris and Forstner interest point operators and propose a method to detect spatiotemporal corner points [20]. Doll'ar et al. identify the weakness of spatiotemporal corners to represent actions in certain domains (e.g., rodent behavior recognition and facial expressions) and propose a detector based on the response of Gabor filters applied both spatially and temporally [8]. Ke et al. extract volumetric features from spatiotemporal neighborhoods and construct a real-time event detector for complex actions of interest with interesting results [49]. Boiman et al. [39] and Zelnik-Manor et al. [33] have used overlapping volumetric neighborhoods for analyzing dynamic actions, detecting salient events and detecting/recognizing human activity. Their methods show the positive effect of using spatiotemporal information in all these applications.

In comparison to the saliency- and non-saliency-based approaches, we use the notion of a centralized saliency map along with an inherent feature competition scheme to provide a computational solution to the problem of region-of-interest (ROI) detection/selection in video sequences. In our framework, a video shot is represented as a solid in the three-dimensional Euclidean space, with time being the third dimension extending from the beginning to the end of the shot. Hence, the equivalent of a saliency map is a volume where each voxel has a certain value of saliency. This saliency volume is computed by defining cliques at the voxel level and use an optimization/competition procedure with constraints coming both from inter- and intra-feature level. Overall, we propose a model useful for providing computational solutions to vision problems, but not for biological predictions. In the following sections, we present the model and elaborate on various aspects including visual feature modification, normalization and fusion of the involved modalities (intensity, color and motion).

Evaluating the efficiency of a saliency operator is rather subjective and difficult, especially when the volume of the data to be processed is large. Researchers have attempted to measure the benefit in object recognition using salient operators [46] or under the presence of similarity transforms [6], but – to the authors' knowledge – no statistical results have been obtained yet for saliency extraction itself. Since any evaluation is strongly application dependent, we choose video classification as a target application to obtain objective, numerical evaluation. The experiment involves multi-class classification of several video clips, where the classification error is used as a metric for comparing a number of approaches either using saliency or not, thus providing evidence that the proposed model provides a tool for enhancing classification performance.

The underlying motivation is that if classification based on features from salient regions is improved when compared to classification without saliency, then there is strong evidence that the selected regions represent well the input sequence. In other words, we assume that if we could select regions in an image or video sequence 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 feature selection/formulation.

To summarize our contribution, we propose a novel spatiotemporal model for saliency computation on video sequences that is based on feature competition enabled through an energy minimization scheme. We evaluate the proposed method by carrying out experiments on scene classification and emphasize on the improvements that saliency brings into the task. Overall, classification based on saliency is achieved by segmenting the saliency volume, ordering them according to their saliency, extracting features from the ordered regions, and create a global descriptor to use for classification. We do not focus on selecting the best set of descriptors, but we consider a fixed set of three descriptors (intensity, color and spatiotemporal orientation) – the features we use to compute saliency – and focus on showing how to exploit histograms of these features for scene classification. Experimental evidence includes several statistical comparisons and results that show the classification performance enhancement using the proposed method against established methods including one of our early spatiotemporal visual attention methods [25,26].

The paper is organized as follows. Section 2 provides an overview of the proposed model, while Section 3 describes the methodology for evaluating the effect of saliency on video classification. In Section 4 the performance of the proposed model is evaluated against state-of-the-art methods, while conclusions are drawn in Section 5.

2. Spatiotemporal visual attention

Attending spatiotemporal events is meaningful only if these events occur inside a shot. Hence, the input is first segmented into shots using a common shot detection technique, which is based on histogram twin comparison of consequent frames [19]. Each of the resulting shots forms a volume in space–time, which is composed of a set of points $q = (x, y, t)$ in 3D Euclidean space. This volume is created by stacking consecutive video frames in time. Under this representation, point q becomes the equivalent of a voxel. Hence, a moving object in such a volume is perceived as occupying a spatiotemporal area. Fig. 1 shows a set of frames cropped from an image sequence of a woman walking along a path. Different views and slices of the spatiotemporal volume are also shown.

Fig. 2 shows an overview of the proposed model with all involved modules: feature extraction, pyramidal decomposition and normalization and computation of the conspicuity volumes (intermediate feature specific salient volumes) and of the final saliency one. The

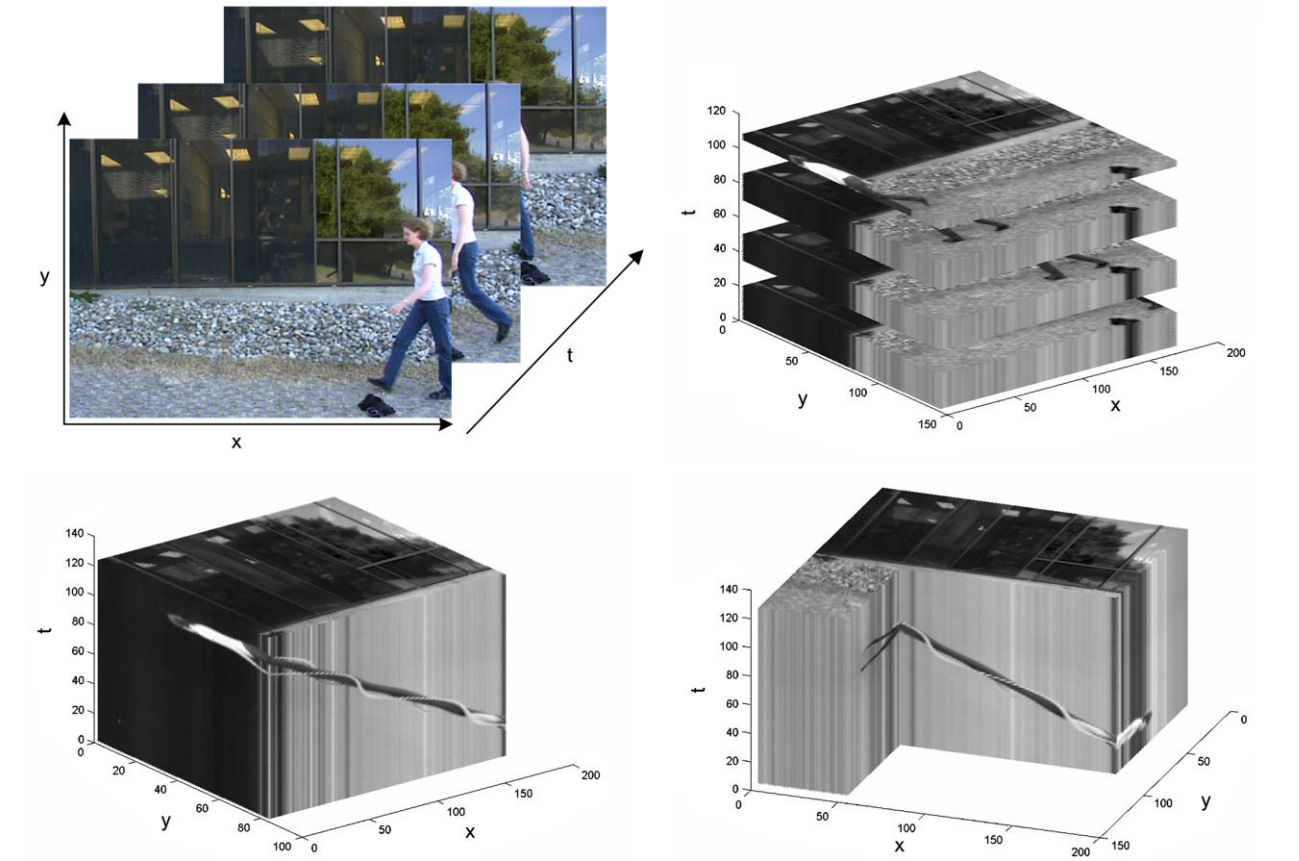

Fig. 1. Representation of a video sequence as a spatiotemporal volume and three different views.

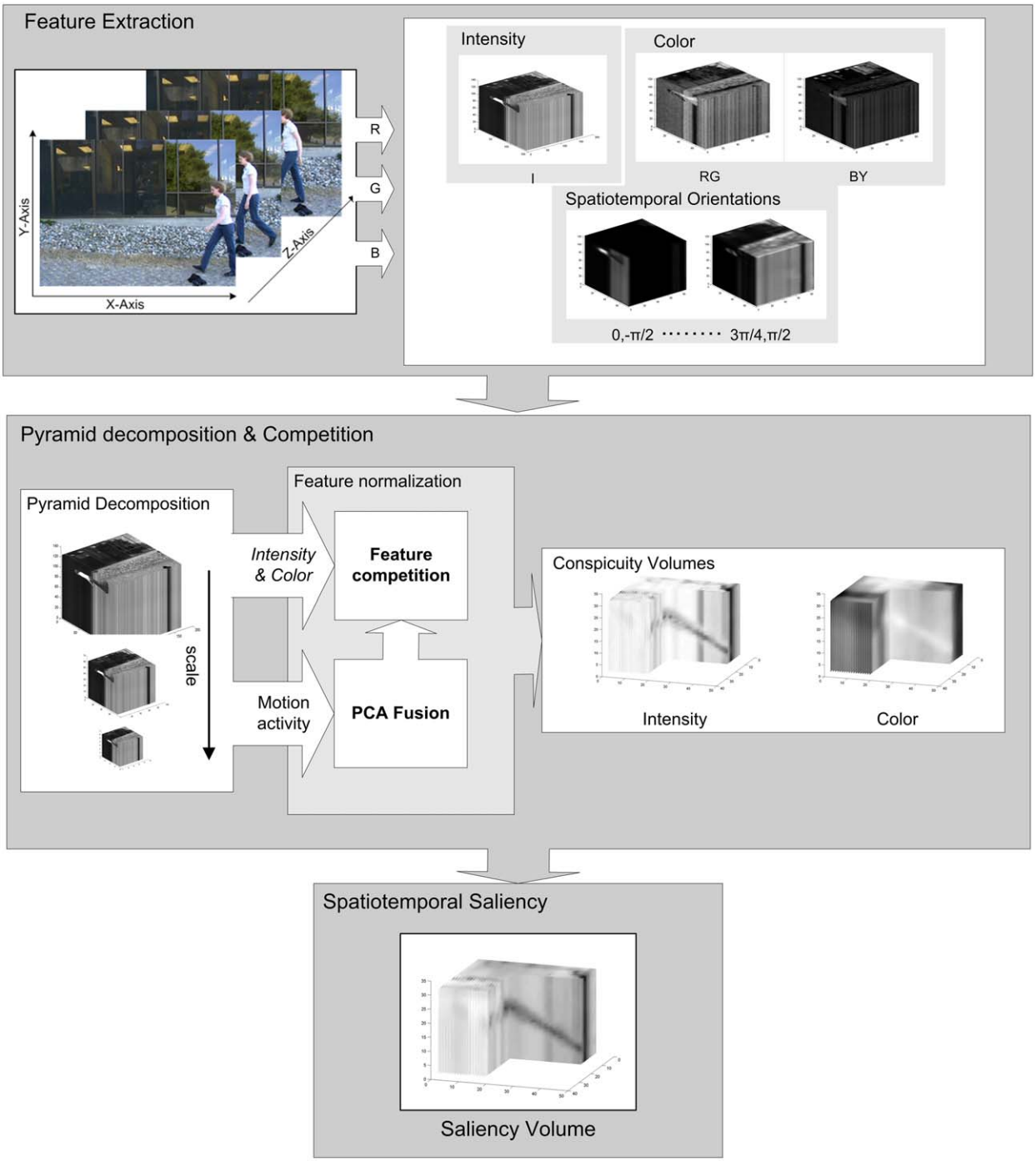

Fig. 2. Spatiotemporal saliency detection architecture.

following subsections provide an in-depth analysis for each module.

2.1. Feature volumes

The spatiotemporal volume is initially decomposed into a set of feature volumes, namely intensity, color and 3D-orientation.

2.1.1. Intensity and color

For the intensity and color features, we adopt the opponent process color theory that suggests the control of color perception by two opponent systems: a blue–yellow and a red–green mechanism [11]. The extent to which these opponent channels attract attention of humans has been previously investigated in detail, both for biological [4] and computational models of attention [50]. The color

volumes r, g and b are created by converting each color frame into its red, green and blue components, respectively, and temporally stacking them. Hence, according to the opponent color scheme the intensity is obtained by

$$I = (r + g + b)/3 \tag{1}$$

and the color ones by

$$\mathrm{RG} = \mathrm{R} - \mathrm{G} \tag{2}$$

$$\mathrm{BY} = \mathrm{B} - \mathrm{Y} \tag{3}$$

where $\mathrm{R} = r-(g+b)/2$, $\mathrm{G} = g-(r+b)/2$, $\mathrm{B} = b-(r+g)/2$ and $\mathrm{Y} = (r+g)/2-|r-g|/2-b$.

2.1.2. Spatiotemporal orientation

Spatiotemporal orientations are related to different motion directions in the video sequence. In our framework, we calculate motion activity (with no direction preference) using spatiotemporal steerable filters [48]. A steerable filter may be of arbitrary orientation and is synthesized as a linear combination of rotated versions of itself. Orientations are obtained by measuring the orientation strength along particular directions θ (the angle formed by the plane passing through the t-axis and the x–t plane) and φ (defined on the x–y plane). The desired filtering can be implemented using three-dimensional filters $G_2^{\theta,\,\phi}$ (i.e. second derivative of a 3D Gaussian) and their Hilbert transforms $H_2^{\theta,\,\phi}$ by taking the filters in quadrature to eliminate the phase sensitivity present in the output of each filter. This is called the oriented energy

$$E_v(\theta,\phi) = [G_2^{\theta,\phi} * I]^2 + [H_2^{\theta,\phi} * I]^2 \tag{4}$$

where $\theta \in \{0, (\pi/4), (\pi/2), (3\pi/4)\}$, $\phi \in \{-(\pi/2), (\pi/4), 0, (\pi/4), (\pi/2)\}$ and I is the intensity volume as defined in Section 2.1.1.

The squared outputs of a set of such oriented subband produce local measures of motion energy, and thus are directly related to motion analysis [9,48]. In the case of axial symmetric steerable filters, used in our model and proposed by Derpanis et al. [28], the functions are assumed to have an axis of rotational symmetry.

By incorporating steerable filters 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., computationally expensive optical flow estimation. Fig. 3a shows neighboring frames of the same video shot, where the players are moving in various directions. Fig. 3b shows part of the steerable filters' outputs. Each image corresponds to the slice of a specific spatiotemporal orientation corresponding to the 3D frame of Fig. 3a. Although part of the oriented filters captures accurately the movements in the scene, there is still a problem of fusing all filter outputs and producing a single volume that will represent the actual spatiotemporal movements.

Our model requires a single volume that is related to the spatiotemporal orientations of the input and that will

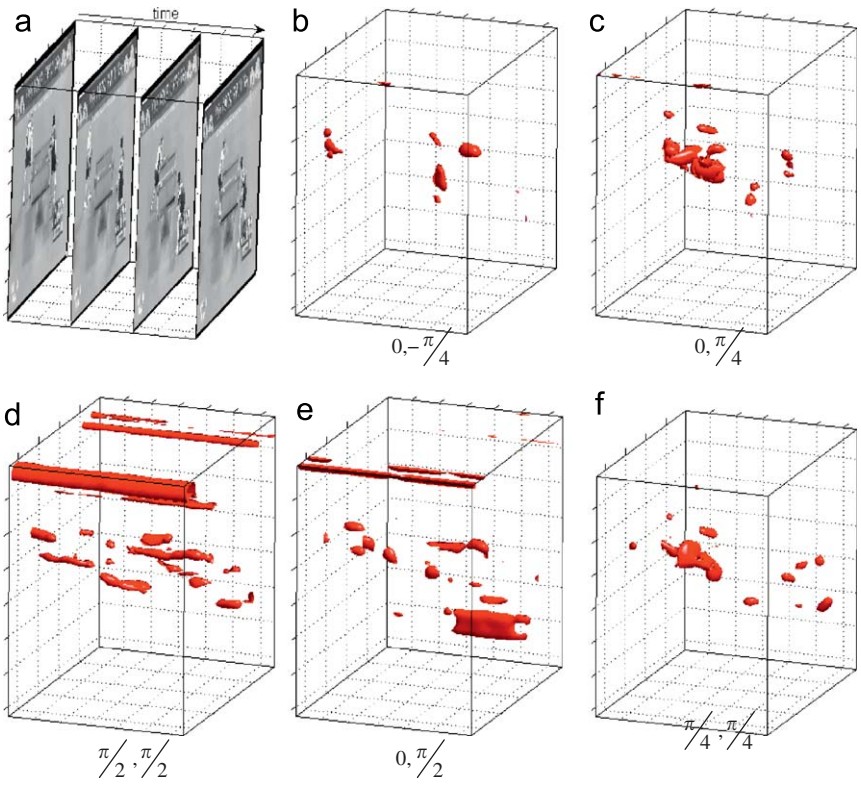

Fig. 3. Initial spatiotemporal volume and high-valued isosurfaces on various filter outputs (better viewed in color).

be remained fixed during the proposed competition procedure. By selecting θ and φ as in Eq. (4), 20 volumes of different spatiotemporal orientations are produced, which should be combined to produce a single one that will be further enhanced and compete against the rest of the features. A common strategy, adopted also by Itti et al. [31], is to produce a normalized average of all orientation bands. In our case, such a simplistic combination is prohibitive due to the large number of different bands. In this work, we use a contrast operator based on principal component analysis (PCA) and generate the spatiotemporal orientation volume V as

$$V = PCA\{E_\nu(\theta,\phi)\} \tag{5}$$

PCA finds orthogonal linear combinations of a set of n features that maximize the variation contained within them, thereby displaying most of the original variation in an equal or smaller number of dimensions sorted in decreasing order. The common strategy is to use part of the high variability data to represent the visual input [12,26]. To fuse the orientation volumes, we first create a matrix S for the set of n block vectors corresponding to the n (i.e. $n=20$) orientations and compute an n-dimensional mean vector μ. Next, the eigenvectors and eigenvalues are computed and the eigenvectors are sorted according to decreasing eigenvalue. Call these eigenvectors e_i with eigenvalues λ_i, where $i=\{1,\ldots,n\}$. The $n\times n'$ projection matrix $\mathbf{W}$ is created to contain n' eigenvectors $e_1,\ldots,e_{n'}$ corresponding to the largest eigenvalues $\lambda_1,\ldots,\lambda_{n'}$ such that $\mathbf{W}=[e_1,\ldots,e_{n'}]$ and the full data set is transformed according to $\mathbf{S}'=W^t(\mathbf{S}-\boldsymbol{\mu})$ so that the coordinates of the initial data set become decorrelated after the transformation [42]. Finally, we keep the average of the first two principal components (the transformed dimensions) that account for most of the variance in the initial data set [41].

2.1.3. Pyramid decomposition of feature volumes

As discussed above, a set of feature volumes for each video shot is generated after proper video decomposition. A multi-scale representation of these volumes is then obtained using Gaussian pyramids. Each level of the pyramid consists of a 3D smoothed and subsampled version of the original video volume. 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 represents the input sequence in decreasing spatiotemporal scales. Every volume simultaneously represents the spatial distribution and temporal evolution of the encoded feature. The pyramidal decomposition of the volumes allows the model to represent shorter and longer "events" in separate scales and enables reasoning about longer term dynamics.

Hence, a set $\mathbf{F}=\{F_{\ell,k}\}$ is created with $k=1,2,3$ and $\ell=1,\ldots,L$. This set represents the coarse-to-fine hierarchy of maximum scale L discussed before with $F_{0,k}$ corresponding to the initial volume of each of the features. Each level of the pyramid is obtained by convolution with an isotropic 3D Gaussian and dyadic down-sampling.

2.2. Spatiotemporal feature competition

Several computational variants have been proposed in the literature for detecting salient regions, i.e. regions that locally pop-out from their surroundings, with the Difference-of-Gaussian (DoG) and Laplacian-of-Gaussian (LoG) being used very often [18,31,35]. In the past, we have used a simple spatiotemporal center-surround difference (CSD) operator based on DoG and implemented it in the model as the difference between fine and coarse scales for a given feature [24,25]. Nevertheless, most of the existing models do not count in efficiently the competition among different features, which according to experimental evidence has its biological counterpart in the HVS [13] (interaction/competition among the different visual pathways related to motion/depth (M pathway) and gestalt/depth/color (P pathway), respectively. In this paper, we propose an iterative minimization scheme that acts on 3D local regions and is based on center-surround inhibition regularized by inter- and intra-feature constraints biased from motion. In our framework, motion activity volume $\tilde{V}$ is obtained by across-scale addition $\oplus$, which consists of reduction of each volume to a predefined scale σ' and point-by-point addition of the reduced volumes

$$\tilde{V} = T\left[\bigoplus_{\ell=1}^{L} V_\ell\right] \tag{6}$$

T is an enhancement operator used to avoid excessive growth of the average mean conspicuity level after the addition operation. In our implementation, we use a simple top-hat operator with a 3D-connected structuring element.

2.2.1. Energy formulation

We formulate the problem by en energy optimization scheme. An energy measure is designed, which consists of a set of constraints related to established notion of saliency, i.e. regions become prominent when they differ from their local surrounding and exhibit motion activity. In a regularization framework, the first term of this energy measure may be regarded as the data term (E_D) and the second as the smoothness one (E_S), since it regularizes the current estimate by restricting the class of admissible solutions [5,27]. Hence, for each voxel q at scale c the energy is defined as

$$E(\mathbf{F}) = \lambda_D \cdot E_D(\mathbf{F}) + \lambda_S \cdot E_S(\mathbf{F}) \tag{7}$$

where λ_D, λ_S are the importance weighting factors for each of the involved terms.

The first term of Eq. (7), E_D, is defined as

$$E_D(\mathbf{F}) = \sum_{c=1}^{L-d}\sum_{k=1}^{3} F_{c,k}(q)\cdot|F_{c,k}(q)-F_{h,k}(q)| \tag{8}$$

where c and h correspond to the center and surround pyramid scales, i.e. to a coarse and a corresponding finer scale of the representation. If the center is at scale $c\in\{1,\ldots,L-d\}$ then the surround is the scale $h=c+\delta$ with $\delta\in\{1,2,\ldots,d\}$, where d is the desired depth of the center-surround scheme. Notice that the first element of set c is the second scale for reasons of low computational

complexity. The difference at each voxel is obtained after interpolating $F_{h,k}$ to the size of the coarser scale. This term promotes areas that differ from their spatiotemporal surroundings and therefore attract our attention. If a voxel changes value across scales, then it will become more salient, i.e., we put emphasis on areas that pop-out in scale-space.

The second term, $E_S(\mathbf{F})$, is a regularizing term that involves competition among voxel neighborhoods of the same volume, so as to allow a voxel to increase its saliency value only if the activity of its surroundings is low enough. Additionally, this term involves a motion-based regularizing term to bias the feature towards moving regions. This term promotes areas that exhibit both intra-feature and motion activity. Due to lack of prior knowledge, we define the surrounding neighborhood N_q to be the set of 26 3D-connected neighbors of each voxel q excluding the closest 6 3D-connected ones and define the second energy term as

$$E_S(\mathbf{F}) = \sum_{c=1}^{L-d}\sum_{k=1}^{3} F_{c,k}(q)\cdot\frac{1}{|N(q)|}\cdot\sum_{\substack{r\in N(q)\\ r\neq q}}(F_{c,k}(r)+\tilde{V}(r)) \tag{9}$$

2.2.2. Energy minimization

Local minima of Eq. (7) may be found using any number of descent methods. For simplicity, we adopt a simple gradient descent algorithm. The value of each feature voxel $F_{c,k}(q)$ is changed along a search direction, driving the value in the direction of the estimated energy minimum

$$F^{\tau}_{c,k}(q) = F^{\tau-1}_{c,k}(q) + \Delta F^{\tau-1}_{c,k}(q)$$
$$\Delta F^{\tau}_{c,k}(q) = -\gamma(\tau)\cdot\frac{\partial E_t(\mathbf{F}^{\tau-1})}{\partial F^{\tau-1}_{c,k}(q)} + \mu\cdot\Delta F^{\tau-1}_{c,k}(q) \tag{10}$$

where $\gamma(\tau)$ is a variable learning rate and μ a momentum term to make the algorithm more stable [34].

Given both terms of the energy function to be minimized, the partial derivative may be computed as

$$\begin{aligned}\frac{\partial E}{\partial F_{c,k}(q)} &= \lambda_D\cdot\frac{\partial E_D}{\partial F_{c,k}(q)} + \lambda_S\cdot\frac{\partial E_S}{\partial F_{c,k}(q)}\\ &= \lambda_D\cdot(|F_{c,k}(q)-F_{h,k}(q)| + sign(F_{c,k}(q))\cdot F_{c,k}(q))\\ &\quad+\lambda_S\cdot\frac{1}{|N_q|}\cdot\sum_{r\in N_q}(F_{c,k}(r)+\tilde{V}(r))\end{aligned} \tag{11}$$

The learning parameter $\gamma(t)$ in Eq. (11) is important both for stability and speed of convergence. In our implementation, we use a varying γ that depends on the sign of $\partial E/\partial F_{c,k}(q)$ and the current value of ${F_{c,k}}^{\tau}(q)$

$$\gamma = \begin{cases} 1-\xi\cdot F^{\tau}_{c,k}(q) & \text{if } \dfrac{\partial E}{\partial F_{c,k}(q)} \geqslant 0\\ \xi\cdot F^{\tau}_{c,k}(q) & \text{otherwise}\end{cases} \tag{12}$$

We normalize each ${F_{c,k}}^{t}$ to lie in the range [0,1] so that the value of γ lies also in the same range. For this reason, it can be seen as a coefficient which reduces the value of the

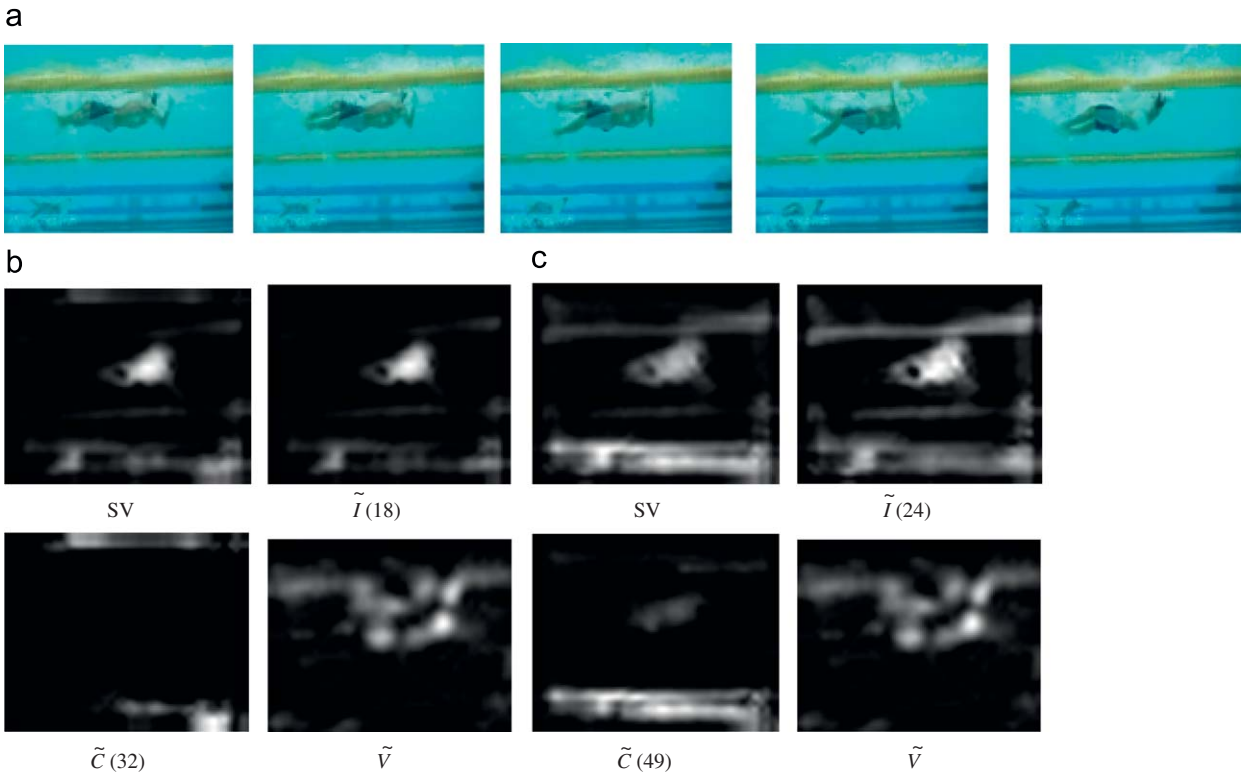

Fig. 4. (a) Frames from the same swimming sequence; (b)–(c) saliency (SV) and conspicuity maps corresponding to the middle frame for $\xi = 1$ and $\xi = 0.5$, respectively; The motion map $\tilde{V}$ is the same and the numbers in parentheses correspond to number of total iterations. (All images are resized and min–max normalized for visualization purposes.)

increment/decrement to reach quickly the desired solution. The parameter ξ controls the size of the learning step and consequently the rate of convergence and the extent of the resulting salient regions. Fig. 4 shows neighboring frames of a swimming sequence along with the derived saliency (SV) and conspicuity maps. The conspicuity maps correspond to intensity, color and spatiotemporal orientation, respectively, and are shown for two different values of ξ. All images are from the slice corresponding to the middle frame of Fig. 4a. For our implementation, we fix this parameter to $\xi = 1$. Practically, few iterations are enough for the estimate to approach a stable solution as shown by the numbers in parenthesis in Fig. 4.

2.3. Conspicuity and saliency generation

To compute the final saliency volume, the conspicuity ones should be appropriately combined. The optimization procedure we adopt produce noise-free results, and thus simple addition of the outputs is adequate. We create conspicuity volumes for the intensity and color features using the same procedure as in Section 2.2.1. The conspicuity volume for the intensity feature is obtained by

$$\tilde{I} = \bigoplus_{c=1}^{L-d} I_c \tag{13}$$

while the color one by combining the *RG* and *BY* channels

$$\tilde{C} = \bigoplus_{c=1}^{L-d} RG_c + \bigoplus_{c=1}^{L-d} BY_c \tag{14}$$

Finally, a linking stage fuses the separate volumes by simple addition and produces a saliency volume that encodes saliency at each voxel as a gray level value

$$SV = \tfrac{1}{2}(\tilde{I} + \tilde{C}) \tag{15}$$

Saliency volumes for a swimming and tennis sequence are shown in Fig. 5. The red isosurfaces correspond to high values of the saliency volume and roughly enclose the most prominent parts of the scenes like the swimmers/players, the TV logos and score boards.

3. Evaluating the effect of saliency on video classification

3.1. Saliency-based classification

As mentioned in the introduction, evaluating the performance of a saliency detector is rather subjective.

To the extent of authors' knowledge there is no benchmarking data publicly available that fits well with such kind of evaluation. Nevertheless, we do not attempt to evaluate attention itself, but rather to measure the

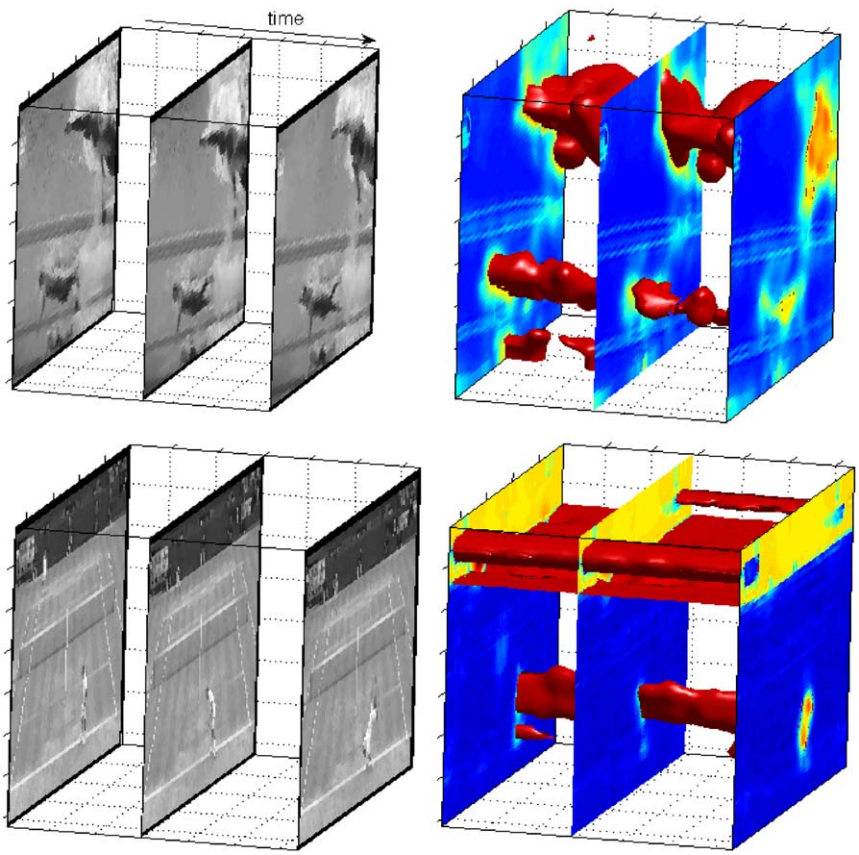

Fig. 5. Examples of slices from the original volume and the corresponding slices from the computed saliency volume. High-valued isosurfaces (in red) on the saliency volume are generated in order to make the most salient regions evident. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

effect of saliency detection in a common computer vision task like classification. We choose to evaluate the performance of the spatiotemporal saliency method by setting up a multi-class video classification experiment and observing the classification error's increase/decrease when compared against other techniques. Input data consists of several sports clips (see Section 4.1), which are collected and manually annotated by the authors.

Obtaining a meaningful spatiotemporal segmentation of a video sequence is not a simple and straightforward task. Nevertheless, if this segmentation is saliency driven, namely if regions of low (or high) saliency should be treated similarly, segmentation becomes easier. The core idea is to incrementally discard regions of similar saliency starting from high values and watch the impact on the classification performance. This procedure may seem contradictory, since the goal of attention approaches is to focus on high- rather than low-saliency areas. In this paper, we exploit the dual problem of attending low-saliency regions. These regions are quite representative, since they are consistent through the shot and are, therefore, important for recognizing the scene (playfield, slowly changing events, etc.). To support this approach, we have to place a soft requirement: regions related to background of the scene should cover a larger area than regions belonging to the foreground. Under this requirement, low-salient regions are related to the background or generally to regions that do not contribute much to the instantaneous interpretation of the observed scene.

The feature extraction stage calculates histograms of the primary features used for computing saliency, namely color, orientation and motion. To keep the feature space low, we calculate the histograms by quantizing them in a small number of bins and form the final feature vector. We use SVM for classifying the data [47]. Given a training set of instance-label pairs (x_i, y_i), $i = 1, \ldots, l$, where $x_i \in \Re^n$ and $y \in \{-1,1\}^l$, the SVMs require the solution of the following optimization problem:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} g_i \; s.t. y_i (w^T \phi(x_i) + b) \geqslant 1 - g_i, \; g_i \geqslant 0 \tag{16}$$

where the training data x_i are mapped to a higher dimensional space by function ϕ and the second term of Eq. (16) is the penalty term with parameter C. Training data correspond to the feature vectors extracted from the salient regions (the length of the feature vector depends on the experiment) and function ϕ is the radial basis function (RBF). Parameters for the RBF kernel are selected by performing a "grid-search" on the regularization parameter $C = \{2^0, 2^1, 2^2, 2^3, 2^4\}$ using five-fold cross validation estimation of the multi-class generalization performance. 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 that is used for the classifiers.

To sum up in a few words, the input video sequence is segmented into one or more regions after discarding a percentage of high-saliency voxels and histograms of pre-calculated features are extracted for each of them. Feature vectors feed an SVM classifier and the outputs of all methods are compared.

3.2. Evaluation of classification performance

To test the robustness and efficiency of the proposed model, we compare it against a method based on a simple heuristic, two methods that share a common notion of saliency, against our early spatiotemporal visual attention model and a fifth one, which is based on PCA and has proven its efficiency in background subtraction approaches [1,38]. The reason of including the last method is two-fold: first to confirm the correctness of our assumption, namely that the background of a scene is more important in recognizing it (meaning that an efficient background subtraction technique should lead to a low classification error) and second, to provide a state-of-the-art comparison. The PCA-based technique is composed of an eigenvalue decomposition stage and the rejection of the eigenvectors that correspond to small eigenvalues. The main idea is that moving objects are typically small, so they do not contribute significantly to the model. Consequently, the portions of a video sequence containing moving objects cannot be well described by this eigenspace model, whereas the static portions of the video sequence can be accurately described as a sum of the various eigenbasis vectors [38].

Our early visual attention model shared the same notion of spatiotemporal saliency, but without the feature competition module. This model has proven its efficiency in enhancing performance of a video classification system [26]. The two other saliency-based methods are the state-of-the-art static saliency-based approach of Itti et al. [30,31], and an extension using a motion map [22], as proposed in the past by several researchers [23,32,45]. Both methods produce a saliency measure per pixel. The static saliency-based approach processes the video sequences in a per frame basis. After producing a saliency map for each frame, we generate a saliency volume by stacking them together. To be fair, we filter this volume with a 3D median filter to improve temporal coherency. The motion map of the extended approach is derived using the motion estimation technique of Black and Annandan, which is based on robust statistics [37]. The same procedure for producing a saliency volume is followed for the PCA-based technique. For the sake of completeness, we also provide results of a method that operates in a heuristic way and is based on the fact that people pay often more attention to the region near the center of the view [15]. At each time step, the initial video volume is incrementally reduced by $p\%$ and a classification error is produced. The reduction is done spatially in an uniform way, which means that we reduce the extent of x–y axes from the edges to the center and leave the temporal dimension intact.

3.3. Experimental methods

In Section 4, we attempt to prove the benefit obtained using saliency by two different experiments. Each of them

is carried out on the same dataset and exploits each methods' results in a different way. The first approach is illustrated in Fig. 6(a) and is composed of 3 steps: (1) discard p% of high-saliency voxels; (2) extract histograms of pre-calculated features; and (3) feed the features to a classifier and obtain an error for each p value. The saliency volume is segmented into two regions, namely a high- and a low-salient one using Otsu thresholding [40] driven by the percentage of the high-saliency pixels to retain. Practically, the volume is iteratively thresholded using a small threshold step until the desired percentage of discarded pixels is approximately obtained. At each step, a salient and a non-salient region are produced. The feature vector generated from features bound to the less salient region is always of the same size and is formed by encoding the color histograms using 32 bins per color channel (i.e., 96 elements per region), and the motion/2D-orientation features using 16 bins. The total size of each feature vector is thus 112.

Intuitively, there exist a number of regions that represent best the underlying scene. For example, in case of sport clips, one region may be representative of the playfield, another one may include the players, the advertisements, the audience, etc. Each of these regions corresponds to a single scene property, but all of them provide a complete scene description. If we follow the reasoning of the previous experiment, we expect that if the appropriate regions are selected, the classification error would be further reduced. Hence, the second experiment segments the saliency volume into a varying number of regions (# clusters) as shown in Fig. 6b. The same incremental procedure is applied with the saliency volume being segmented into more than two regions at each iteration. After segmenting the input, the resulting regions are ordered in terms of saliency and the most salient one is discarded. This scenario has an intrinsic difficulty, since, if the number of regions is not constant for each video clip, the size of the feature vector will not be constant. Thus, direct comparison between vectors of different clips would not be straightforward. To overcome this problem, we segment the saliency volume into pre-determined number of regions using unsupervised clustering. In this framework, we use a clustering technique that allows for non-hard thresholding and labeling. *K-means* is used to partition the saliency volume into regions of different saliency. Voxels are clustered in terms of their saliency value and a predefined number of clusters are extracted. Afterwards, we order the clusters in increasing order of saliency, discard the last one, and label the rest using 3D connectivity. The optimal number of clusters, in terms of classification error minimization, is found using ROC curve analysis. At this scenario, 8 bins

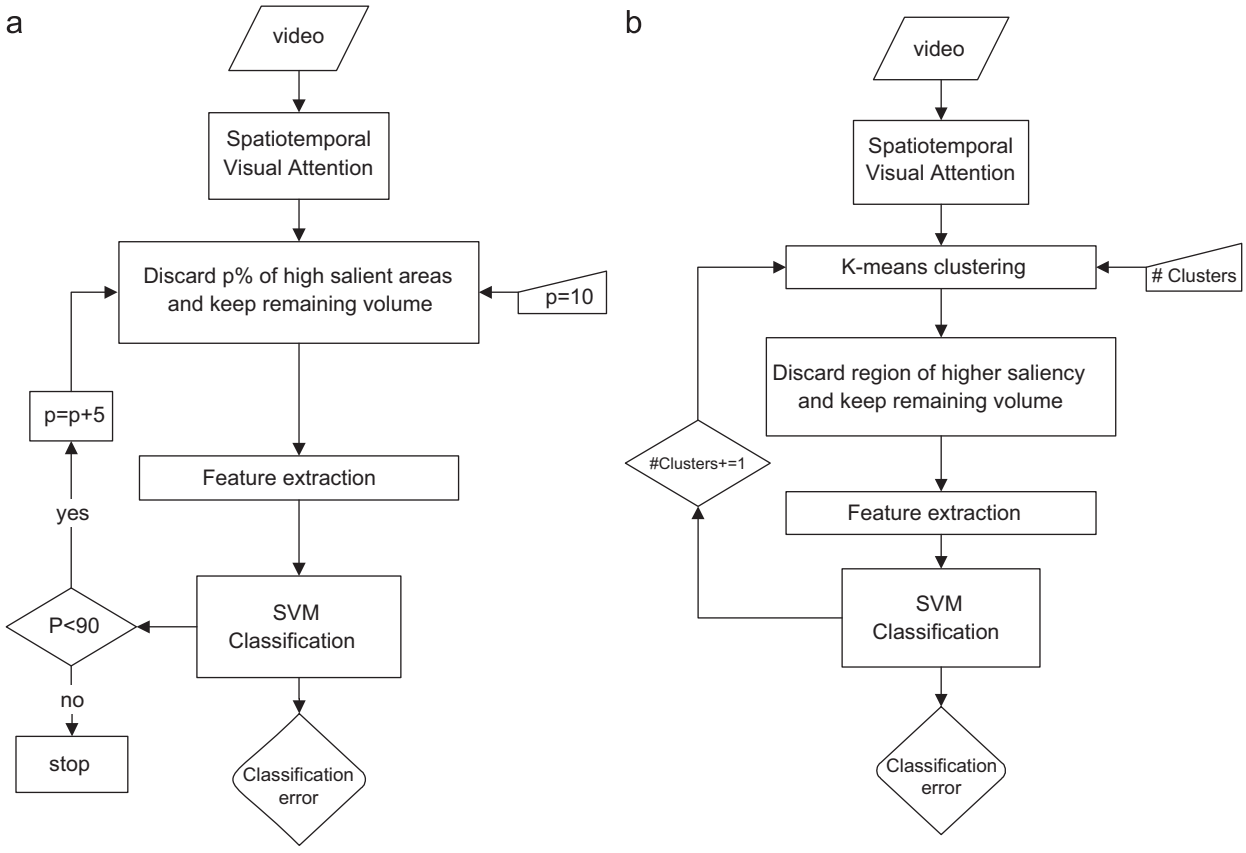

Fig. 6. Saliency-based classification: (a) based on foreground/background detection; and (b) based on >1 salient regions.

per color channel (i.e., 24 elements per region) and 4 bins for motion/2D-orientation are used. Hence, the total size of the feature vector is 32 for each region.

4. Experimental results

4.1. Experimental setup

To demonstrate the potential of the proposed scheme, we collected 924 video shots from seven different sport clips. Soccer (SO), swimming (SW), basketball (BA), boxing (BO), snooker (SN), tennis (TE) and table-tennis (TB) are the seven classes of shots we use for conducting our experiments (The abbreviations in parentheses are used to present the classification results in the tables). Most of the clips are from the Athens Olympic Games 2004. Each class includes far- and near-field views and frames where all the playfield, players and part of the audience are present. Additionally, many clips include minor camera motions (pan, zoom) due to the diversity of the video clips and the entailed errors in shot detection. The length of the shots ranges from 5 to 7 s ($\approx$120–168 frames). All clips, each consisting of a single shot, are resized to have the same spatial dimensions and manually annotated as belonging to either of the given classes. The spatiotemporal saliency volume was obtained using the proposed algorithm. Fig. 7 shows indicative frames of each class. The third column shows the segmentation of the saliency mask using automatic thresholding and the fourth one shows the segmentation of the saliency map into three regions using unsupervised clustering. The darker regions correspond to the least salient ones.

The proposed model and PCA-based background subtraction along with SVM classification are implemented in Matlab, while the saliency maps for Itti et al.'s approach are obtained using the saliency toolbox that is publicly available [10]. To acquire a better feeling about the results, e.g., video sequences with the corresponding saliency volumes are available at http://www.image.ntua.gr/~rap/saliency/.

4.2. Results

The following subsections provide a rigorous statistical analysis of the results that aid the evaluation/comparison of the techniques discussed so far. The main criteria are the precision/recall measures and various statistics on the classification error $\varepsilon(r)$, where r stands for the dependent variable. The statistics and the corresponding equations are given below

$$\mu_{err} = \frac{1}{\nu \cdot |r|} \sum_{\nu} \sum_{r} \varepsilon(r) \quad \text{Average error of } \nu\text{-fold validation}$$

$$\sigma_{err} = \frac{1}{|r|} \sum_{r} \left(\varepsilon(r) - \frac{1}{|r|} \sum_{r} \varepsilon(r) \right)^2$$

Standard deviation of ν-fold validation

$$MCE = \arg\min_{r}(\varepsilon(r)) \quad \text{Minimum classification error}$$

$$\bar{\sigma}_{err} = \frac{1}{|r|} \left(\frac{1}{\sqrt{\nu}} \sum_{r} (\varepsilon(r) - \frac{1}{|r|} \sum_{r} \varepsilon(r))^2 \right)$$

Average standard error of ν-fold validation

$fp/(tp+fp)$ Classification error

$tp/(tp+fp)$ Precision

$tp/(tp+fn)$ Recall

where *tp, fn* and *fp* are the true-positive, false-negative and false-positive rates, respectively.

4.2.1. Classification based on salient regions

A spontaneous question when dealing with saliency detection approaches in images/video sequences is how useful they are; especially when they are based on subjective measures of saliency determined by current constrained biological evidence. This section explores the effect of the proposed model when using it for video classification.

Fig. 8 shows the classification error plot for all tested methods. Each point on the graph represents the error at the specific ratio of discarded voxels and is obtained after a five-fold cross validation classification procedure. In case of the heuristic method, the ratio represents the portion of the discarded regions starting from the borders (see Section 3.2). The short lines at each point correspond to the standard error intervals obtained after the cross validation procedure.

Both Itti et al. and ITTI-motion methods provide similar results in terms of absolute error after a 30% ratio is discarded, with the ITTI-motion having higher $\bar{\sigma}_{err}$ imposing, therefore, higher uncertainty on the results. Rapantzikos et al. approach performs almost between the proposed model and the Itti et al.'s one.

The PCA-based approach has stronger fluctuations (notice the abrupt changes from 10% to 20% and 30% to 40%, respectively), but achieves a slightly smaller classification error at $p \approx 55\%$. The spatiotemporal saliency model produces an error that is always smaller than the rest with each $\bar{\sigma}_{err}$ interval being always smaller. Table 1 provides statistics for all tested methods. Tables 2 and 3 provide the confusion matrices for the PCA and the proposed methods, respectively, at the ratio, where both achieve the lowest error ($p \approx 55\%$). Although the global error improvement is not high, 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 basketball–boxing or basketball–table-tennis, have similar global characteristics due to the similar color of the playfield and the Athens 2004 advertisements (blue–white). Careful interpretation of the confusion matrices reveals the fact that the PCA method differentiates these pair of classes less efficiently than the proposed one.

Overall, classification driven by spatiotemporal saliency results to improved statistics as can be also seen from the precision/recall values in Tables 2 and 3. Although differences in *MCE* magnitude are not tremendous, two

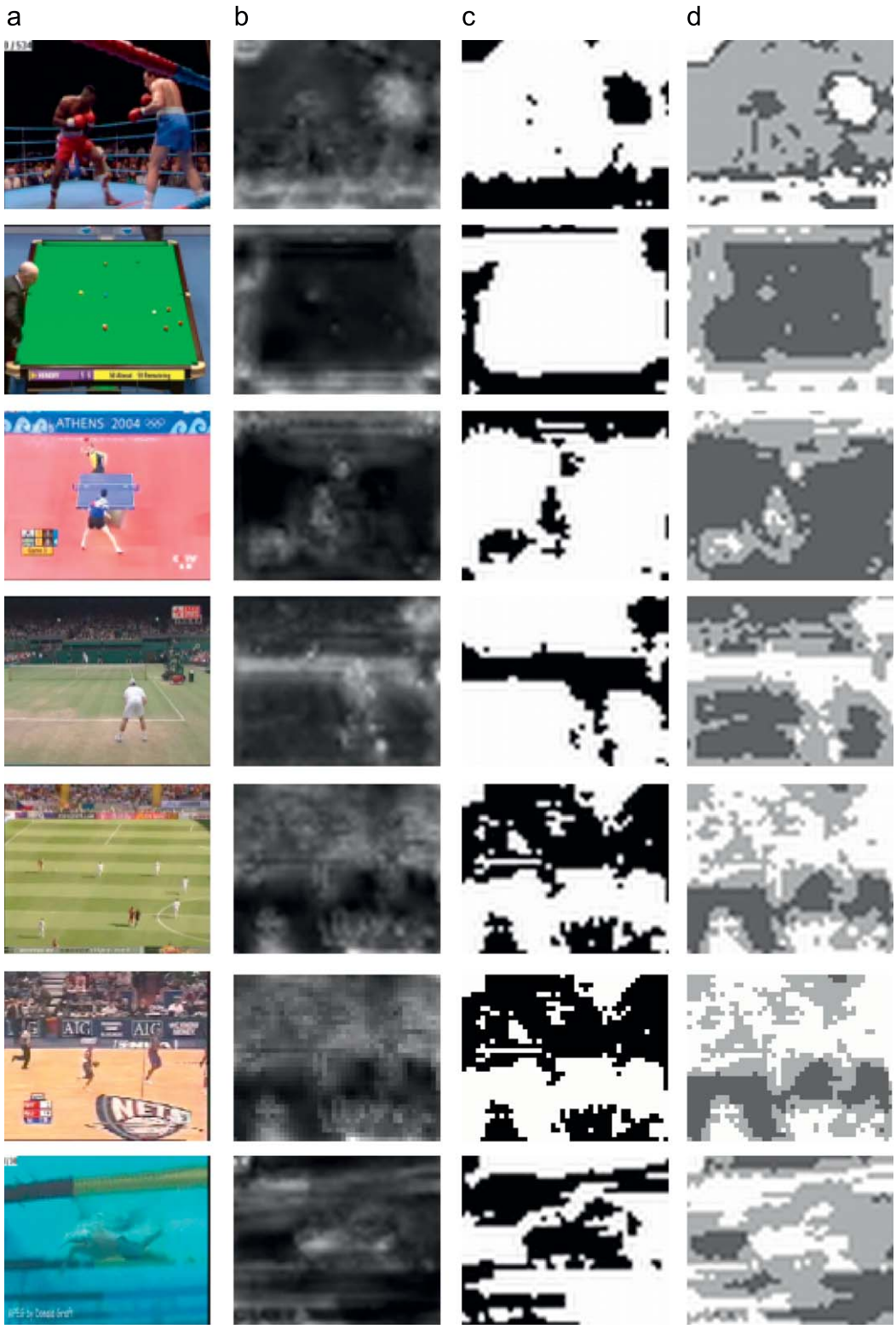

Fig. 7. (a) Original frame; (b) saliency map; (c) thresholded saliency mask; and (d) segmented saliency mask.

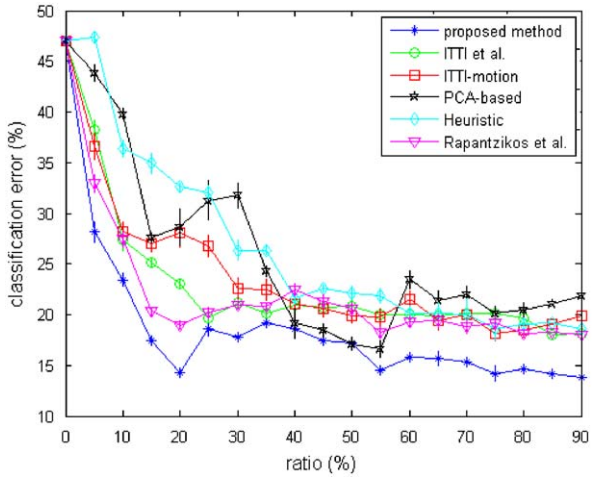

Fig. 8. Experiment I—classification error along with standard error intervals for all tested methods when varying the size of the discarded region (ratio represents the percent of discarded high-saliency voxels).

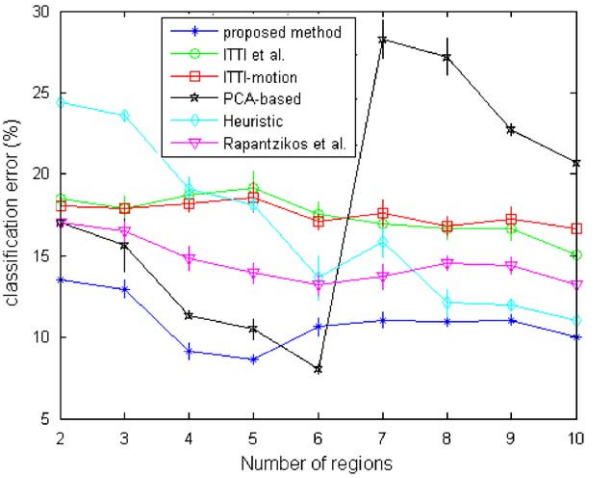

Fig. 9. Classification error along with standard error intervals when varying the number of regions (error versus number of regions used to segment the volumes).

Table 1
Statistics on saliency-based classification.

	MCE (%)	μ_{err} (%)	$\tilde{\sigma}_{err}$ (%)	σ_{err} (%)
Proposed method	**13.78±0. 14**	**16.58**	**0.48**	2.51
Itti et al.	17.96±0.63	20.86	0.67	2.35
ITTI-motion	18.10±0.93	21.93	0.93	3.43
PCA-based	16.56±0.63	23.83	0.94	6.16
Heuristic	18.60±1.1	24.26	0.75	6.09
Rapantzikos et al.	18.04±0.3	20.14	0.58	**2.27**

Table 2
Confusion matrix for the PCA method ($p \approx 55\%$).

	SN	SW	BA	BO	SO	TE	TB
SN	40	0	5	0	15	0	0
SW	0	40	0	10	0	0	0
BA	0	0	75	20	0	0	5
BO	0	0	12	30	0	0	0
SO	10	0	5	0	60	5	0
TE	0	0	20	0	10	45	0
TB	0	0	10	0	0	0	40
Precision	0.800	1.000	0.591	0.500	0.706	0.900	1.000
Recall	0.667	0.800	0.789	0.714	0.750	0.600	0.800

Table 3
Confusion matrix for the proposed method ($p \approx 55\%$).

	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	36	6	0	0
SO	0	0	5	0	70	5	0
TE	0	0	5	0	0	70	0
TB	0	0	5	0	0	0	45
Precision	0.909	1.000	0.789	0.783	0.864	0.824	1.000
Recall	0.833	1.000	0.750	0.857	0.875	0.933	0.900

Table 4
Statistics on region-based classification.

	MCE (%)	μ_{err} (%)	$\tilde{\sigma}_{err}$ (%)	σ_{err} (%)
Proposed method	8.61±0.27	**10.88**	**0.43**	1.59
Itti et al.	17.51±0.69	18.63	0.74	1.27
ITTI-motion	17.07±0.48	18.44	0.56	**0.66**
PCA-based	**8.03±0.27**	16.85	0.78	7.31
Heuristic	13.65±1.35	18.46	0.78	4.37
Rapantzikos et al.	13.21±0.58	14.60	0.59	1.35

facts become evident when evaluating the saliency against the non-saliency-based methods: first, salient-based approaches seem to provide more consistent results when varying the selected variable (Fig. 8); second, the proposed method outperforms all other techniques since it allows feature extraction from areas bound to actual spatiotemporal saliency regions.

The second experiment illustrates the effect on classification performance when using features bound to more than one salient region, as explained in Section 3.3. This experiment corresponds to the flow diagram shown in Fig. 6b. Fig. 9 shows the obtained classification error versus the number of segmented regions. The two approaches based on Itti et al.'s approach perform equally well without fluctuations. Although, the average error is almost equal (Table 4), the average standard error for the static approach is higher. This is an expected result, since the ITTI-motion approach provides temporally more coherent regions. The heuristic method reaches a lower error after an almost sharp downturn (6 regions), but has high average standard error. The Rapantzikos et al. model outperforms both previous saliency-based approaches, while the proposed and the PCA-based techniques perform overall better than the rest. The PCA method has a slightly lower *MCE*, but has high fluctuations when varying the number of regions. The PCA method results in a sparser range of values compared to the denser values

of the saliency methods, and thus segmentation in a large number of regions (>6 as shown in Fig. 9) leads to not meaningful areas. This is the reason for the strong fluctuation in Fig. 9.

5. Conclusions

Saliency-based image and video processing contributes in several aspects to solving common computer vision problems. This work presents a computational model for saliency detection that exploits the spatiotemporal structure of a video stream and produces a per voxel saliency measure based on a feature competition approach. This measure provides evidence about important and non-important regions in the sequence. The benefits of the model obtained when using it as a pre-processing step in video classification are examined. The performance analysis is based on several experiments that illuminate different aspects of the method against other established techniques that either share a common notion of saliency or not. Two experiments show the improvement in classification error when selecting only part of the video stream based on saliency. Future work will focus on the application of the proposed model to computer vision problems that could benefit from the proposed spatio-temporal formulation like generic event detection and salient point detectors. Precisely, human action recognition based on salient points/regions is our current goal of research [52].

References

[1] A. Monnet, A. Mittal, N. Paragios, V. Ramesh, Background modelling and subtraction of dynamic scenes, IEEE Int. Conf. Comput. Vision (ICCV'03), vol. 2, 2003, pp. 1305–1313.

[2] A. Torralba, Contextual influences on saliency, in: L. Itti, G. Rees, J. Tsotsos (Eds.), Neurobiology of Attention, Academic Press/Elsevier, New York, 2005, pp. 586–593.

[3] A. Torralba, Contextual priming for object detection, Int. J. Comput. Vision 53 (2) (2003) 169–191.

[4] A.M. Treisman, G. Gelade, A feature integration theory of attention, Cognitive Psychol. 12 (1) (1980) 97–136.

[5] A.N. Tikhonov, V.Y. Arsenin, Solution of Ill-Posed Problems, W. H. Winston, Washington DC, 1977.

[6] B.A. Draper, A. Lionelle, Evaluation of selective attention under similarity transforms, International Workshop on attention and performance in computer vision (WAPCV'03), April 2003.

[7] C. Koch, S. Ullman, Shifts in selective visual attention: towards the underlying neural circuitry, Hum. Neurobiol. 4 (4) (1985) 219–227.

[8] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: Second Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.

[9] David J. Heeger, Optical flow using spatiotemporal filters, Int. J. Comput. Vision (1988) 279–302.

[10] Dirk Walther, Christof Koch, Modeling attention to salient proto-objects, Neural Networks 19 (2006) 1395–1407.

[11] E. Hering, Outlines of a Theory of the Light Sense, Harvard Univ. Press, Cambridge, Mass., 1964.

[12] E. Sahouria, A. Zakhor, Content analysis of video using principal components, IEEE Trans. Circuits Syst. Video Technol. 9 (8) (1999).

[13] E.R. Kandel, J.H. Schwartz, T.M. Jessell, Essentials of Neural Science and Behavior, Appleton & Lange, Stamford, Connecticut, 1995.

[14] F. Stentiford, An Attention Based Similarity Measure With Application to Content-Based Information Retrieval, SPIE Electronic Imaging, Santa Clara, 2003.

[15] F.-F. Li, R. VanRullen, C. Koch, P. Perona, Rapid natural scene categorization in the near absence of attention, Proc. Natl. Acad. Sci. 99 (2002) 8378–8383.

[16] F.H. Hamker, A dynamic model of how feature cues guide spatial attention, Vision Res. 44 (2004) 501–521.

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

[18] H. Schneiderman, T. Kanade, Probabilistic modeling of local appearance and spatial relationships for object recognition. IEEE Conf. Comput. Vision Pattern Recognition, July 1998, pp. 45–51.

[19] H.J. Zhang, A. Kankanhalli, S.W. Smoliar, Automatic partitioning of full-motion video, Multimedia Syst. 1 (1993) 10–28.

[20] I. Laptev, T. Lindeberg, Space-time interest points, in: Proc. ICCV'03, Nice, France, 2003, pp. 432–443.

[21] J.K. Tsotsos, S.M. Culhane, Y.K.W. Winky, L. Yuzhong, N. Davis, F. Nuflo, Modelling visual attention via selective tuning, Artif. Intell. 78 (1) (1995) 507–545.

[22] K. Rapantzikos, N. Tsapatsoulis, Enhancing the robustness of skin-based face detection schemes through a visual attention architecture, in: Proceedings of the IEEE International Conference on Image Processing (ICIP), September 2005.

[23] K. Rapantzikos, N. Tsapatsoulis, On the implementation of visual attention architectures, Tales of the Dissapearing Computer, Santorini, June 2003.

[24] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, spatiotemporal visual attention architecture for video analysis, in: Proceedings of IEEE International Workshop On Multimedia Signal Processing (MMSP'04), 2004.

[25] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, S. Kollias, A bottom-up spatiotemporal visual attention model for video analysis, IET Image Process. 1 (2) (2007) 237–248.

[26] K. Rapantzikos, Y. Avrithis, An enhanced spatiotemporal visual attention model for sports video analysis, International Workshop on Content-based Multimedia indexing (CBMI'05), June 2005.

[27] K. Rapantzikos, M. Zervakis, Robust optical flow estimation in MPEG sequences, IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), March 2005.

[28] K.G. Derpanis, J.M. Gryn, Three-Dimensional *n*th derivative of gaussian separable steerable filters, ICIP'05, vol. 3, 2005, pp. 553–556.

[29] L. Itti, Automatic foveation for video compression using a neurobiological model of visual attention, IEEE Trans. Image Process. 13 (10) (2004) 1304–1318.

[30] L. Itti, C. Koch, A saliency-based search mechanism for overt and covert shifts of visual attention, Vision Res. 40 (2000) 1489–1506.

[31] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (11) (1998) 1254–1259.

[32] L. Itti, P. Baldi, A Principled Approach to Detecting Surprising Events in Video, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2005, pp. 631–637.

[33] L. Zelnik-Manor, M. Irani, Statistical analysis of dynamic actions, IEEE Trans. Patt. Anal. Mach. Intell. 28 (9) (2006) 1530–1535.

[34] M. Riedmiller, Advanced supervised learning in multi-layer perceptrons-from backpropagation to adaptive learning algorithms, Int. J. Comput. Stand Interfaces, Special Issue Neural Networks 16 (1994) 265–278.

[35] M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex, Nat. Neurosci. 2 (11) (1999) 1019–1025.

[36] M. Ristivojević, J. Konrad, Space-time image sequence analysis: object tunnels and occlusion volumes, IEEE Trans. Image Process. 15 (2006) 364–376.

[37] M.J. Black, P. Anandan, The robust estimation of multiple motions: parametric and piecewise-smooth flow fields, CVIU 63 (1) (1996) 75–104.

[38] N.M. Oliver, B. Rosario, A.P. Pentland, A Bayesian computer vision system for modelling human interactions, IEEE Trans. Pattern Anal. Mach. Intell. 22 (8) (2000).

[39] O. Boiman, M. Irani, Detecting irregularities in images and in video, IEEE International Conference on Computer Vision (ICCV), October 2005.

[40] Otsu Nobuyuki, A threshold selection method from gray level histograms, IEEE Trans. Syst., Man Cybern. SMC-9 (1) (1979).

[41] P. Groves, P. Bajcsy, Methodology for hyperspectral band and classification model selection, IEEE Workshop on Advances in Techniques for Analysis of Remotely sensed data, October 2003, pp. 120–128.

[42] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, New York, 2001.

[43] S. Corchs, G. Ciocca, R. Schettini, Video summarization using a neurodynamical model of visual attention, Multimedia Signal Processing (MMSP'04), October 2004, pp. 71–74.

[44] T. Avraham, M. Lindenbaum, Attention-based dynamic visual search using inner-scene similarity: algorithms and bounds, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 251–264.
[45] T.J. Williams, B.A. Draper, An Evaluation of Motion in Artificial Selective Attention, IEEE Conf. Comput. Vision Pattern Recognition (CVPRW'05), June 2005, pp. 85.
[46] U. Rutishauer, D. Walther, C. koch, P. Perona, Is bottom-up attention useful for object recognition?, CVPR'04, July 2004, pp. 37–44.
[47] V. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[48] W.T. Freeman, E.H. Adelson, The design and use of steerable filters, IEEE Trans. Pattern Anal. Mach. Intell. 13 (9) (1991) 891–906.
[49] Y. Ke, R. Sukthankar, M. Hebert, Efficient visual event detection using volumetric features, Int. Conf. Comput. Vision, vol. 1, October 2005, pp. 166–17.
[50] Y.-F. Ma, L. Lu, H.-J. Zhang, M. Li, A user attention model for video summarization," ACM Multimedia Conf., 2002, pp.533–542.
[51] Yu-Fei Ma, Xian-Sheng Hua, Lie Lu, Hong-Jiang Zhang, A generic framework of user attention model and its application in video summarization, IEEE Trans. Multimedia 7 (5) (2005) 907–919.
[52] K. Rapantzikos, Y. Avrithis, S. Kollias, Dense saliency-based spatiotemporal feature points for action recognition, IEEE Conf. Comput. Vision Pattern Recognition, 2009, accepted.