Cogn Comput (2011) 3:167–184
DOI 10.1007/s12559-011-9097-0

Spatiotemporal Features for Action Recognition and Salient Event Detection

Konstantinos Rapantzikos · Yannis Avrithis · Stefanos Kollias

Received: 30 April 2010 / Accepted: 10 February 2011 / Published online: 11 March 2011

Abstract Although the mechanisms of human visual understanding remain partially unclear, computational models inspired by existing knowledge on human vision have emerged and applied to several fields. In this paper, we propose a novel method to compute visual saliency from video sequences by counting in the actual spatiotemporal nature of the video. The visual input is represented by a volume in space–time and decomposed into a set of feature volumes in multiple resolutions. Feature competition is used to produce a saliency distribution of the input implemented by constrained minimization. The proposed constraints are inspired by and associated with the Gestalt laws. There are a number of contributions in this approach, namely extending existing visual feature models to a volumetric representation, allowing competition across features, scales and voxels, and formulating constraints in accordance with perceptual principles. The resulting saliency volume is used to detect prominent spatiotemporal regions and consequently applied to action recognition and perceptually salient event detection in video sequences. Comparisons against established methods on public datasets are given and reveal the potential of the proposed model. The experiments include three action recognition scenarios and salient temporal segment detection in a movie database annotated by humans.

Keywords Spatiotemporal visual saliency · Volumetric representation · Action recognition · Salient event detection

K. Rapantzikos (✉) · Y. Avrithis · S. Kollias
National Technical University of Athens, Athens, Greece
e-mail: rap@image.ntua.gr

S. Kollias
e-mail: stefanos@cs.ntua.gr

Introduction

Current evidence supports the fact that human visual understanding is initiated by focusing attention to some parts of the input. Such parts are usually representative enough so that attending them reduces complexity and enhances understanding [1–3]. Although the biological mechanisms involved in this process are not clear yet, computational models, which are either based on strict or relaxed correspondences to well-founded neurophysiological counterparts, have been developed and applied to several computer vision fields.

In this paper, we propose a computational framework for spatiotemporal saliency detection and exploit it to tackle the problems of representing, detecting and recognizing salient events in video sequences. The nature of the model makes it more relevant to the work by Itti et al. [4] and the early work by Milanese et al. [5] The volumetric extensions of Itti's model we have proposed in the past proved to be quite efficient in a series of video analysis applications [6–8], and thus we follow a similar video representation. Milanese et al. used a relaxation process to optimize an energy measure consisting of four terms that model different interactions between pixels to detect salient regions in images. Our method is also based on optimization, though in this case competition is performed across features, scales, and voxels in our volumetric representation. The *competition* is implemented through constrained minimization with the constraints being inspired by the Gestalt laws. From the mathematical point of view, the computational model we use is the same with the one we proposed in [9]. In this manuscript, we provide a detailed walkthrough of the model, highlight the main principles, and give a better insight of the method. We augment the action recognition results with standard datasets in the field

and discuss them in detail. Furthermore, in this work, we test our detector on salient event detection using a human annotated movie database, which tends to become a standard in the field.

The Gestalt theory, we were inspired from, refers to a unified configuration of objects/events that has specific attributes, which are greater than the simple sum of its individual parts that share common properties [10, 11]. For example, we automatically perceive a walking event as a complete human activity rather than a set of individual parts like moving legs, swinging arms, etc. Each subaction is clearly an individual unit, but the greater meaning depends on the arrangement of the subactions into a specific configuration (walking action). Specifically, the *constraints* are related to the *figure/ground* (identify objects as distinct from their background), *proximity* (group elements that exhibit spatial or temporal proximity), *closure* (visually close gaps in a form), *similarity* (group similar elements depending on form, color, size or brightness), and *common fate* (elements with similar movements are perceived as a whole) Gestalt laws. The output of our model is a spatiotemporal distribution with values related to the saliency of the individual visual units.

Using the proposed model, our first objective is to identify salient configurations of small visual units that form a coherent surface in space–time possibly belonging to a single event. The second objective is to represent and recognize this event or measure its saliency against other salient configurations in the same or different video clips. Experiments, using public datasets and comparison against established methods, illustrate the potential of the method. Action detection is based on detecting salient spatiotemporal points on the derived saliency distribution and implemented through a bag-of-words approach. Statistics on three diverse datasets demonstrate the performance of the method. Generic salient event detection is demonstrated by detecting salient parts of movie clips and evaluated using a recently released database [MUSCLE]. The database is composed of a set of movie clips with accompanying human annotations of saliency. We evaluate the proposed saliency computation method, using these annotations as ground truth, examine the potential of our method to produce perceptually meaningful results and present comparisons.

Related Work

One of the most widely used computational models of *spatial visual attention*, namely the one by Itti et al. [4], was based on the neurally plausible computational architecture for controlling visual attention proposed by Koch and Ullman [3]. The core idea is the existence of a saliency map that combines information from several feature maps into a global measure, where points corresponding to one location in each feature map project to single units in the saliency map. Attention bias is then reduced to drawing attention toward high-activity locations. The Guided Search model of Wolfe [12], also inspired by the previous work, hypothesizes that attentional selection is the net effect of feature maps weighted by the task at hand and of bottom-up activation based on local feature differences. Hence, local activity of the saliency map is based both on feature contrast and top-down feature weight. Tsotsos et al. proposed a selective tuning model that applies attentional selection recursively during a top-down process. First, the strongest item is selected at the top level of the hierarchy using a Winner-Take-All (WTA) process (the equivalent of a saliency map), and then, the hierarchy of a WTA process is activated to detect and localize the strongest item in each layer of representation [13].

Although not always explicitly said, extraction of *points of interest* is usually based on a saliency-related representation of the input. For example, the famous point detector by Harris et al. detects points as local maxima of a distribution that measures the image intensity change in multiple directions [14]. Similarly, the Hessian-affine detectors involving computation of second derivatives give strong (salient) response on blobs and ridges [15]. The detector of Kadir and Brady is explicitly based on the notion of saliency and aims at detecting representative, discriminative, and therefore salient regions in the image [16]. A review and a well-grounded comparison of spatial detectors is given by Mikolajcyk et al. [17].

Recently, object/event detection and recognition are often based on points of interest combined with *bag-of-words* approaches according to which the visual input can be well represented by a set of small visual regions (words) extracted around salient points like the ones referred before. Csurka et al. [18] use visual words around regions extracted with the Harris affine detector [19] to represent and detect visual objects [20]. Their approach demonstrates robustness to background clutter and good classification results. Wang et al. proposed a similar recognition framework and consider the problem of describing the action being performed by human figures in still images [21]. Their approach exploits edge points around the shape of the figures to match pairs of images depicting people in similar body poses. Bosch et al. [22] also propose a bag-of-words-based method to detect multiple object categories in images. Their method learns categories and their distributions in unlabeled training images using probabilistic Latent Semantic Analysis (pLSA) and then, uses their distribution in test images as a feature vector in a supervised k-NN scheme.

Spatiotemporal event representation and *action recognition* have recently attracted the interest of researchers in the field. One of the few interest point detectors of a spatiotemporal nature is an extension of the Harris corner detection to 3D, which has been studied quite extensively by Laptev and Lindeberg [23] and further developed and used in [24, 25]. A spatiotemporal corner is defined as an image region containing a spatial corner whose velocity vector is changing direction. They proposed also a set of image descriptors for representing local space–time image structures as well as a method for matching and recognizing events and activities based on local space–time interest points [23]. Dollár et al. [26] proposed a framework, which is based on a bag-of-words approach, where a visual word is meant as a cuboid. They developed an extension of a periodic point detector to the spatiotemporal case and test the performance on action recognition applications. They show how the use of cuboid prototypes, extracted from a spatiotemporal video representation, gives rise to an efficient and robust event descriptor by providing statistical results on diverse datasets. Niebles et al. [27] use the same periodic detector and propose an unsupervised learning method for human action categories. They represent a video sequence by a collection of spatiotemporal words based on the extracted interest points and learn the probability distribution of the words using pLSA.

Oikonomopoulos et al. [28] use a different measure and propose a spatiotemporal extension of the salient point detector by Kadir and Brady. They relate the entropy of space–time regions to saliency and describe a framework to detect points of interest at their characteristic scale determined by maximizing their entropy. This detector is evaluated on a dataset of aerobic actions, and promising results are reported. Wong and Cipolla [29] report a more thorough evaluation of the latter and propose their own detector based on global information. Their detector is evaluated against the state-of-the-art in action recognition and outperforms the ones proposed by Laptev et al., Dollár et al., and Oikonomopoulos et al. on standard datasets highlighting the importance of global information in space–time interest point detection. Willems et al. propose a space–time detector based on the determinant of the 3D Hessian matrix, which is computationally efficient (use of integral videos) and is still on a par with current methods. Quite recently, Rapantzikos et al. [9] proposed and evaluated the potential of spatiotemporal saliency to represent and detect actions in video sequences. Their method ranks among the best on standard human action datasets.

Action recognition using a different framework is presented in the work by Blank et al. [30], who consider human actions as three-dimensional shapes forming a *spatiotemporal volume* modeled by a Poisson equation. Their method utilizes properties of the solution to the Poisson equation to extract space–time features, but requires successful foreground to background segmentation. Schlectman et al. [31, 32] tackle the same problem in a different way: They measure similarity between visual entities (images or videos) based on matching internal self-similarities. These similarities are measured densely throughout the image/video at multiple scales. Such correlation-based methods overcome the assumption of the foreground to background segmentation, since they search the entire space for the query event, but they are not easy to scale-up due to the computational burden. Ke et al. propose an event detector by learning a cascade of filters for each action of interest and scan the video sequences in space and time to detect the event [33]. The authors claim that the proposed detector recognizes actions that are traditionally problematic for interest point methods, such as smooth motions where insufficient space–time interest points are available and prove this using a variety of examples. The statistics are not high, compared to the state-of-the-art, but it seems that their detector is more generic than the counterpart.

Another class of methods that is related to our work is *perceptually salient event detection* in video sequences. We use visual saliency to produce an attention curve that summarizes saliency across frames and thus allows decisions about the importance of events occurring in the stream. Such a process shares certain similarities with two fields: (a) unusual/surprising event detection methods that compute a saliency measure associated with the visual stream and explore the evolution of this measure to detect events [34–36], (b) work in movie summarization using attention curves [37, 38]. We should differentiate between two definitions of unusual activities: (a) activities dissimilar to regular ones and (b) rare activities, with low similarity among other usual ones.

Most attempts that are built around the first definition tackle the problem by pre-defining a particular set of activities as being usual, model it in some way, and then detect whether an observed activity is anomalous [39, 40]. Researchers working on rare event detection assume that unusual events are sparse, difficult to describe, hard to predict and can be subtle, but given a large number of observations it is easier to verify if they are indeed unusual. Zhong et al. divide the video into equal length segments and classify the extracted features into prototypes [36]. Measures on a prototype-segment co-occurrence matrix reveal unusual formations that correspond to rare events. Using the similar notion of unusual activity, but under a quite different framework, Adam et al. automatically analyze the video stream from multiple cameras and detect unusual events by measuring the likelihood of the observation with respect to the probability distribution of the observations stored in the buffer of each camera [41].

Going one step further toward human perception of saliency, Ma et al. [38] propose a framework for detecting the salient parts of a video based on user attention models. They use motion, face, and camera attention along with audio attention models (audio saliency and speech/music) as cues to capture salient information and identify the audio and video segments to compose the summary. In a similar fashion, Evangelopoulos et al. [42, 43] and Rapantzikos et al. [44] fuse audio, visual, and text saliency measures into a single attention curve and select prominent parts of this measure to generate a summary.

We formulate the problem in the section "Problem Formulation" and discuss volumetric representation of features and conspicuity in the next section. In "Volumetric Saliency Competition" section, we provide details and visual examples of the framework for saliency computation, while a description of datasets and a detailed description of the experimental methodology and results is given in "Applications and Experiments". The final section concludes our work.

Problem Formulation

We define saliency computation in image sequences as a problem of assigning a measure of interest to each visual unit. This means that a saliency measure is produced by taking into account the actual spatiotemporal evolution of the input. Inspired by theories of grouping and perceptual organization, we propose a model based on a volumetric representation of the visual input where features are grouped together according to several criteria related to the Gestalt laws.

Figure 1 illustrates the proposed spatiotemporal saliency architecture. The input is composed of a stack of consequent frames represented in the model as a volume in space–time. This volume is decomposed into a set of perceptual features, and spatiotemporal pyramids are formed. Each pyramid represents the spatiotemporal evolution of the specific feature for the whole sequence at different scales (resolutions). In order to have a common conspicuity encoding among all features and enable a fair competition, we normalize each pyramid by a conspicuity operator based on voxel proximity and contrast. For ease of visualization, only the conspicuity pyramids are shown in Fig. 1 as gray volumetric shapes. The arrows related to the illustrated pyramids are indicative of the three way voxel competition allowed in the model: (a) Intra-feature competition between voxels of the same feature and same scale, (b) inter-scale competition between voxels of the same feature but different scale, and (c) inter-feature competition between voxels of different features. Specifically, the green arrows indicate competition at the intra-feature level, the blue ones at the inter-scale level, and the red ones at the inter-feature level. This competition is realized in the model as a constrained energy minimization problem that leads to a final saliency volume. Broadly speaking, the constraints enhance coherency of similar and neighboring voxels according to each of the criteria. Figure 1 shows also indicative input and output examples of the model. The video clip used in the illustration depicts an ice-skater sliding on the ice while the camera is tracking him. Actually, the output images are slices of the saliency spatiotemporal volume and encode salient areas as the ones with high values (red) and non-salient as the ones with low values (blue).

Precisely, let V be a video volume created by a set of consequent frames defined on a set of points Q with $q = (x, y, t)$ being an individual space–time point. Points $q \in Q$ form a grid in the discrete Euclidean 3D space defined by their Cartesian coordinates. Under this representation, point q becomes the equivalent to a voxel in this volume. Voxels in V related to a moving object are perceived as occupying a spatiotemporal region in the volume. The value of volume V at point q will be denoted as $V(q)$.

Volume V is decomposed into a set of feature volumes F_i with $i = 1, 2, \ldots, M$. The entire competition process can

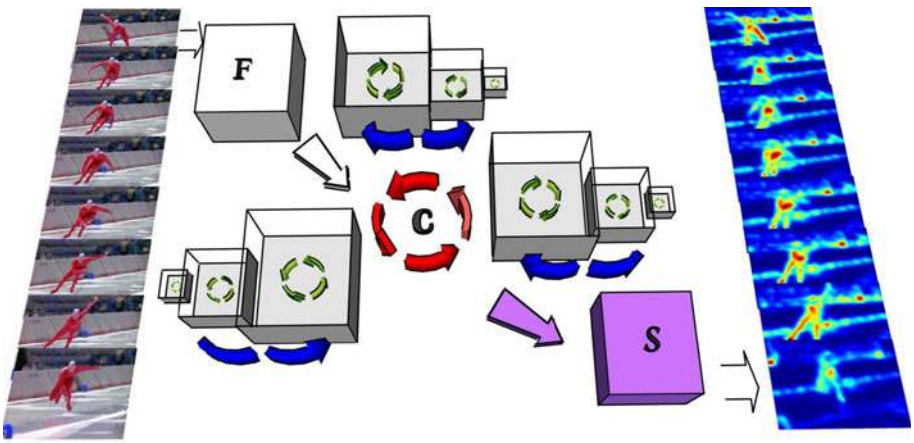

Fig. 1 Proposed architecture. The input space–time volume F is formed by stacking consequent frames and is decomposed into a set of multi-scale conspicuity features C. The arrows correspond to inter-, intra-feature and inter-scale competition that produces the saliency volume S (better viewed in *color*) (Color figure online)

be generalized for an arbitrary number of features. However, in this work, we focus on three specific features ($M = 3$), namely intensity (F_1), color (F_2), and spatiotemporal orientation (F_3). Intensity and color features are based on color opponent theory, and spatiotemporal orientations are related to moving stimuli. Each feature volume is further decomposed to create a volumetric pyramid, and a set $\mathbf{F} = \{F_{i,\ell}\}$ is created with $i = 1, 2, 3$ and $\ell = 0, 1, \ldots, L$. This decomposition enables a fine-to-coarse representation of maximum scale L, where the conventional Gaussian image pyramid is generalized to a space–time pyramid [6, 45]. In order to establish a common conspicuity encoding, a conspicuity operator is applied to all feature volumes. This operator marks a voxel as more salient when its local feature value differs from the average feature value in the surrounding region. A set of conspicuity volumes $\mathbf{M} = \{M_{i,\ell}\}$ is then formed.

The proposed model assigns to every voxel q a saliency value derived through interaction of the conspicuity volumes $\mathbf{M}$. We formulate this procedure as an optimization of an energy function E composed of a data term E_d and a smoothness one E_s:

$$E(\mathbf{M}) = \lambda_d \cdot E_d(\mathbf{M}) + \lambda_s \cdot E_s(\mathbf{M}) \quad (1)$$

In a regularization framework, the first term of Eq. 1 is regarded as the observation term and the second as the smoothness one, since it regularizes the current estimate according to the selected constraints [46, 47]. The data term models the interaction between the observation and the predicted value, while the smoothness term is composed of the three similarity constraints. The final saliency volume S is obtained by averaging the final conspicuity volumes after the minimization of E. In order to get an intuition of a saliency volume, Fig. 2 shows an indicative example of saliency computation for the skater sequence. Figure 2a shows a set of frames, Fig. 2b and c depict semi-transparent versions of the computed saliency volume, while Fig. 2d depicts an ISO surface that contains the most salient parts of the scene.

Feature and Conspicuity Formulation

Features

The initial volume is decomposed into a set of feature volumes, namely intensity, color, and spatiotemporal orientation, that represent a specific property of the input video. 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 [48]. The extent to which these opponent channels attract attention of humans has been previously investigated in detail, both for biological [2] and computational models of attention [4, 37]. According to the opponent color scheme, if r, g, b are the color volumes (color components of the video volume V), the luminance and color opponent volumes are obtained by

$$F_1 = \frac{1}{3} \cdot (r + g + b) \quad (2)$$

$$F_2 = RG + BY \quad (3)$$

where $RG = R - G$, $BY = B - Y$ and $R = r - (g + b)/2$, $G = g - (r + b)/2$, $B = b - (r + g)/2$, $Y = (r + g)/2 - |r - g|/2 - b$.

The orientation feature volume is computed using spatiotemporal steerable filters tuned to respond to moving stimuli. This method of filtering is similar to the quadrature sine and cosine pair of spatiotemporal energy filters developed by Freeman et al. [49]. The desired filtering is implemented by convolving the intensity volumes $F_{1,\ell}$ with the second derivatives G_2 of a 3D Gaussian filter and their Hilbert transforms H_2 and then, taking the quadrature of the response to eliminate phase variation:

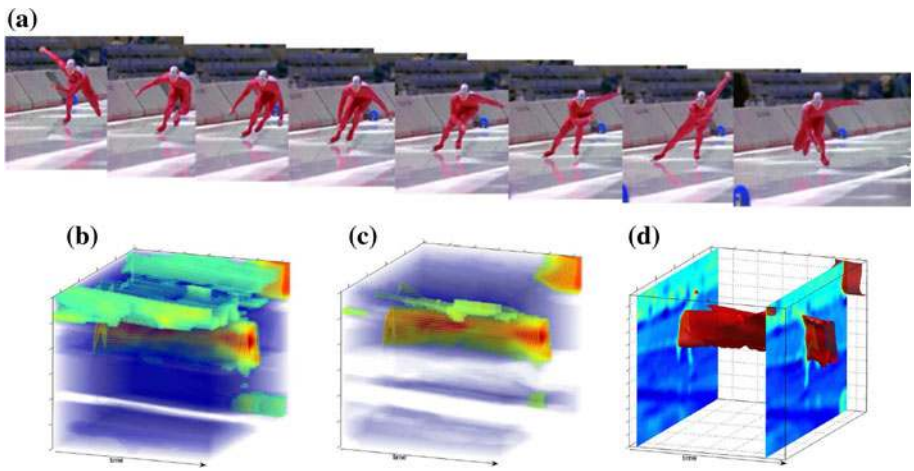

Fig. 2 **a** Consequent frames of the scater sequence illustrated also in Fig. 1, **b**, **c** semi-transparent versions of the saliency volume (with increasing transparency), **d** ISO surface at a specific threshold showing its most salient part of the saliency volume (better viewed in *color*) (Color figure online)

$$E_v^\theta = [G_2^\theta * F_{1,\ell}]^2 + [H_2^\theta * F_{1,\ell}]^2 \quad (4)$$

where $\theta \equiv (\alpha, \beta, \gamma)$ and α, β, γ are the direction cosines according to which the orientation of the 3D filter kernel is steered. The filters used in our model are proposed by Derpanis et al. [50]. In order to limit the orientations to the most important ones, we follow the rationale by Wildes et al. [51] and compute energies at directions θ related to rightward $(\frac{1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})$, leftward $(\frac{-1}{\sqrt{2}}, 0, \frac{1}{\sqrt{2}})$, upward $(0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and downward $(0, \frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ motion. In order to get a purer measure of orientation, the response of each filter is normalized by the sum and the final spatiotemporal orientation volume is computed by

$$F_3(q) = \frac{\sum_\theta E_v^\theta(q)}{\sum_r \sum_\theta E_v^\theta(r)} \quad (5)$$

All features are then decomposed into volumetric pyramids that represent the feature volumes in multi-scale in order to create the set $\mathbf{F} = \{F_{i,\ell}\}$, where $i = 1, 2, 3$ and $\ell = 0, 1, \ldots, L$. The volume in each scale of the pyramid is a 3D-smoothed and subsampled version of the input volume. The required smoothing and subsampling operations are obtained by a spatiotemporal Gaussian low-pass filter and decimation in all three dimensions by consecutive powers of two. The final result is a hierarchy of video volumes that represent the input sequence in decreasing spatiotemporal scales (resolutions). Each volume simultaneously represents the spatial distribution and temporal evolution of the encoded feature.

Conspicuity

Inspired by the figure/ground law of Gestalt theory that is related to foreground/background separation of regions, we apply a local contrast enhancement operator in order to enhance voxels that stand-out from their local surrounding (high valued among low valued and vice-versa). This operator ensures also a common importance/conspicuity range among all features so that they are comparable and thus correctly applied as initial solutions for the minimization procedure that follows. This means that e.g. the most salient voxel in the intensity volume must have the same value as the most salient voxel in the color volume. Ideally, each voxel should be compared against the whole feature volume in order to test its conspicuity against the whole sequence. Nevertheless, there are two reasons for not doing this: (a) The sequence may not be composed of a single shot and therefore the comparison would be non-homogenous and (b) the computational barrier. Conspicuity is calculated as

$$C_{i,l}(q) = \left| F_{i,l}(q) - \frac{1}{|N_q|} \sum_{r \in N_q} F_{i,l}(r) \right| \quad (6)$$

where N_q is the set of the 26-neighbors of the voxel q in the 3D space. The 26-neighborhood is the direct extension in 3D of the standard 8-neighborhood in the 2D image space.

Volumetric Saliency Competition

Energy Formulation

Each of the conspicuity volumes encodes the saliency of the contained voxels according to the corresponding feature only. These volumes should interact in order to produce a single saliency measure for each voxel. The proposed model achieves this through a regularization framework, whereby conspicuity volumes compete along a number of directions, namely interaction among voxels at the intra-feature, inter-scale and inter-feature level. As discussed in the introduction, the different interactions are implemented as a competition modeled by energies inspired by the Gestalt laws. Specifically, proximity and closure laws give rise to the intra-feature constraint, according to which voxels that are located near each other tend to be part of a group and small gaps are closed due to the induced forces. The similarity law is related to all energies, since voxels similar in terms of intra-, inter-feature, and inter-scale value tend to group. Finally, the common fate law is related to the whole minimization approach that produces space–time regions that can be perceived as coherent and homogenous. In the following, a detailed analysis of the regularization framework and the involved data and smoothness terms in Eq. 1 is given along with illustrative examples of their use and performance in "Energy Minimization".

The data term, E_d, preserves a relation between the observed and initial estimate in order to avoid excessive smoothness of the result, since the energies involved in the smoothness term E_s tend to smooth the visual input according to different criteria. This constraint is formulated as an energy relating the observed to the initial voxel values. The data term for a set of conspicuity volumes $\mathbf{C}$ is defined as

$$E_d(\mathbf{C}) = \sum_{i,\ell,q} (C_{i,\ell}(q) - C_{i,\ell}^0(q))^2 \quad (7)$$

where $C_{i,\ell}^0(q)$ is the initial estimate, $i = 1, 2, 3$, $l = 1, 2, \ldots, L$, $q \in Q$. The sum limits are omitted in the following for simplicity.

The smoothness term is formulated as

$$E_s(\mathbf{M}) = E_1(\mathbf{M}) + E_2(\mathbf{M}) + E_3(\mathbf{M}) \quad (8)$$

where E_1, E_2, E_3 denote the constraints for the intra-feature, inter-feature, and inter-scale competition respectively. The *intra-feature energy constraint* E_1 is

related to intra-feature coherency, i.e. defines the interaction among neighboring voxels of the same feature at the same scale. It enhances voxels that are incoherent with their neighborhood in terms of feature value and is minimized when the value of the examined voxel is equal to the average value of its 26-neighborhood N_q:

$$E_1(\mathbf{C}) = \sum_{i,\ell,q} \left(C_{i,\ell}(q) - \frac{1}{|N_q|} \sum_{r \in N_q} C_{i,\ell}(r) \right)^2 \quad (9)$$

Ideally, this constraint will close small gaps in the feature volume and allow similar—in terms of value—voxels to form larger coherent regions.

The *inter-feature energy constraint* is related to inter-feature coherency, i.e. it enables competition for saliency among different features so that voxels being conspicuous across all feature volumes are grouped together and form coherent regions. It involves competition between a voxel in one feature volume and the corresponding voxels in all other feature volumes:

$$E_2(\mathbf{C}) = \sum_{i,\ell,q} \left(C_{i,\ell}(q) - \frac{1}{M-1} \sum_{j \neq i} F_{j,\ell}(q) \right)^2 \quad (10)$$

This constraint introduces another force that biases the examined voxels toward the average of their competitors.

The *inter-scale energy constraint* operates on the pyramid scales of each individual feature and is related to coherency among ever coarser resolutions of the input, i.e. aims to enhance voxels that are conspicuous across different pyramid scales. This means that if a voxel retains a high value along all scales, then it should become more salient.

$$E_3(\mathbf{C}) = \sum_{i,\ell,q} \left(C_{i,\ell}(q) - \frac{1}{L-1} \sum_{n \neq l} C_{i,n}(q) \right)^2 \quad (11)$$

Figures 3, 4, 5, and 6 illustrate the operation of the proposed scheme both when the intra-feature, inter-feature, inter-scale, and all constraints are involved, respectively. Figures 3a, 4a, 5a, and 6a show the enabled interactions between the multi-scale conspicuity volumes involved in the minimization process. The horizontal axis corresponds to features and the vertical one to scales. The grayed-out volumes are the ones not involved in the minimization process when using the specified constraint.

Energy Minimization

Minimization of Eq. 1 can be achieved using any descent method. We adopt a steepest gradient descent algorithm where the value of each feature voxel is updated along a search direction, driving the value in the direction of the estimated energy minimum

$$C^{\tau}_{i,\ell}(q) = C^{\tau-1}_{i,\ell}(q) + \delta C^{\tau-1}_{i,\ell}(q) \quad (12)$$

with

$$\delta C^{\tau-1}_{i,\ell}(q) = -\gamma \cdot \frac{\partial E(\mathbf{C}^{\tau-1})}{\partial C^{\tau-1}_{i,\ell}(q)} + \mu \cdot \delta C^{\tau-1}_{i,\ell}(q) \quad (13)$$

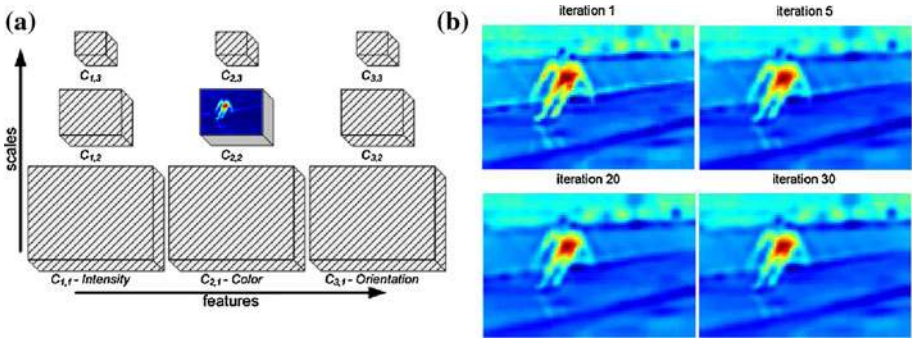

Fig. 3 **a** Conspicuity volumes involved in minimization using intra-feature constraint only, **b** result for $M_{2,2}$ after 1, 5, 20 and 30 iterations

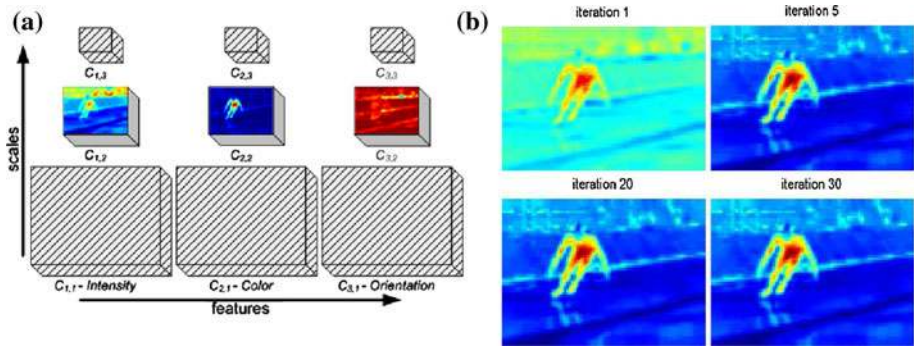

Fig. 4 **a** Conspicuity volumes involved in minimization using inter-feat constraint only, **b** result for $M_{2,2}$ after 1, 5, 20 and 30 iterations

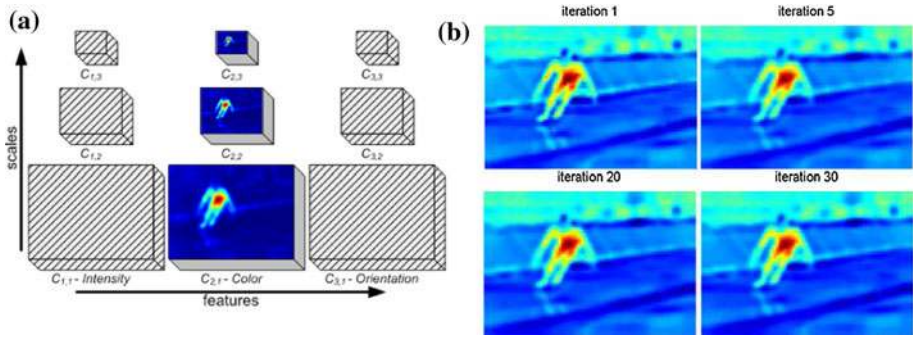

Fig. 5 **a** Conspicuity volumes involved in minimization using inter-scale constraint only, **b** result for $M_{2,2}$ after 1, 5, 20 and 30 iterations

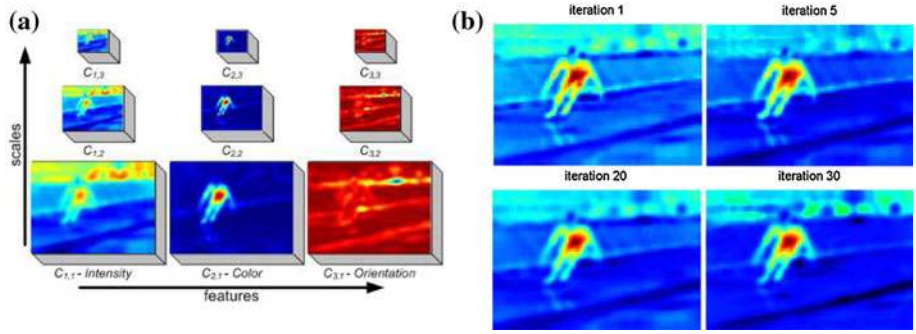

Fig. 6 **a** Conspicuity volumes involved in minimization using all constraints, **b** result for $M_{2,2}$ after 1, 5, 20 and 30 iterations

where τ is the iteration number, γ is the learning rate, and μ a momentum term that controls the algorithm's stability [52]. These two parameters are important both for stability and speed of convergence. Practically, a few iterations are enough for the estimate to reach a near optimal solution. In order to keep notations simple, we omit the iteration symbol τ in the following.

Minimization of Eq. 1 requires the computation of the energy partial derivative involved in Eq. 13

$$\begin{aligned}\frac{\partial E()}{\partial C_{k,m}(s)} &= \lambda_d \cdot \frac{\partial E_d()}{\partial C_{k,m}(s)} + \lambda_d \cdot \frac{\partial E_s(\mathbf{C})}{\partial C_{k,m}(s)} \\ &= \lambda_d \cdot \frac{\partial E_d(\mathbf{C})}{\partial C_{k,m}(s)} + \lambda_s \cdot \sum_{c=1}^{3} \frac{\partial E_c(\mathbf{C})}{\partial C_{k,m}(s)}\end{aligned} \tag{14}$$

where $k = 1, \ldots, M$, $s \in Q$ and E_c with $c = 1, 2, 3$ the three energy constraints of the smoothness term.

Computation of Derivatives

Detailed computations of the derivatives appearing in Eq. 14 are provided in "Appendix". In particular, the partial derivative of E_d is computed as

$$\frac{\partial E_d}{\partial C_{k,m}(s)} = 2 \cdot \sum_{q} \left(C_{k,m}(s) - C^0_{k,m}(s) \right) \tag{15}$$

The partial derivative of the intra-feat constraint in Eq. 9 becomes

$$\begin{aligned}\frac{\partial E_1}{\partial C_{k,m}(s)} = 2 \cdot \Bigg[& C_{k,m}(s) - \frac{1}{|N_q|^2} \\ & \cdot \sum_{q \in N(s)} \left(2N \cdot C_{k,m}(q) - \sum_{r \in N_q} C_{i,\ell}(r) \right) \Bigg]\end{aligned} \tag{16}$$

Figures 3b, 4b, 5b, and 6b show indicative results for the conspicuity feature $M_{2,2}$ (color feature at scale 2) after 1, 5, 20, and 30 iterations. Red values correspond to the most conspicuous voxels and the blue to the least conspicuous. Results are shown for the video in Fig. 1. As mentioned earlier, the clip depicts an ice-scater with a moving camera tracking him; a moving camera clip is intentionally chosen in order to illustrate the robustness of the proposed approach to background motion.

Specifically, as illustrated in Fig. 3b, voxels of the same feature at each individual scale are involved in the minimization process when only the intra-feature constraint is used. No interactions across scales or across features are enabled. Notice how the voxels corresponding to the torso of the ice-skater are grouped together during the evolution of the minimization process. The torso corresponds to the most coherent and salient—in terms of intensity, color, and orientation region across frames.

The derivative of the inter-feature constraint is computed as

$$\frac{\partial E_2}{\partial C_{k,m}(s)} = 2 \cdot \frac{M}{M-1} \cdot \left(C_{k,m}(s) - \frac{1}{M-1} \cdot \sum_{j \neq k} C_{j,m}(s) \right) \tag{17}$$

As shown in Fig. 4, voxels of different features at the same scale are allowed to compete for saliency when only the inter-feature constraint is used. Notice that the motion feature is quite uninformative since the relative motion between the camera and the ice-skater is small. Nevertheless, voxels belonging to the object of interest (scater) become salient due to the across-feature competition, since background voxels are not conspicuous according to the rest of the features.

Finally, the derivative of the inter-scale constraint becomes:

$$\frac{\partial E_3}{\partial C_{k,m}(s)} = 2 \cdot \frac{L}{L-1} \cdot \left[C_{k,m}(s) - \frac{1}{L-1} \cdot \sum_{n \neq \ell} C_{k,n}(s) \right] \tag{18}$$

As shown in Fig. 5, voxels of the same feature across all scales are allowed to compete for saliency when only the inter-scale constraint is used. Notice how the voxels corresponding to the "skeleton" of the ice-skater become more evident across iterations. The legs, arms, and torso form the most coherent regions across scales. Figure 6 shows the final result of the proposed model when all constraints are enabled. The background regions become ever less salient across iterations, while voxels belonging to the object of interest are grouped and enhanced.

Saliency Computation

After the minimization process has reached a desired error value δC, the output is a set of modified conspicuity multi-scale volumes $\hat{\mathbf{C}} = \{\hat{C}_{i,\ell}\}$ that should be fused to form the final saliency volume. Saliency $\mathbf{S} = \{ S_\ell \}$, where S_ℓ for level ℓ is

$$S_\ell = \frac{1}{M} \cdot \sum_{i=1}^{M} \hat{C}_{i,\ell}, \tag{19}$$

is computed as the average of all volumes.

Slices of the final saliency volumes are shown in Figs. 6, 7 and throughout "Applications and Experiments". Figure 7 shows the effect of the proposed method when applied to a moving stimuli illusion, where a running horse is illustrated under the heavy presence of random valued dots.[1] The perceptibility of the horse is possible mainly due to the coherency across small spatiotemporal neighborhoods rather than to its spatial shape (it is hard to detect the horse when looking at individual frames only). Figure 7a shows one representative frame, while the rest of the images show the evolution of the proposed method. The horse has become evident after 25 iterations (Fig. 7d) due to the enhancement of the aforementioned coherency.

Applications and Experiments

We evaluate the proposed model by setting up experiments in two different domains: action recognition using a variety of datasets and perceptually salient event detection using a human annotated movie database. The aim is twofold: First to experimentally validate that the proposed model does enhance representative parts of the visual input and second to examine the relation of the model to perceptual visual understanding, as this is established through comparison against human evaluation.

Action Recognition

Datasets and Algorithms for Comparison

We used three datasets to evaluate our method in an action recognition task. The first is the KTH dataset [53], one of the largest of its kind available online,[2] which consists of six types of human actions (walking, jogging, running, boxing, hand-waving and hand-clapping) performed by 25 subjects in four different scenarios: outdoors (s_1), outdoors with scale variation (s_2), outdoors with different clothes (s_3), and indoors (s_4), giving rise to 600 sequences. Actually, the number of publicly available sequences is 591, since few are missing. All sequences were recorded with a static camera at a 25 fps rate and have a size of 160 × 120. The second is the *facial expression* dataset [26], which is also publicly available.[3] It contains two individuals performing six different facial expressions, namely anger, disgust, fear, joy, sadness, and surprise. Each individual was asked to repeat each of the six expressions eight times under two different lightning setups, thus giving a total of 192 sequences. The third dataset is the HOHA one that is composed of 51 sequences of human actions in complex scenes and was provided by Ivan Laptev [25]. Most of the sequences contain complex background, background motion, clutter and occlusions. Additionally, changes in spatial scale and view variations are present.

We use the same recognition framework with that by Dóllar et al. [26] in order to provide fair comparisons and isolate the effect of the proposed salient point detector. We compare against two different interest point detectors,

[1] http://www.viperlib.york.ac.uk/, search for filename "horse2.mov".

[2] http://www.nada.kth.se/cvap/actions/.

[3] http://www.vision.ucsd.edu/∼pdollar/.

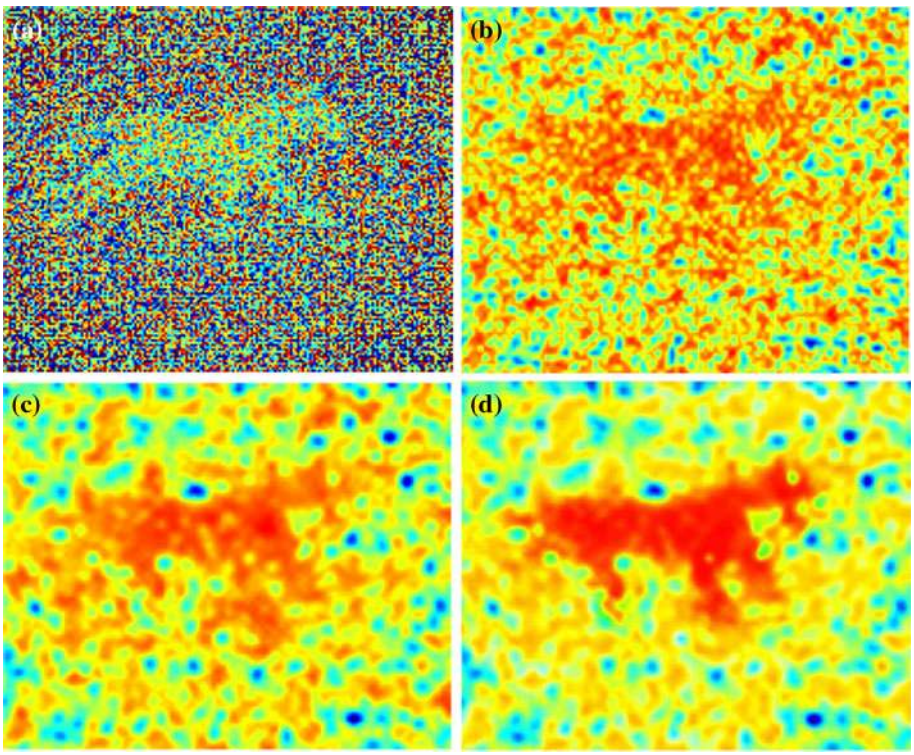

Fig. 7 **a** Initial frame of the moving horse illusion, saliency after **b** 1, **c** 10, and **d** 25 iterations. The horse becomes evident due to the enhancement of local spatiotemporal coherency

namely the one proposed by Laptev [23] and the one proposed by Dollár [26]. In both cases, we used existing implementations, which are publicly available by the authors. Laptev et al. detect local space–time interest points by extending the Harris corner detector to the spatiotemporal domain. The detector is based on the second-moment matrix, which describes the local gradient distribution at several spatial and temporal scales. The local maxima of this distribution—the counterpart of the proposed saliency volume—correspond to the points of interest. The rationale is that spatiotemporal corners, i.e. points at which the velocity vector is reversing direction, are well suited to represent human motions. Dollár et al. propose and evaluate a recognition framework that is based on a periodic point detector, which is tuned to fire whenever variations in local image intensities contain periodic frequency components. This is achieved by spatially smoothing the intensity volume and then temporally filtering it with a quadrature pair of 1D Gabor filters. Intuitively, the derived periodicity distribution is large where intensity changes temporally at a specific rate (specified by the frequency and scale of the Gabor filters).

Methodology and Results

Each of the action detection datasets is split into a training set D_{train} and testing set D_{test}. The training set $D_{\text{train}} = \{d_m\}$, $m = 1, \ldots, M$ is composed of M clips. Each training clip is represented by a set of interest points, which in our method are detected as the local maxima of the saliency volume **S**. There is no need to specify a threshold to alleviate non-important points as in the other two methods, because low saliency areas are quite smooth and no local maxima are found. A cuboid of fixed size is defined around each point, and a generalized PCA-SIFT descriptor [33] is extracted from each of them. In order to find the most representative points, we create a codebook $W = \{w_k\}$, $k = 1, \ldots, K$ of K visual words using k-means clustering on the set of descriptors from all points. The main idea is that although two instances of the same behavior may vary significantly in terms of their overall appearance and motion, many of the interest points they give rise to are similar. Finally, the salient points of each testing clip in D_{test} are associated with the most similar visual word, and a set of histograms is extracted by computing the histogram of visual words of each video. This is the counterpart of a co-occurrence matrix of size $K \times M$ with each indicating the occurrence of code-word w_k in the clip d_m.

For the *KTH* and *facial expressions* datasets, we compute a leave-one-out estimate of the error by training on all clips minus one, testing on the remaining one, and repeating the procedure for all clips, as in [26]. Codebooks are generated using a variable number of clusters, and classification is performed using an *1*-NN classifier. Figures 8 and 9 show the overall detection results for these two datasets. For the KTH dataset, the proposed detector performs better than the other two with an overall rate of

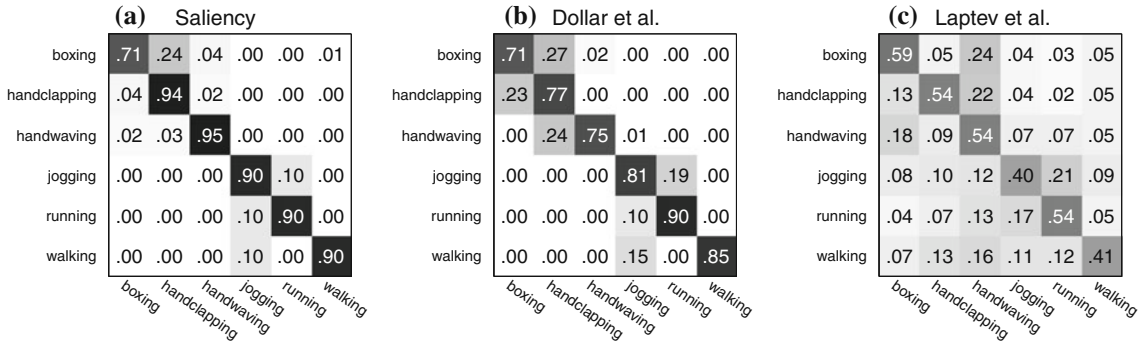

Fig. 8 Action recognition results on the KTH dataset. **a** Proposed model (overall rate: 88.3%), **b** Dollár et al. (overall rate: 79.83%) and **c** Laptev et al. (overall rate: 50.33%)

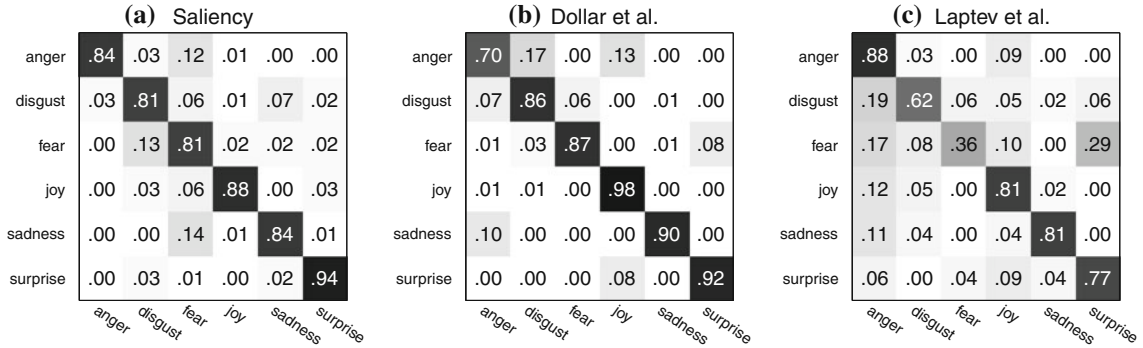

Fig. 9 Facial expression recognition results. **a** Proposed model (overall rate: 85.3%), **b** Dollár et al. (overall rate: 87.16%) and **c** Laptev et al. (overall rate: 70.83%)

88.3%. It achieves rates equal to or higher than 90% for all actions except the boxing one. The periodic detector achieves lower rates, with the ones related to the more dynamic actions (jogging, running, walking) being higher than the rest. It seems that the inherent periodicity of these actions is well represented by the Gabor filters.

The method by Laptev et al. ranks third with an overall rate of 50.33%. It seems that the spatiotemporal points of this method should be further adapted to scale and velocity, as the author propose, in order to achieve better performance. Figure 10 depicts detected features for the three detectors on a "jogging" sequence. Table 1 summarizes the results on the KTH dataset published so far. Currently, two recognition frameworks seem to be quite common, namely Nearest-Neighbor (NNC) and Support Vector Machines (SVM). Our detector achieves the best results among methods relying on NNC and is second best among all.

Conclusions from the statistics on *facial expression* recognition are slightly different. The periodic detector achieves a slightly higher rate (by 2% on average) than the proposed one with the model by Laptev et al. ranking third. It seems that the periodicity of the considered expressions is quite representative of the underlying actions, and therefore, it is better captured by this detector. Nevertheless, the salient detector achieves a rate higher than 80% for all actions and still performs better in actions "surprise" and "anger". Figure 11 depicts detected features for the three detectors on an "anger" sequence.

The *complex actions dataset* is the most challenging one, since it requires both successful—in terms of location—point detection and appropriate grouping in order to separate points belonging to different actions or generic events (e.g. background motion). Although grouping of points is out of the scope of this paper, we test the potential of the proposed method to operate efficiently in complex environments and use pLSA to recognize human actions in such sequences. The pLSA method has been used in the past both for spatial object [58] and spatiotemporal action recognition [27]. Classical pLSA was originally used in document analysis; in our case, documents correspond to video clips and visual words to cuboids. Apart from the codebook and visual words, which are observable from data and are extracted as described before, there also exists an unobservable variable, z, to describe the Z topics found in the training clips and associated with the visual words. The pLSA algorithm is used to learn the associations $P(w_k|z)$ (visual words that can be generated from a topic)

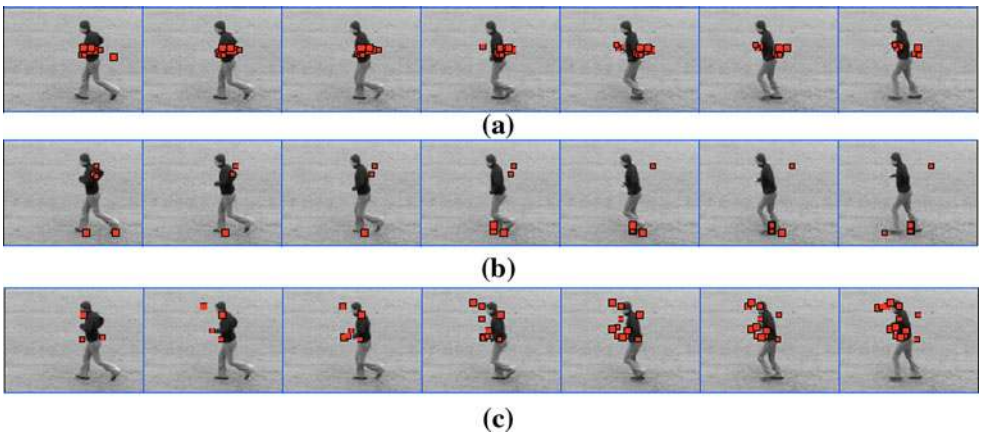

Fig. 10 Cuboids detected on a "jogging" sequence for the **a** proposed, **b** stHarris, and **c** periodic detectors

Table 1 Average recognition accuracy on the KTH dataset for different classification methods

Method	Accuracy (%)	Classifier
Schuldt et al. [53](reported in [29])	26.95	NNC
Schuldt et al. [53] (implemented by us)	50.33	NNC
Oikonomopoulos et al. [28] (reported in [29])	64.79	NNC
Wong et al. [29]	80.06	NNC
Dollár et al. [26] (implemented by us)	79.83	NNC
Dollár et al. [26]	81.20	NNC
Ours	88.30	NNC
Ke et al. [55]	80.90	SVM
Schuldt et al. [53]	71.70	SVM
Niebles et al. [27]	81.50	pLSA
Willems et al. [56]	84.36	SVM
Jiang et al. [57]	84.40	LPBOOST
Laptev et al. [54]	91.80	mc-SVM

Notice that the results proposed by Laptev et al. [54] are not directly comparable to ours, since the authors have used an extra optimization over a set of different descriptors

and $P(z|d)$ (the topic of clip d) through an Expectation-Maximization (EM) algorithm [59] applied to the co-occurrence matrix. As in the previous datasets, we used again the interest point extraction tools of [23, 26]; however, for recognition, we used our implementation of the approach described earlier.

Figure 12 shows the 20 most salient spatiotemporal points detected on a complex action clip for each method. The clip is labeled as "boxing" and shows a boxing action in the foreground with moving cars and someone running at the background. The periodic detector focuses more on the background, since periodic actions like the running man and the moving cars are present. The proposed method and the one by Laptev et al. detect more points on both observed actions and, therefore, make the distinction easier. This remark is confirmed for most complex clips, and this is the reason why the performance of both methods is higher. Statistics for the complex action recognition dataset are shown in Fig. 13 for the three tested methods. Overall, performance is quite lower compared with the first dataset, due to the diversity of the content and the lack of filtering or grouping of inconsistent points. Nevertheless, the results are indicative of the ability of the methods to concentrate either successfully or unsuccessfully around points that represent the observed action. We should emphasize here that the statistics of the method by Laptev et al. are lower than the ones reported in [25], because the recognition methods are different. The authors in [25] do not use pLSA and use a subset of features that match events in a similar training sequence rather than the whole set. Having said this, the proposed method performs considerably better that both other methods, by 20% on average.

Perceptually Salient Event Detection

Dataset for Evaluation

Evaluation of methods related to visual understanding has always been a challenge for researchers in the field, since human annotation is usually missing and—if existing—quite subjective. In an attempt to evaluate the proposed model, we use an online human annotated movie database [60] built within the MUSCLE Network of Excellence.[4] The movie clips of this database were viewed and annotated according to several cues including the audio, visual, and audio-visual saliency of their content. This means that parts of the clip were annotated as *salient* and *non-salient* depending on the importance of their content. The viewers

[4] Network of Excellence on Multimedia Understanding through Semantics, Computation and Learning (MUSCLE), FP6-507752, 2004–2007.

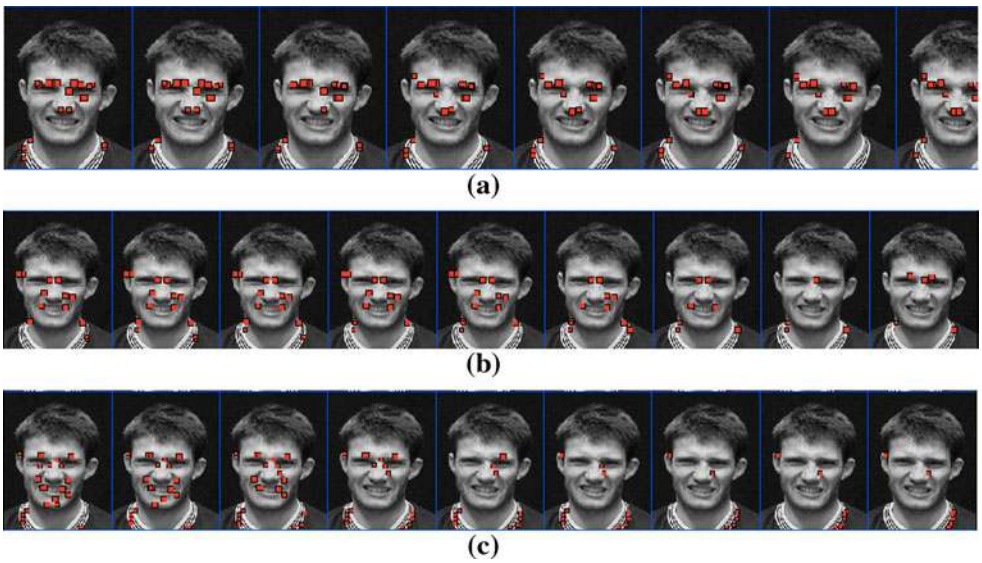

Fig. 11 Cuboids detected on an "anger" sequence for the **a** proposed, **b** stHarris and **c** periodic detectors

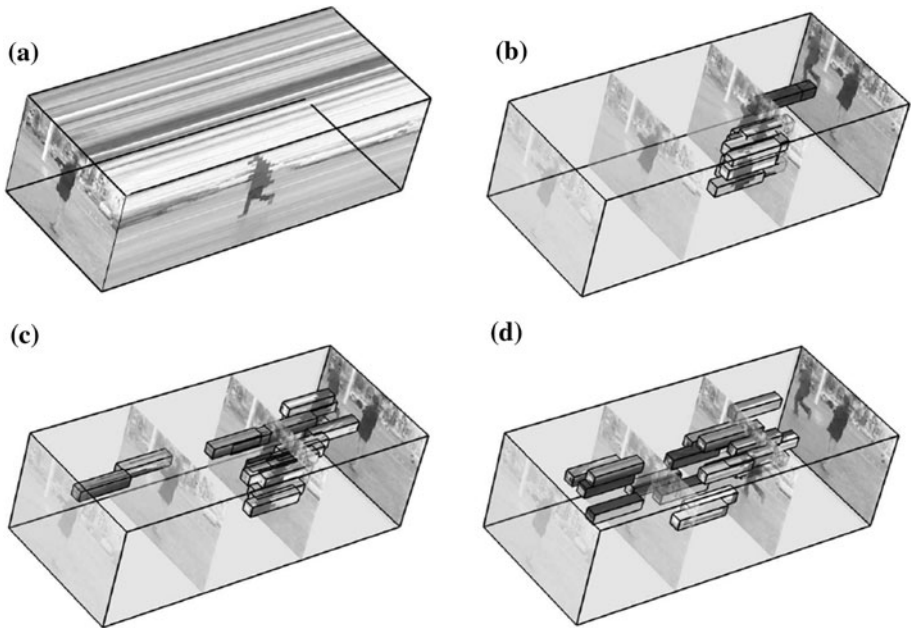

Fig. 12 Action recognition results on complex sequences. **a** Initial volume of a boxing action with complex background and the 20 most salient interest points of the **b** proposed method, **c** Dollár et al. and **d** Laptev et al

were asked to assign a saliency factor to these parts according to loose guidelines; strict rules cannot be applied due to the high subjectivity of the task. The guidelines are related to the audio, visual, and audio-visual saliency and are not related to the semantic interpretation of the content. This means that the viewers were asked to rely—as much as possible—on their spontaneous understanding of the scene (e.g. abrupt visual changes, loud noise, etc) and not on scene importance assignment due to content interpretation (e.g. a car on the highway is more important than others, because it is driven by the leading actor). We exploit the visual saliency annotation and provide experimental results on salient event detection in movie clips.

The database contains three annotated clips of 10–13 min each, selected from the films "300" (10701 frames), "Lord of the Rings I" (14001 frames), and "Cold Mountain" (16150 frames). Due to their short length compared to the whole movie, these clips are not exactly representative of the movie genre. The clip from "300" contains many fighting scenes from the original movie and

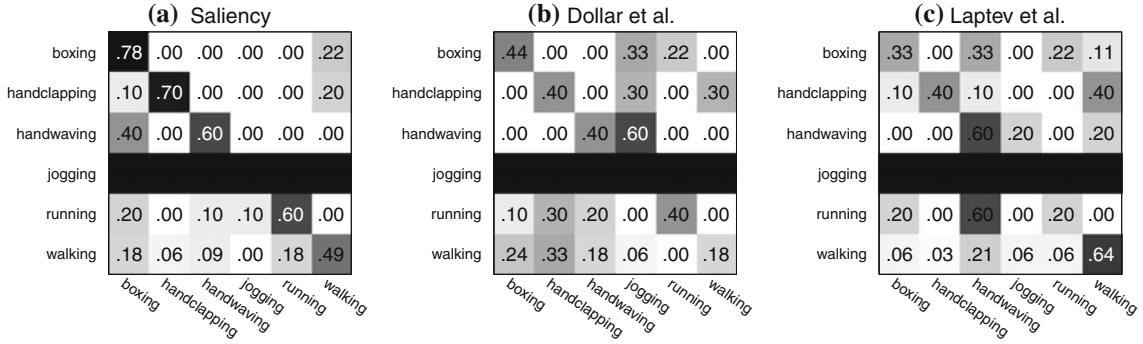

Fig. 13 Action recognition results on the complex sequence dataset (the jogging action is not present). **a** Proposed model (overall rate: 63.2%), **b** Dollár et al. (overall rate: 35.6%) and **c**Laptev et al. (overall rate: 43.4%)

therefore can be easily considered as an action clip. The selected part of the "Lord of the Rings I" is taken from the beginning of the movie and does not contain any intense battle scenes, as in the rest of the film. It is mainly composed of indoor scenes with people discussing and few "peaceful" outdoor scenes, thus it can be rather considered as an atmospheric fantasy film with short action parts. The last clip extracted from "Cold Mountain" consists of outdoor scenes with almost no action and can be considered as a drama clip.

Methodology

The target here is to create an attention curve that will encode the saliency factor of a specific temporal video segment, which can be of a fixed frame length. For this, we follow a rationale similar to the action recognition case; However, a single saliency value is assigned to each of the temporal segments in order to create the attention curve. The notion of saliency in movies is slightly different than the one we have used so far: salient/unusual events occur in the background of usual events. This means that throughout the movie, there are typical events that occur quite frequent and sparse—or even rare—events that are usually associated with interesting parts. We model this notion of saliency in movie clips as an inlier/outlier representation, where salient events are considered as outliers of the attention curve.

Following the framework described in "Methodology and Results", we create a separate codebook for each movie, representing typical events. The codebook is created by clustering all interest points using K-means. If a clip cannot be represented well in terms of this codebook, this means that the clip is rather an outlier or unusual event. We split each clip in segments of length 64 frames ($\simeq$2.5 s, since all videos have a frame rate of 25 frames/s). The saliency value of a segment is then obtained by computing the Hausdorff distance between the set of prototypes in W and the set of cuboids U extracted from the segment. Hence, the attention curve A is equal to the Hausdorff distance between these finite point sets and is defined as

$$A(W, U) = \max(h(W, U), h(U, W)) \tag{20}$$

where

$$h(W, U) = \max_{w \in W} \min_{u \in U} \rho(w, u) \tag{21}$$

is the directed Hausdorff distance from W to U. The distance ρ between w and u is the Euclidean one.

The comparison against the available ground truth is not a straightforward task. During the annotation procedure, the observers are able to almost automatically integrate and detect salient visual information across many frames. The result of this operation is not directly comparable with the results of the proposed method, since the spatiotemporal nature of the visual part depends highly on the chosen frame neighborhood rather than on biological evidence. The output of this annotation is a saliency indicator function I_{sal}, which is depicted as the green line in Fig. 14. A value of 1 means that the underlying temporal range is salient, while the opposite signifies a non-salient area.

Experiments

We formulate the problem of comparing the visual saliency curve against the ground truth as a problem of tuning two different parameters, namely the number of visual words K and a threshold T used to create the binary decision about salient (outlier) and non-salient (inlier) parts of the curve. Experiments revealed a reasonable relation between these two parameters. A large number of visual words will eventually describe better the underlying movie structure and produce a peaky attention curve (salient events become sparser and more dissimilar to the background), while a

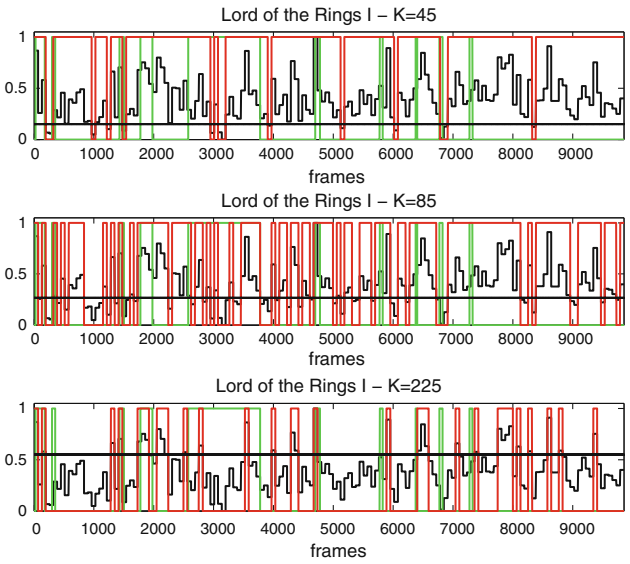

Fig. 14 Salient temporal segment detection using the proposed model for the movie "Lord of the Rings I". Each plot shows the human Annotation (*green*), the attention Curve (*black*), the detected salient segments (*red*) and the computed threshold (*horizontal black*) (Color figure online)

small number will produce a smoother curve. A peaky curve should be accompanied by a high threshold and a less peaky one by a lower threshold. We, therefore, relate these parameters using an exponential function

$$T(K) = (1/B)\exp(-K/B) \tag{22}$$

where B is the scale. For the experiments, we set $B = 0.5$ and normalize $T(K)$ so as $T(0) = 1$.

This behavior is illustrated in Fig. 14, where three different results for one movie corresponding to three different values of K are shown. The attention curve is shown as black, the ground truth as green, the detected salient temporal segments as red, and the computed threshold as the horizontal black line. Notice how the detected salient events become sparser as the number of visual words increases. Determining the appropriate length of the salient events is not straightforward. In [44], under a quite different framework, a median filter of varying length is used and the one giving the best statistics is selected. In this work, we prefer not to insert another parameter in the approach and we use directly the output of the thresholding scheme as the length of the underlying salient event.

Figure 15 shows precision–recall curves for the three annotated clips. The curves are computed by varying the number of visual words K used to create the codebook. As expected, the recall values are higher over a broader range, since high precision in such a subjective application is hard to achieve. Nevertheless, all methods reach an adequate level of precision for all movies, with the proposed one being apparently higher for two of the three movies. Interestingly, enough in the "300" case, all methods

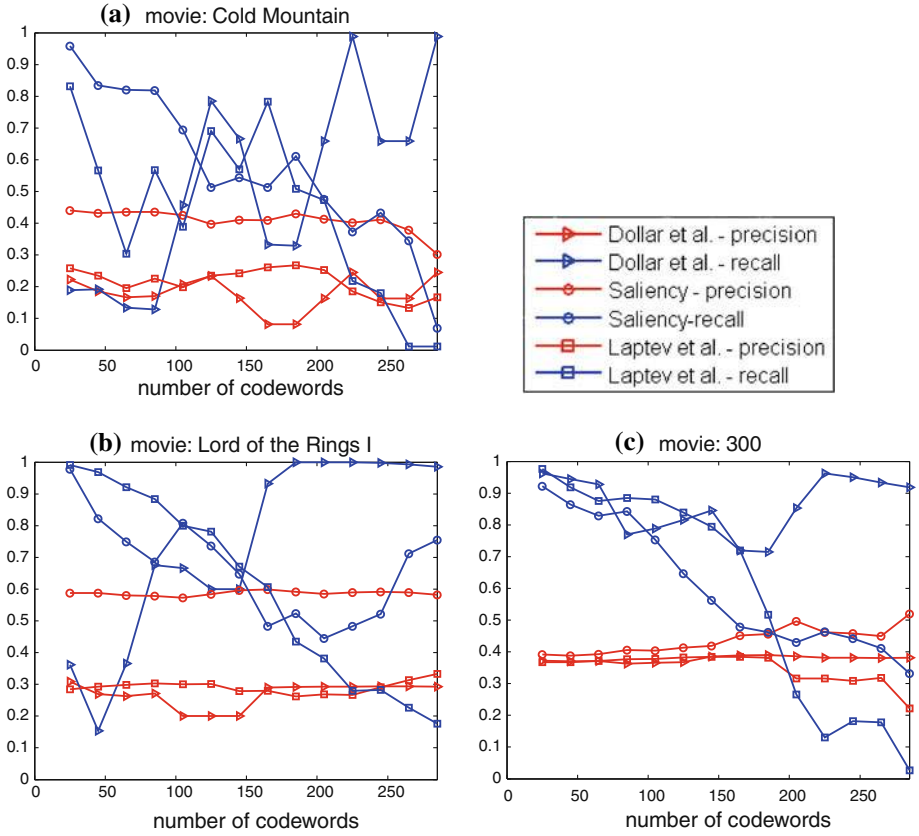

Fig. 15 Precision (*red*) and recall (*blue*) curves for movies **a** "300", **b** "Lord of the Rings I", and **c** "Cold Mountain" (Color figure online)

perform almost equally both in terms of precision and recall. It seems that the action content of this movie and the peculiarities of the digital scenery (sharp colors, cartoon like characters, etc.) give rise to similar spatiotemporal points for the three methods. For example, fight actions in the battle field—as depicted in the film—give rise to pop-out events in terms of color/motion/intensity (proposed detector), periodic movements (Dollár et al. detector) and abrupt motion direction changes that thoroughly or partially correspond to spatiotemporal corners (Laptev et al. detector).

Conclusion

We have proposed a computational model for visual saliency based on spatiotemporal analysis. The model is derived by combining ideas of established visual saliency models and applying them to a volumetric representation of the video sequence. The iterative competition between different levels of the representation is implemented using constraints inspired from the Gestalt laws. We demonstrate the model on action recognition and salient event detection in video sequences. The results reveal behavioral details of the proposed model and provide a rigorous analysis of the advantages and disadvantages of all methods involved in the comparisons. It seems that the periodic detector by Dollár et al. fits well with actual periodic actions like the facial expressions, while the spatiotemporal points by Laptev et al. need further adaptation to scale or velocity to perform well, as the same authors propose. Overall, the proposed model performs quite well to the considered applications and either outperforms the techniques it is compared to or performs almost equally to them.

Appendix

The partial derivative of the intra-feature constraint in Eq. 16 is obtained as follows: Given that

$$A_{k,m}(s) = C_{k,m}(s) - \frac{1}{|N_q|} \cdot \sum_{r \in N_q} C_{k,m}(r)$$

$$B_{k,m}(q,s) = \frac{\partial}{\partial C_{k,m}(s)} \cdot \left[\frac{1}{|N_q|} \cdot \sum_{r \in N_q} C_{k,m}(r) \right],$$

$$B_{k,m}(q,s) = \begin{cases} 1 & \text{if} \quad s \in N_q \\ 0 & \text{if} \quad s \neq N_q \end{cases}$$

the partial derivative is computed by

$$\begin{aligned}
\frac{\partial E_1(\mathbf{C})}{\partial C_{k,m}(s)} &= \sum_s \frac{\partial}{\partial C_{k,m}(s)} \left(C_{k,m}(s) - \frac{1}{|N_q|} \sum_{r \in N_q} C_{k,m}(r) \right)^2 \\
&= 2 \cdot \sum_s A_{k,m}(s) \cdot \frac{\partial}{\partial C_{k,m}(s)} \left(A_{k,m}(s) \right) \\
&= 2 \cdot \left[\left(A_{k,m}(s) \cdot \frac{\partial}{\partial C_{k,m}(s)} (C_{k,m}(s)) \right) - \sum_s A_{k,m}(s) \cdot B_{k,m}(q,s) \right] \\
&= 2 \cdot \left[A_{k,m}(s) - \sum_{q \in N(s)} \frac{1}{|N_q|} A_{k,m}(q) \right] \\
&= 2 \cdot \left[C_{k,m}(s) - \frac{1}{N^2} \cdot \sum_{q \in N(s)} \left(2N \cdot C_{k,m}(q) - \sum_{r \in N_q} C_{k,m}(r) \right) \right]
\end{aligned}$$

The partial derivative of the inter-feature constraint in Eq. 16 is obtained as

$$\begin{aligned}
\frac{\partial E_2}{\partial C_{k,\ell}(s)} &= \sum_i \frac{\partial}{\partial C_{k,\ell}(s)} \left(C_{k,m}(s) - \frac{1}{M-1} \sum_{j \neq i} C_{j,\ell}(s) \right)^2 \\
&= 2 \cdot \sum_i C_{k,m}(s) - \frac{1}{M-1} \cdot \sum_{j \neq i} C_{j,\ell}(s) \cdot \frac{\partial}{\partial C_{k,\ell}(s)} \\
&\quad \times \left(C_{k,m}(s) - \frac{1}{M-1} \sum_{j \neq i} C_{k,m}(s) \right) \\
&= 2 \cdot \left[\left(C_{k,\ell}(s) - \frac{1}{M-1} \sum_{j \neq k} C_{j,\ell}(s) \right) \right. \\
&\quad \left. - \frac{1}{M-1} \cdot \sum_{i \neq k} \left(C_{k,m}(s) - \frac{1}{M-1} \cdot \sum_{j \neq i} C_{j,\ell}(s) \right) \right] \\
&= 2 \cdot \frac{M}{M-1} \cdot \left(C_{k,\ell}(s) - \frac{1}{M-1} \cdot \sum_{j \neq k} C_{j,\ell}(s) \right)
\end{aligned}$$

with

$$\frac{\partial}{\partial C_{k,\ell}(s)} \left(C_{k,m}(s) \right) = \begin{cases} 1 & \text{if} \quad i = k \\ 0 & \text{if} \quad i \neq k \end{cases}$$

$$\frac{\partial}{\partial C_{k,\ell}(s)} \left(\sum_{j \neq i} C_{k,m}(s) \right) = \begin{cases} 1 & \text{if} \quad i \neq k \\ 0 & \text{if} \quad i = k \end{cases}$$

$$\sum_{i \neq k} \sum_{j \neq i} C_{j,\ell}(s) = (M-1) \cdot C_{k,\ell}(s) + (M-2) \cdot \sum_{j \neq k} C_j(s)$$

References

1. James W. The principles of psychology; 1890.
2. Treisman A. A feature-integration theory of attention. Cogn Psychol. 1980;12(1):97–136.
3. Koch C. Shifts in selective visual attention: towards the underlying neural circuitry. Comput Vis Image Underst 1985;4:219–27.

4. Itti L, Koch C. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 1998;20(11):1254–59 doi:10.1109/34.730558.
5. Milanese R, Gil S. Attentive mechanisms for dynamic and static scene analysis. Opt Eng 1995;34(8):2428–34.
6. Rapantzikos K, Tsapatsoulis N. Spatiotemporal visual attention architecture for video analysis. In: Multimedia signal processing, 2004 IEEE 6th workshop on 2004. p. 83–6.
7. Rapantzikos K. Coupled hidden markov models for complex action recognition. In: An enhanced spatiotemporal visual attention model for sports video analysis, international workshop on content-based multimedia indexing (CBMI); 2005.
8. Rapantzikos K, Tsapatsoulis N. A bottom-up spatiotemporal visual attention model for video analysis. IET Image Process. 2007;1(2):237–48.
9. Rapantzikos K, Avrithis Y. Dense saliency-based spatiotemporal feature points for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE; 2009.
10. Koffka K. Principles of Gestalt psychology. New York: Harcourt; 1935.
11. Wertheimer M. Laws of organization in perceptual forms (partial translation). A sourcebook of Gestalt psychology; 1938.
12. Wolfe J. Visual search in continuous, naturalistic stimuli. Vis Res(Oxf). 1994;34(9):1187–95.
13. Tsotsos JK, Culhane SM. Modeling visual attention via selective tuning. Artif Intell. 1995;78(1–2):507–45.
14. Harris C. A combined corner and edge detector. In: Alvey vision conference 1988. vol 15, p. 50.
15. Lindeberg T. Feature detection with automatic scale selection. Int J Comput Vis. 1998;30(2):79–116.
16. Kadir T. Saliency, scale and image description. Int J Comput Vis 2001;45(2):83–105.
17. Mikolajczyk K, Tuytelaars T. A comparison of affine region detectors. Int J Comput Vis 2005;65(1):43–72.
18. Csurka G, Dance C. Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV; 2004. p. 1–22.
19. Mikolajczyk K. An affine invariant interest point detector. Lect Notes Comput Sci. 2002;128–142.
20. Koch C. Visual categorization with bags of key-points. In: European conference on computer vision; 2004. p. 1–22.
21. Hao YW. Unsupervised discovery of action classes. Comput Vis Pattern Recognit. 2006;2:17–22
22. Bosch A, Zisserman A. Scene classification via plsa. In: European conference on computer vision; 2006. p. 517–30.
23. Laptev I. Interest point detection and scale selection in space-time. Lect Notes Comput Sci. 2003;372–387.
24. Laptev I. On space-time interest points. Int J Comput Vis. 2005;64(2):107–23.
25. Laptev I, Caputo B. Local velocity-adapted motion events for spatio-temporal recognition. Comput Vis Image Underst. 2007; 108(3):207–29.
26. Dollár P, Rabaud V. Behavior recognition via sparse spatio-temporal features. In: 2nd Joint IEEE International Workshop on visual surveillance and performance evaluation of tracking and surveillance, 2005; 2005. p. 65–72.
27. Niebles JC, Wang H. Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis. 2008;79(3):299–318.
28. Oikonomopoulos A, Patras I. Spatiotemporal salient points for visual recognition of human actions. IEEE Trans Syst Man Cybern B Cybern. 2006;36(3):710–9.
29. Wong SF. Extracting spatiotemporal interest points using global information. In: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007; 2007. p. 1–8.
30. Blank M, Gorelick L. Actions as space-time shapes. In: Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005; 2005. p. 2.
31. Shechtman E. Space-time behavior based correlation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005; 2005. p. 1.
32. Shechtman E. Matching local self-similarities across images and videos. In: Proceedings of CVPR; 2007.
33. Ke Y. PCA-SIFT: a more distinctive representation for local image descriptors. In: IEEE Compuer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 1999; 2004. p. 2.
34. Hamid R, Johnson A. Detection and explanation of anomalous activities: representing activities as bags of event n-Grams. In: IEEE Compuer Society Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 1999; 2005. vol 1, p. 1031.
35. Itti L. A principled approach to detecting surprising events in video. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR05); 2005. p. 1.
36. Zhong H, Shi J. Detecting unusual activity in video. In: IEEE Compuer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society 1999; 2004. p. 2.
37. Ma YF, Lu L. A user attention model for video summarization. In: Proceedings of the tenth ACM international conference on Multimedia. New York: ACM; 2002. p. 533–42.
38. Ma YF, Hua XS. A generic framework of user attention model and its application in video summarization. IEEE Trans Multimed. 2005;7(5):907–19.
39. Boiman O. Detecting irregularities in images and in video. In: IEEE International Conference on Computer Vision (ICCV); 2005.
40. Stauffer C. Learning patterns of activity using real-time tracking. IEEE Trans Pattern Anal Mach Intell. 2000;22(8):747–57.
41. Adam A, Rivlin E. Robust real-time unusual event detection using multiple fixed-location monitors. IEEE Trans Pattern Anal Mach Intell. 2008; 555–60.
42. Evangelopoulos G, Rapantzikos K. Movie summarization based on audiovisual saliency detection. In: 15th IEEE International Conference on Image Processing, 2008. ICIP 2008; 2008. p. 2528–31.
43. Evangelopoulos G. Video event detection and summarization using audio, visual and text saliency. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP09); 2009. p. 3553–6.
44. Rapantzikos K, Evangelopoulos G. An audio-visual saliency model for movie summarization. In: Proceedings of IEEE International Workshop on Multimedia Signal Processing (MMSP07); 2007.
45. Caspi Y. Spatio-temporal alignment of sequences. IEEE Trans Pattern Anal Mach Intell (PAMI). 2002;24:1409–24.
46. Tikhonov AN. Solution of ill-posed problems. Washington, DC: W. H. Winston; 1977.
47. Rapantzikos K. Robust optical flow estimation in MPEG sequences. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005. (ICASSP'05); 2005. p. 2.
48. Hering E. Outlines of a theory of the light sense. Cambridge: Harvard University Press; 1964.
49. Freeman WT. The design and use of steerable filters. IEEE Trans Pattern Anal Mach Intell. 1991;13(9):891–906.
50. Derpanis KG. Three-dimensional nth derivative of gaussian separable steerable filters. In: IEEE International Conference on Image Processing; 2005.
51. Wildes R. Qualitative spatiotemporal analysis using an oriented energy representation. Lect Notes Comput Sci. 2000;768–84.
52. Riedmiller M. Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms.

Int J Comput Stand Interf Spec Issue Neural Netw. 1994;16: 265–78.
53. Schuldt C, Laptev I. Recognizing human actions: a local SVM approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. 2004. p. 3.
54. Laptev I, Marszalek M. Learning realistic human actions from movies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2008.
55. Ke Y, Sukthankar R. Spatio-temporal shape and flow correlation for action recognition. In: 7th International Workshop on Visual Surveillance; 2007.
56. Willems G, Tuytelaars T. An efficient dense and scale-invariant spatio-temporal interest point detector. In: Lecture Notes in Computer Science, Marseille, France; 2008. p. 650–3.
57. Jiang H, Drew MS. Successive convex matching for action detection. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition; 2006. p. 1646–53.
58. Sivic J, Russell BC, Efros A, Zisserman A, Freeman WT. Discovering object categories in image collections. In: International Conference on Computer Vision (ICCV05); 2005.
59. Dempster AP, Laird NM. Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc 1977;39(1):1–38.
60. MUSCLE. Movie dialogue dataBase v1. 1. Aristotle University of Thessaloniki, AIILab; 2007.