Chapter 11
Vision, Attention Control, and Goals Creation System

Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kolias

Abstract Biological visual attention has been long studied by experts in the field of cognitive psychology. The Holy Grail of this study is the exact modeling of the interaction between the visual sensory and the process of perception. It seems that there is an informal agreement on the four important functions of the attention process: (a) the *bottom-up* process, which is responsible for the *saliency* of the input stimuli; (b) the *top-down* process that bias attention toward known areas or regions of predefined characteristics; (c) *the attentional selection* that fuses information derived from the two previous processes and enables focus; and (d) the *dynamic evolution* of the attentional selection process. In the following, we will outline established computational solutions for each of the four functions.

11.1 Overview

Most of our impressions and memories are based on vision. Nevertheless vision mechanisms and functionalities are still not apparent. How do we perceive shape, color, or motion and how do we automatically focus on the most informative parts of the visual input? It has been long established that primates, including human, use focused attention and fast saccades to analyze visual stimuli based on the current situation or the desired goal. Neuroscientists have proven that neural information related to shape, motion, and color is transmitted through, at least, three parallel and interconnected channels to the brain rather than a single one. Hence a second question arises related to how these channels are "linked" in order to provide useful information to the brain.

The Human Visual System (HVS) creates a perceptual representation of the world that is quite different than the two dimensional depiction of the retina.

K. Rapantzikos (✉)
Image, Video and Multimedia Systems Laboratory, Computer Science Division,
School of Electrical and Computer Engineering, National Technical University of Athens,
Iroon Polytexneiou 9, 15780 Zografou, Greece
e-mail: rap@image.ece.ntua.gr

V. Cutsuridis et al. (eds.), *Perception-Action Cycle: Models, Architectures, and Hardware*, Springer Series in Cognitive and Neural Systems 1,
DOI 10.1007/978-1-4419-1452-1_11, © Springer Science+Business Media, LLC 2011

This perceptual representation enables us to, e.g., identify same objects under quite different conditions (illumination, noise, perspective distortion, etc.), especially when we are searching for a target object (*goal*). These abilities led the experts to distinguish between *bottom-up* (stimuli-based) and *top-down* (goal-oriented) processes in the HVS. Bottom-up attentional selection is a fast, and often compulsory, stimulus-driven mechanism. Top-down attentional selection initiates from the higher cognitive levels in the brain that influence the attentional system to bias the selection in favor of a particular (or a combination of) feature(s). Only information about the region that is preattentively extracted can be used to change the preferences of the attentional system.

11.2 Computational Models of Visual Attention

From the computer vision point of view, all aspects related to the visual attention mechanism are quite important. Creating constraints about what and where to "look" is a great benefit for, e.g., object detection and recognition tasks, visual surveillance, and learning of unknown environments (robotic vision). Computational modeling of visual attention is mainly related to *selective visual attention* that includes *bottom-up* and *top-down* mechanisms, *attentional selection*, and *dynamic evolution* (Sternberg 2006).

11.2.1 Bottom-Up Visual Attention

The *Feature Integration Theory* (*FIT*), introduced in 1980 (Treisman and Gelade 1980), is considered as the seminal work for computational visual attention based on a *master map*. The theory inspired many researchers and evolved toward current research findings. The main idea is that "different visual features are registered automatically and in parallel across the visual field, while objects are identified separately and only thereafter at a later stage, which requires focused attention" (Treisman and Gelade 1980).

According to this theory, the visual features (e.g., intensity, color, orientation, size, and shape) are linked after proper attentional processing. This observation is in accordance with the two-stage attentional process of James (James 1890/1981), namely the preattentive and attentive ones. Information from the *feature maps* – topographical maps that highlight saliency according to the respective feature – is collected in the master map. This map specifies *where* (in the image) the entities are situated, but not *what* they are. Scanning serially through this map directs the focus of attention toward selected scene entities and provides data useful for higher perception tasks. Information about the entities is gathered into so-called *object files*. An object file is a midlevel visual representation that "sticks" to a moving object over time on the basis of spatiotemporal properties and stores (and updates) information about that object's properties.

Based on the FIT model, Koch and Ullman introduced a complete computational architecture of visual attention (Koch and Ullman 1985). The idea is that several features are computed in parallel, and their conspicuities are collected in a saliency map (the equivalent of a master map). A *Winner-Take-All* (WTA) network determines the most salient location in this map, which is routed to a central representation, where more complex processing might take place. One of the most widely exploited computational models is the one introduced by Itti et al. (1998), which is based on the model of Koch and Ullman and hypothesize that various visual features feed into a unique saliency map (Koch and Ullman 1985) that encodes the importance of each minor visual unit. Among the wealth of methods based on saliency maps, the differences are mainly due to the selection of feature maps to be included in the model [contrast (May and Zhang 2003), color (Itti et al. 1998), orientation (Itti et al. 1998), edges (Park et al. 2002), size, skin maps (Rapantzikos and Tsapatsoulis 2005), texture, motion (Rapantzikos and Tsapatsoulis 2005; Milanese et al. 1995; Abrams and Christ 2003), etc.], the linear, or nonlinear operations [multiscale center-surround (Itti et al. 1998), entropy (Kadir and Brady 2001), etc.] applied to these maps and the final fusion to produce the saliency map (min/max/sum fusion, learning, etc.). Recent methods use more advanced features like wavelet responses learned through Independent Component Analysis (ICA) (Bruce and Tsotsos 2009; Torralba et al. 2006).

Computing saliency involves processing the individual feature maps to detect distinguished regions and fusing them to produce the final saliency distribution of the input. A well known biologically motivated operator is the center-surround one that enhances areas that pop out from the surroundings (Itti et al. 1998). This operator presumes a multiscale representation of the input with the center lying on a fine level and the surround on a coarser one. Kadir et al. (2001) use an entropy measure to compute saliency and derive salient regions that represent well the underlying scene. Their detector is based on the entropy measure in a circle around each pixel (spatial saliency) and in a range of scales (scale saliency). Improvements of this detector have been proposed focusing mainly on making it robust under affine transformations of the input or computationally lighter (Shao et al. 2007).

Recently, information theory has been used to model the importance, saliency, or surprise [a notion introduced by Itti and Baldi (2009)]. As the authors claim "Life is full of surprises, ranging from a great Christmas gift or a new magic trick, to wardrobe malfunctions,…" (Itti and Baldi 2009), and therefore Itti and Baldi proposed a mathematical model of surprise based on Bayesian theory. Based on knowledge or experience, an observer is watching a scene having a priori thoughts about it. Hence, surprise is defined as a great change between what is expected and what is observed. The expectations of the observer could either be modeled using a priori information or by the temporary scene context. Toward this direction, Bruce and Tsotsos (2006, 2009) compute saliency based on information maximization. They define saliency by quantifying Shannon's self information of a local image patch. In order to produce a reasonable estimate of the probability distribution – necessary to compute the information measure – the authors perform ICA to a large sample of patches from natural images. If a patch can be adequately predicted by this distribution, then it is not considered salient. Overall, entropy-based approaches put emphasis on the structural complexity of the scene, while information maximization methods enhance regions that differ from their context. Both approaches are biologically plausible as shown by experiments in the field (Treisman and Gelade 1980; Leventhal 1991; Sillito and Jones 1996).

A biologically plausible solution of attentional selection has been proposed by Tsotsos et al. (1995) that is also based on the spatial competition of features for saliency. The *selective tuning model* exploits a hierarchy of WTA networks that represent the input at increasing scales. The higher levels (coarse) guide the lower ones (finer) toward selecting salient regions of increasing detail, which in turn provide feedback to their "ancestors." Practically, each level is a saliency map that communicates with the neighboring levels to define the saliency value of each neuron. The main advantage of this model is the fact that there is no need for a single – and often oversimplifying – centralized saliency map, since saliency is computed through *distributed competition*.

11.2.2 Top-Down Visual Attention

Bias toward specific features or areas of interest can be incorporated through a top-down channel independently of the bottom-up architecture. Top-down influence can be modeled either by biasing specific areas of the saliency map or by adapting the weights of the intermediate conspicuity maps. Correspondingly in the selective tuning model, the top-down weights of the hierarchy can be used to bias attention toward a known target. Such an approach has certain similarities with the Guided-Search Model (GSM) proposed by Wolfe et al. (Wolfe 1994, 2007; Wolfe et al. 1989), where the weights of the features or the neurons are changed according to existing knowledge. Unlike FIT this model does not follow the idea of separate maps for each feature type, it defines only one map for each feature dimension, and within each map different feature types are represented. Comparable to the saliency map of location in FIT, there is an activation map in which the feature maps are fused. But in contrast to at least the early versions of FIT, in GSM the attentive part profits from the results of the preattentive one. The fusion of the feature maps is done by summing them. Additionally to this bottom-up behavior, the model includes the influence of a priori knowledge about the target by maintaining a top-down map that selects the feature type which distinguishes the target best from its distracters. Frintrop et al. have recently proposed similar attention models and applied them to real applications (Frintrop et al. 2005; Frintrop and Cremers 2007). Lee et al. (2005) use a neural network-based model with higher level cues guiding the attention (shapes, faces, etc.).

11.2.3 Attentional Selection: Attention as a Controller

The simpler approach to select the region to attend is the use of a fixed shape (e.g., a circle of fixed radius) located at the maximum of the saliency distribution.

Moving the focus-of-attention (FOA) is an important part of the models. In order to avoid revisiting the same area, most of them use an *inhibition-of-return* approach that does not permit successive focus of the same region. This is the approach, e.g., of the standard model of Itti and Koch, where FOA is the only active one at each iteration (the rest are ignored). This is the approach adopted by most computational models so far.

Nevertheless, experimental evidence exists that supports the direct selection of objects to attend, rather than a fixed shape region around them (Duncan 1984) (Duncan's Integrated Competition Hypothesis). This is related to the notion of proto-objects introduced by Walther and Koch (2006), where the FOA is adapted to the objects we expect to find in the scene. The shape of the FOA is formed by thresholding the conspicuity map of the Itti's model that contributes more to the saliency map. The resulting FOA mask is then used as a top-down bias to the attentional selection process. Sun and Fisher (2003) use a different notion of attentional selection based on Gestalt groupings. When a potential object is attended, the model extracts its groupings (collections of pixels, features, patches, etc.) and compares them against the next FOA. If the groupings of the attended areas remain similar, then the model considers the underlying FOA as belonging to the same object. On the contrary, if the groupings are dissimilar, then another object is attended.

Recently, Cutsuridis proposed a cognitive model of saliency overt attention, and natural picture scanning that unravels the neurocomputational mechanisms of how human gaze control operates during active real-world scene viewing. It is based on both the resonance of top-down and bottom-up processes (Cutsuridis 2009). The resonance is accomplished by a value module (dopamine) that ensures the similarity of the top-down and bottom-up salient representation by increasing/decreasing the SNR of neural representations. The model is heavily supported by the neuroscientific evidence and addresses important questions in active vision (overt attention). It also provides the insights on the neural mechanisms of inhibition of return and focus of attention.

11.2.4 CODAM: COrollary Discharge of Attention Movement

Narrowing down the role of attention to controlling the evolution of the FOA after creating a saliency map may work well for few applications, but it is far from the real. Attention is the result of a complex interplay between numbers of brain modules across parts of the brain, as well as complex temporal dynamics involved in this interplay, which are not necessarily modeled in the previous systems. As said, attention acts as a filter to concentrate processing resources on a particular salient stimulus and remove distracters. The attended stimulus, held for a period on the attended state estimator buffer, can then be processed by further higher level mechanisms as might be involved, for example, in thinking, reasoning, comparison, imagining, and so on.

CODAM proposed by Taylor et al. (2000, 2003, 2007) is a neural network model for the brain-based creation of awareness or consciousness that uses the attention copy signal as the crucial component. The attention movement signal generator (inverse model controller or IMC), biased by a set of predefined goals, sends a new attention signal to lower level cortical stimulus activity; this can be summarized as a two-stage model, in which the higher level control system generators (goals and IMC) send attention signals to lower level cortical representations (Taylor et al. 2007). The Control model is composed of the following modules: (a) input modules (early visual hierarchy in vision); (b) inverse model controller (IMC), as the generator of a feedback attention control signal to amplify the attended stimulus activity and reduce that of distracters (acting as a saliency map, running a competitive process to attend to the most salient stimulus input); and (c) goals module, to allow for endogenous bias to the IMC for goals created from other sources in a top-down manner.

11.3 Applications

Applications are numerous, since saliency fits well fit with many of the existing approaches in computer vision. We present a concise report on applications of both spatial and spatiotemporal saliency models.

11.3.1 Scene/Object Recognition

Rutishauer et al. (2004) 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. They show that recognition performance for objects in highly cluttered scenes can be improved dramatically and that other problems, such as learning multiple objects from single images, are only possible using attention. This evidence is further confirmed in Walther et al. (2004, 2005). Specifically, the authors combine Itti's model with a recognition model based on Lowe's SIFT features (Lowe 1999). The final process includes a top-down inhibition-of-return method that prohibits areas of nonproto objects (Walther and Koch 2006) to become salient. The notion of *proto objects* is described by Rensink as the volatile units of visual information that can be bound into a coherent and stable object when accessed by focused attention (Rensink 2000).

Torralba (Koch and Ullman 1985; Treisman and Gelade 1980) integrates saliency with context information (task driven focus-of-attention) and introduces a simple framework for determining regions-of-interest within a scene. They use context learned by training to guide attention (Torralba 2003; Torralba et al. 2006). Various features are extracted from a large collection of images, which are used to train a classifier. The top-down process consists of the classifier that guides the attention toward scene parts that most probably contain the target object. The main concept is that the location of certain natural objects is constrained by context, since, e.g., a car is expected to be found on the road rather than on the sky. A top-down process that biases bottom-up feature weights has been also incorporated to further enhance target objects.

Navalpakkam and Itti (2005, 2006) propose a modification of the basic Itti's model inspired by the Guided Search of Wolfe. They carry out a range of experiments both for visual search (fast search in pop-out tasks, slow search in conjunction tasks, slow search when the target is similar to the distracters, etc.) and visual recognition. With little computational cost incurred through multiplicative top-down weights on bottom-up saliency maps, their model combines both stimulus-driven and goal-driven attention to optimize speed of guidance to likely target locations, while simultaneously being sensitive to unexpected stimulus changes.

In all methods discussed so far, the attention model operates as a front-end to an existing recognition one. Nevertheless, Rothenstein and Tsotsos (2008) claim, based on experimental evidence, that attention and recognition should be interdependent in a bidirectional feedback process which results both in the detection and recognition of the object. An early example of such a model is the one proposed by Rybak et al. (1998), where the set of edges extracted at each fixation provides potential targets for the next gaze fixation. Recognition may then be achieved hierarchically and even from a part of the image (from a fraction of the scan-path belonging to this part) when the image is partly perturbed or the object is occluded. Obviously, the stability of recognition increases with the number of fixations.

11.3.2 Novelty Detection and Video Summarization

Attention selection and saliency computation methods are often implicitly used in common computer vision tasks like novelty detection or visual summarization of an image sequence. Generally, unusual activities like the rarity of an event, the sudden appearance/disappearance, the abrupt behavioral change of an object, etc. are defined as salient in various applications. 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 predefining a particular set of activities as being usual, model it in some way, and then detect whether an observed activity is anomalous (Boiman and Irani 2005; Stauffer and Grimson 2000). 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. Measures on a prototype-segment cooccurrence matrix reveal unusual formations that correspond to rare events. Using a 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 (Adam et al. 2008). Intuitively similar are the methods proposed in Hamid et al. (2005), Itti and Baldi (2005), and Zhong et al. (2004).

Going one step further toward human perception of saliency, Ma et al. (2002) 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. Rapantzikos et al. (Evangelopoulos et al. 2008, 2009) build further on visual, audio, and textual attention models for visual summarization. The authors form a multimodal saliency curve integrating the aural, visual, and textual streams of videos based on efficient audio, image, and language processing and employ it as a metric for video event detection and abstraction. The proposed video summarization algorithm is based on the fusion of the three streams and the detection of salient video segments. The algorithm is generic and independent of the video semantics, syntax, structure, or genre. Subjective evaluation (user attention curves as ground-truth) showed that informative and pleasing video skims can be obtained using such multimodal saliency indicator functions.

11.3.3 Robotic Vision

If vision is important for humans, the same holds for robots. Robotic systems of different categories are equipped with visual sensory to interact with the environment and achieve predefined goals. Naturally, such robotic systems benefit from the incorporation of visual attention mechanisms, since time and efficiency are important in the field.

Attention mechanisms prove beneficial for significant robotic applications like *Simultaneous Localization and Mapping*, where an unknown environment should be mapped and the location of the robot should be efficiently detected. Visual attention is used to limit the amount of visual input and focus on the most informative parts of the scene. VOCUS (Visual Object detection with a Computational attention System) (Frintrop et al. 2007; Siagian and Itti 2007) is a successful example that includes a bottom-up and top-down process, while incorporating further sensory input (laser, etc.) (Frintrop et al. 2005). The experimental evaluation of the system shows the robot's behavioral improvement when using the attention mechanism.

The CODAM model has been recently used as part of an autonomous cognitive system called GNOSYS (Taylor et al. 2009). The basic platform is a Pioneer P3AT robot with a Katana arm and the visual stimuli consist of sets of cylinders and sticks as well as a ball and a cube. The robot had to accomplish several simple and complex tasks related to the objects (e.g., find and grasp and stack the cylinders). Many useful conclusions were drawn that highlight the crucial role of perception, as modeled by a group of modules including CODAM, in performing tasks. More details are given in Taylor et al. (2009).

11.4 Volumetric Saliency by Feature Competition

Apparently computational models, which are either based on strict or relaxed correspondences to well-founded neurophysiological counterparts of the visual attention system, have been proposed and applied to several computer vision fields. Saliency in a natural image sequence can be computed either in a frame-by-frame basis using either of the above models or by incorporating the temporal dimension to the model and let it interact tightly with the spatial one. Computational modeling of spatiotemporal saliency has recently become an interesting topic in the literature, since most video analysis algorithms may benefit from the detection of salient spatiotemporal events (e.g., events characterized by strong variations of the data in both the spatial and temporal dimensions).

Most models are inspired by biological mechanisms of motion-based perceptual grouping and compute saliency by extending operators previously proposed for static imagery. Specifically, we extend the spatial center-surround operator of Itti et al. in a straightforward manner by using volumetric neighborhoods in a spatiotemporal Gaussian scale-space (Rapantzikos et al. 2007). In such a framework, a video sequence is treated as a volume, which is created by stacking temporally consequent video frames. Hence, a moving object in such a volume is perceived as occupying a spatiotemporal area. Evaluation is performed for a surveillance application using a public available dataset. 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 3D segmentation, where the notion of "object tunnel" is used to describe the volume carved out by a moving object in this volume (Ristivojević and Konrad 2006). Okamoto et al. used a similar volumetric framework for video clustering, where video shots are selected based on their spatiotemporal texture homogeneity (Okamoto et al. 2002). Quite recently, Mahadevan and Vasconcelos (2009) proposed a biologically plausible model for spatiotemporal saliency. Under their formulation, the saliency of a location is equated to the power of a predefined set of features to discriminate between the visual stimuli in a center and a surround window, centered at that location. The features are spatiotemporal video patches and are modeled as dynamic textures to achieve a principled joint characterization of the spatial and temporal components of saliency. Their model is evaluated on a background subtraction task and compares well with the competition. Nevertheless, as the authors claim, computationally efficiency is an issue.

Most existing models do not count in efficiently the competition among different features, which according to experimental evidence has its biological counterpart in the HVS (Kandel et al. 1995) (interaction/competition among the different visual pathways related to motion/depth (M pathway) and gestalt/depth/color (P pathway), respectively). The early work of Milanese et al. (1995), one of the first computational visual saliency models, was based on such a competition that was implemented in the model as a series of continuous inter- and intrapixel differences. Toward this direction we implement competition by a constrained minimization approach that involves interscale, intrafeature, and interfeature constraints

(Rapantzikos et al. 2009). A centralized saliency map along with an inherent feature competition scheme is to provide a computational solution to the problem of region-of-interest (ROI) detection/selection in video sequences. As stated before, 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 uses an optimization/competition procedure with constraints coming both from inter-, intrafeature, and interscale level. Our method is based on optimization, though in this case competition is performed across features, scales, and voxels in our volumetric representation. The competition is implemented through a constrained minimization method with the constraints being inspired by the Gestalt laws.

Gestalt theory 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 (Koffka 1935; Wertheimer 1923). 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 the 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.

11.5 Problem Formulation

Saliency computation in video is a problem of assigning a measure of interest to each spatiotemporal visual unit. We propose a volumetric representation of the visual input where features interact and compete to attain a saliency measure. Figure 11.1 depicts a schematic diagram of the method. The input is a sequence of frames represented in our model as a volume in space-time. This volume is decomposed into a set of conspicuity features, each decomposed into multiple scales.

A three-way interaction scheme is implemented, which allows voxel competition in the following ways: (a) intrafeature (proximity), between voxels of the same feature and same scale; (b) interscale (scale), between voxels of the same feature but different scale; and (c) interfeature (similarity), between voxels of different features. The stable solution of the energy minimization leads to the final saliency volume.

Let V be a volume representing a set of consequent input 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

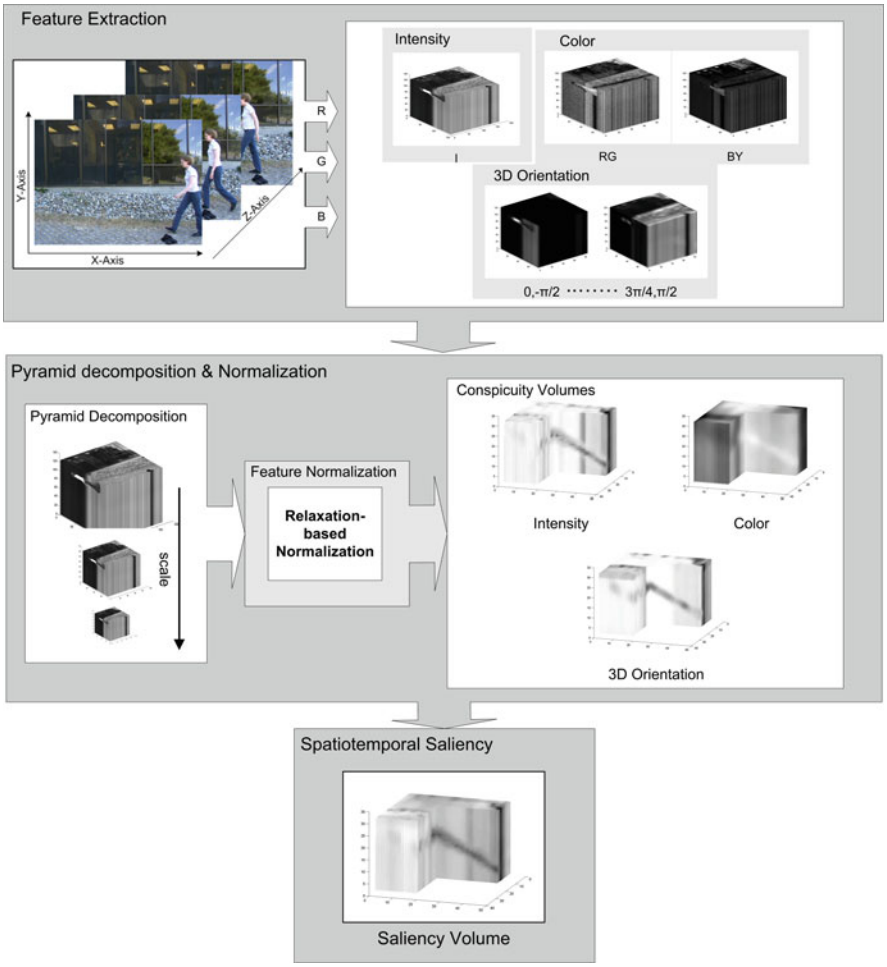

Fig. 11.1 Schematic diagram of the proposed method

coordinates. Under this representation point q becomes the equivalent to a voxel in this volume. Let $V(q)$ be the value of volume V at point q.

V is decomposed into a set of conspicuity volumes C_i with $i = 1, \ldots, M$ corresponding to three different features, namely intensity, color, and motion. Intensity and color features are based on color opponent theory, and spatiotemporal orientation (motion) is computed using 3D steerable filters. Each conspicuity volume is further decomposed into multiple scales ℓ and a set $\mathbf{C} = C_{i\ell}$ is created with $i = 1, \ldots, M$ and $\ell = 1, \ldots, L$ representing a Gaussian volume pyramid.

The final saliency distribution is obtained by minimizing an energy function E composed of a data term E_D and a smoothness term E_S:

$$E(\mathbf{C}) = \lambda_D \cdot E_D(\mathrm{C}) + \lambda_S \cdot E_S(\mathbf{C}). \quad (11.1)$$

The data term models the interaction between the observation and the current solution, while the smoothness term is composed of the three following constraints each related to a different saliency dimension.

E_1 models intrafeature coherency, i.e., defines the interaction among neighboring voxels of the same feature at the same scale and enhances voxels that are incoherent with their neighborhood:

$$E_1(\mathbf{C}) = \sum_i \sum_f \sum_q \left(C_{i,\ell}(q) - \frac{1}{|N_q|} \sum_{r \in N_q} C_{i,\ell}(r) \right)^2 . \tag{11.2}$$

E_1 produces small spatiotemporal blobs of similar valued voxels.

E_2 models interfeature coherency, i.e., it enables interaction 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:

$$E_2(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left(C_{i,\ell}(q) - \frac{1}{M-1} \sum_{j \neq i} C_{j,\ell}(q) \right)^2 . \tag{11.3}$$

E_3 models interscale coherency among ever coarser resolutions of the input, i.e., aims to enhance voxels that are conspicuous across different pyramid scales. If a voxel retains a high value along all scales, then it should become more salient.

$$E_3(\mathbf{C}) = \sum_i \sum_\ell \sum_q \left(C_{i,\ell}(q) - \frac{1}{L-1} \sum_{n \neq \ell} C_{i,n}(q) \right)^2 . \tag{11.4}$$

To minimize (11.1), we adopt a steepest gradient descent algorithm. Detailed information is given in Rapantzikos et al. (2009). Visual examples of saliency volumes are given throughout this chapter (Figs. 11.2 and 11.6b).

11.6 Saliency-Based Video Classification

We choose video classification as a target application to obtain objective, numerical evaluation of the proposed model. The experiment involves multiclass 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 motivation is that if classification based on features from salient regions is improved

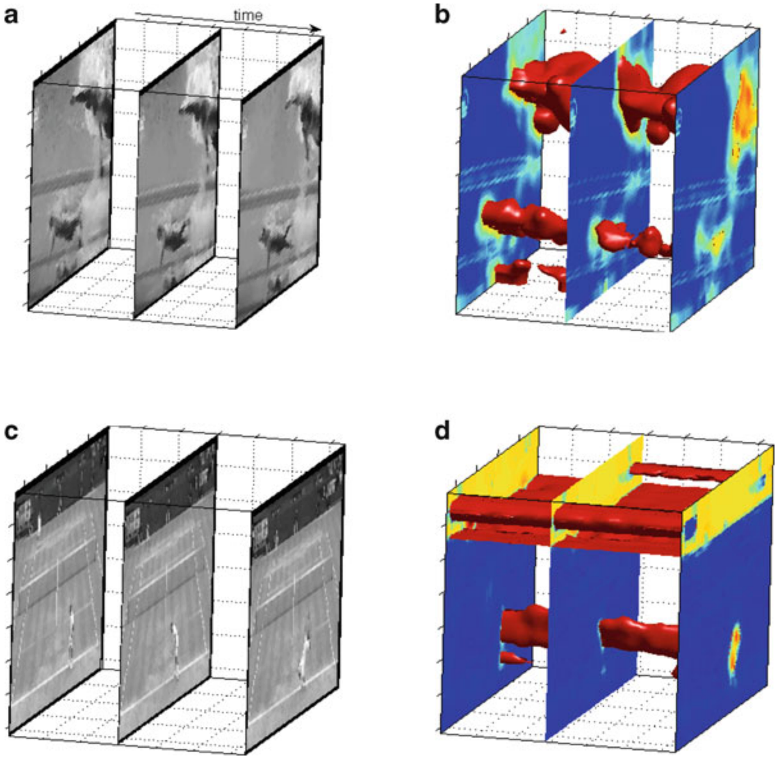

Fig. 11.2 (**a**) (**c**) Examples of slices from the original volume and the corresponding slices from the computed saliency volume. (**b**) (**d**) Corresponding isosurfaces, where the red color corresponds to high values. The salient regions become evident (better viewed in color)

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.

We evaluate the performance of the spatiotemporal saliency method by setting up a multiclass video classification experiment and observing the classification error's increase/decrease when compared against other techniques. Input data consists of several sports clips, which are collected and manually annotated by the authors (Rapantzikos et al. 2009). 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.). In order 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 Support Vector Machines (SVMs) for classifying the data. We train the classifiers using a radial basis function (RBF) kernel after appropriately selecting a model using fivefold cross-validation estimation of the multiclass 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 parameters.

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 precalculated features are extracted for each of them. Feature vectors feed an SVM classifier, and the outputs of all methods are compared.

11.7 Evaluation of Classification Performance

In order 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 (Oliver et al. 2000).

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 (Rapantzikos and Avrithis 2005). The two other saliency-based methods are the state-of-the art static saliency-based approach of Itti et al. (1998) and an extension using a motion map (Rapantzikos and Tsapatsoulis 2005). Both methods produce a saliency measure per pixel. The static saliency-based approach processes the videos in a per frame basis. After producing a saliency map for each frame, we generate a saliency volume by stacking them together. We filter this volume with a 3D median filter to improve temporal coherency. The motion map of the extended approach is derived using a motion estimation technique, which is based on robust statistics (Black and Anandan 1996). The same procedure for producing a saliency volume is followed for the PCA-based technique. For the sake of completeness, we also provide results

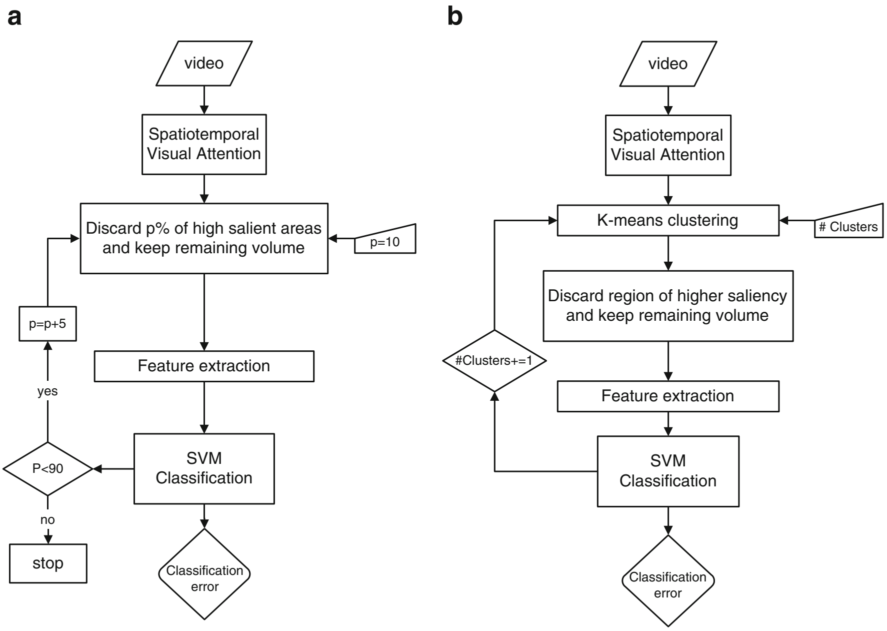

Fig. 11.3 Saliency-based classification: (**a**) based on foreground/background detection; (**b**) based on >1 salient regions

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. At each iteration, the initial video volume is incrementally reduced by $p\%$ and a classification error is produced. The reduction is done spatially in a 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.

We 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. 11.3a and is composed of three steps: (1) discard voxels of high saliency until a volume of $p\%$ of the whole volume remains; (2) extract histograms of precalculated 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 automatic thresholding 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 nonsalient 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.

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 not all of them are required in order to 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. 11.3b. 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 a predetermined number of regions using unsupervised clustering. In this framework, we use a clustering technique that allows for nonhard 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 per color channel (i.e., 24 elements per region) and 4 bins for motion/2D-orientation are used.

Figure 11.4 shows the classification error along with standard error intervals when varying the size of the discarded region. The plot shows the results of evaluated methods. Each point on the graphs should be interpreted as follows: we discard, e.g., 10% of the high salient pixels (x-axis) and obtain a classification error (y-axis) using fivefold cross validation (standard error interval for each point). In case of the heuristic method, the ratio represents the portion of the discarded regions starting from the borders. Classification driven by spatiotemporal saliency provides improved statistics over the other methods, since for all measured values of sensitivity it achieves lower false positive rates. Although differences in magnitude are not tremendous, two facts become evident: first, salient-based approaches seem to provide more consistent and overall better results than the nonsalient one; second, the proposed model compares well against the other three-commonly established-approaches.

The second experiment illustrates the effect on classification performance when using features bound to more than one salient region. This experiment corresponds to the flow diagram shown in Fig. 11.3b. Figure 11.5 shows the obtained classification error vs. the number of segmented regions. The proposed and the PCA-based techniques perform overall better than the rest with the first having lower fluctuations and achieving the lowest average error.

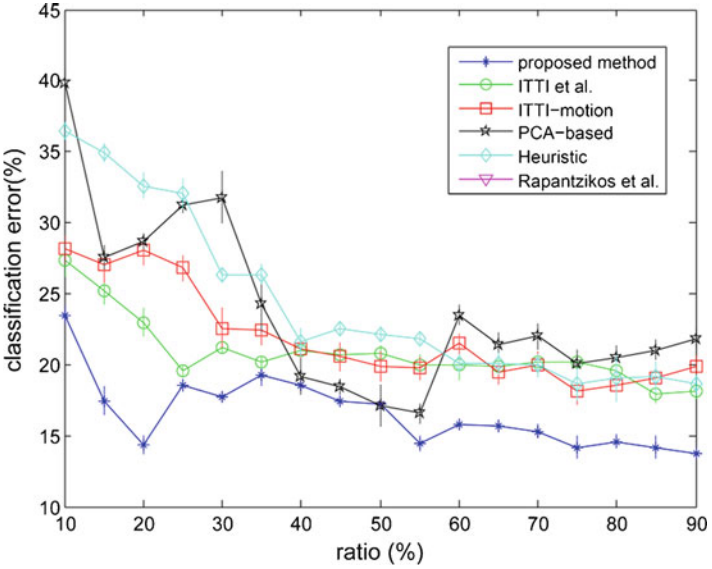

Fig. 11.4 Experiment I – Classification error along with standard error intervals for all tested methods when varying the size of the discarded region (ratio represents the percent of discarded high saliency voxels)

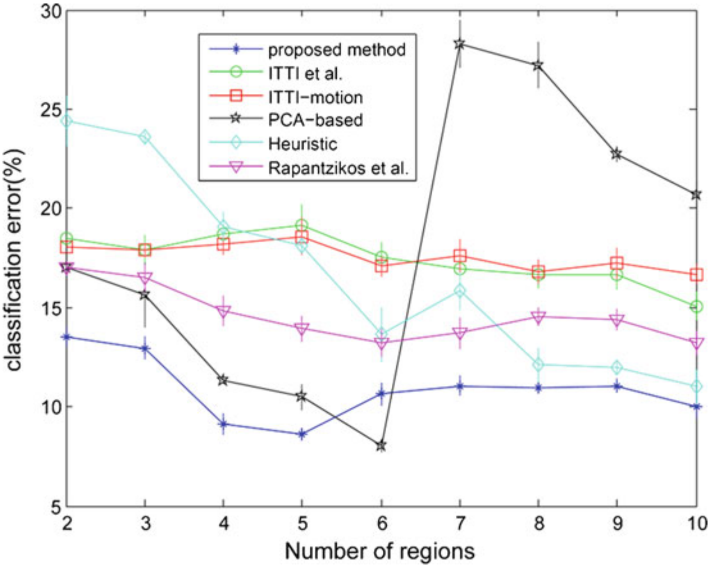

Fig. 11.5 Classification error along with standard error intervals when varying the number of regions (error vs. number of regions used to segment the volumes)

11.8 Action Recognition

Although not always explicitly said, an important computer vision field, namely interest point detection, is also based on a saliency-related representation of the input. For example, the famous point detector of Harris et al. detects points as local maxima of a distribution that measures the image intensity change in multiple directions (Harris and Stephens 1988). Similarly, the Hessian-affine detectors involving computation of second derivatives give strong (salient) response on blobs and ridges (Lindeberg 1998). 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 (Kadir and Brady 2001). A review and a well-grounded comparison of spatial detectors is given by Mikolajczyk et al. (2005). Recently, points of interest combined with bag-of-words approaches have been also used for object/event detection and recognition. Such methods represent the visual input by a set of small visual regions (words) extracted around salient points like the ones referred before. Csurka et al. use visual words around regions extracted with the Harris affine detector (Mikolajczyk and Schmid 2002) to represent and detect visual objects (Csurka et al. 2004). 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 (Wang et al. 2006). 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. also propose a bag-of-words based method to detect multiple object categories in images (Bosch et al. 2006). 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 (2003) and further developed and used in Laptev et al. (2007). A spatio-temporal 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. Dollár et al. proposed a framework, which is based on a bag-of-words approach, where a visual word is meant as a cuboid (Dollár et al. 2005). They developed an extension of a periodic point detector to the spatio-temporal 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. use the same periodic detector and propose an unsupervised learning method for human action categories (Niebles et al. 2006). 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.

11.9 Spatiotemporal Point Detection

Salient points are extracted as the local maxima of the spatiotemporal saliency distribution obtained after minimization of (11.1). Such points are located at regions that exhibit high compactness (proximity), remain intact across scales (scale), and pop-out from their surroundings due to feature conspicuity (similarity). Hence we expect that the points will not be only located around spatiotemporal corners, but also around smoother space-time areas with distinguishing characteristics that are often important for action recognition. Figure 11.6 shows an example of such points.

We evaluate the proposed model by setting up experiments in the action recognition domain using two action datasets, namely the KTH dataset[1] and the Hollywood Human Actions (HOHA) one.[2] Both are public and available online. We provide a qualitative evaluation of the proposed detector, short descriptions of the datasets, and the corresponding recognition frameworks and devote the rest of the section to quantitative analysis. For comparison purposes we use two state-of-the-art detectors, namely the periodic one proposed by Dollár et al. (2005) and the space-time point detector of Laptev and Lindeberg (2003), which are publicly available. In the following, we will denote the first one by "periodic" and the second one by "stHarris" (Fig. 11.7).

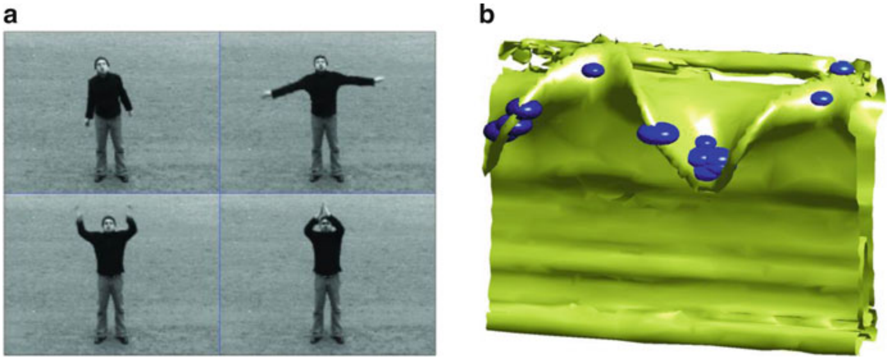

Fig. 11.6 (**a**) Indicative slices of a handwaving sequence and an ISO surface with the detected points overlaid for the (**b**) proposed

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

[2] http://www.irisa.fr/vista/Equipe/People/Laptev/download.html

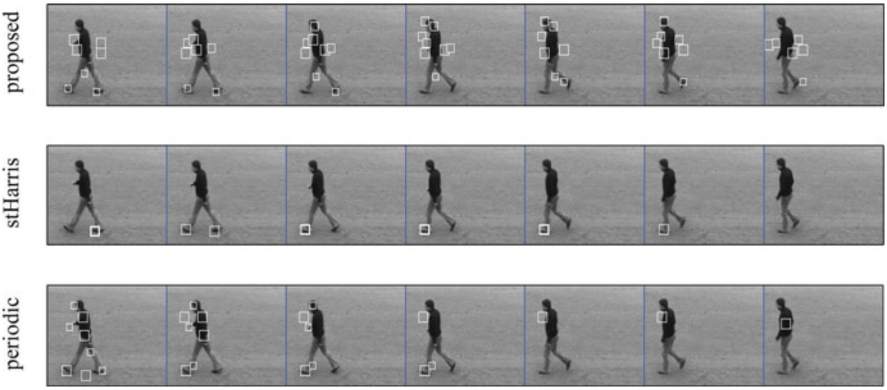

Fig. 11.7 Cuboids detected on a walking sequence from KTH dataset

Overall, saliency is obtained as the solution of an energy minimization problem that is initiated by a set of volumetric feature conspicuities derived from intensity, color, and motion. The energy is constrained by terms related to spatiotemporal proximity, scale, and similarity, and feature points are detected at the extrema of the saliency response. Background noise is automatically suppressed due to the global optimization framework, and therefore the detected points are dense enough to represent well the underlying actions. We demonstrate these properties in action recognition using two diverse datasets. The results reveal behavioral details of the proposed method and provide a rigorous analysis of the advantages and disadvantages of all methods involved in the comparisons. Our detector performs quite well in all experiments and either outperforms the state-of-the-art techniques it is compared to or performs among the top of them depending on the adopted recognition framework. More details can be found in Rapantzikos et al. (2009).

11.10 Discussion

Saliency-based image and video processing contribute in several aspects to solve common computer vision problems. The proposed volumetric saliency-based model for saliency computation exploits the spatiotemporal structure of a video sequence and produces a per voxel saliency measure based on a feature competition approach. This measure provides evidence about important and nonimportant regions in the sequence. Saliency is obtained as the solution of an energy minimization problem that is initiated by a set of volumetric feature conspicuities derived from intensity, color, and motion. The energy is constrained by terms related to spatiotemporal proximity, scale, and similarity. Background noise is automatically suppressed due to the global optimization framework, and therefore the salient regions or points represent well the underlying sequence/actions. The experimental results reveal behavioral details of the proposed method and provide a rigorous analysis of the advantages and disadvantages of all methods involved in the comparisons. In the future, motivated by recent works, we will focus on computational efficiency issues and more applications in the field. Furthermore, it would be interesting to explore top-down approaches in order to guide our saliency computation framework toward desired goals.

References

Abrams, R.A., Christ, S.E., "Motion onset captures attention", Psychological Science, vol. 14, pp. 427–432, 2003.

Adam, A., Rivlin, E., Shimshoni, I., Reinitz, D., "Robust Real-Time Unusual Event Detection using Multiple Fixed-Location Monitors", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 30, no. 3, pp. 555–560, Mar 2008.

Black, M.J., Anandan, P., "The Robust Estimation of Multiple Motions: Parametric and Piecewise-Smooth Flow Fields", CVIU, vol. 63, no. 1, pp. 75–104, 1996.

Boiman, O., Irani, M., "Detecting Irregularities in Images and in Video", IEEE International Conference on Computer Vision (ICCV), Beijing, 2005.

Bosch, A., Zisserman, A., Munoz, X., "Scene Classification via pLSA", ECCV06, pp. 517–530, 2006.

Bruce, N.D.B., Tsotsos, J.K., Saliency, attention, and visual search: An information theoretic approach. Journal of Vision, vol. 9, no. 3, pp. 1–24, 2009.

Bruce, N., Tsotsos, J., "Saliency based on information maximization", Advances in Neural Information Processing Systems, vol. 18, pp. 155–162, 2006.

Csurka, G., Bray, C., Dance, C., Fan, L., "Visual categorization with bags of key-points", pp. 1–22, Workshop on Statistical Learning in Computer Vision, ECCV, 2004.

Cutsuridis, V., "A Cognitive Model of Saliency, Attention, and Picture Scanning", Cognitive Computation, vol. 1, no. 4, pp. 292–299, Sep. 2009.

Dollár, P., Rabaud, V., Cottrell, G., Belongie, S., "Behavior Recognition via Sparse Spatio-Temporal Features", VS-PETS, pp. 65–72, Oct 2005.

Duncan, J., "Selective attention and the organization of visual information", Journal of Experimental Psychology: General, vol. 113, no. 4, pp. 501–517, 1984.

Evangelopoulos, G., Rapantzikos, K., Potamianos, A., Maragos, P., Zlatintsi, A., Avrithis, Y., "Movie Summarization Based On Audio-Visual Saliency Detection", Proceedings International Conference on Image Processing (ICIP), San Diego, California, 2008.

Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., Maragos, P., Avrithis, Y., "Video event detection and summarization using audio, visual and text saliency", IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3553–3556, 2009.

Frintrop, S., Cremers, A., "Top-down attention supports visual loop closing", In. Proceedings Of European Conference On Mobile Robotics (ECMR'05), 2007.

Frintrop, S., Backer, G., Rome, E., "Goal directed search with a top-down modulated computational attention system", LCNS, vol. 3663, no. 117, 2005.

Frintrop, S., Rome, E., Nuchter, A., Surmann, H., "A bimodal laser-based attention system", Computer Vision and Image Understanding, vol. 100, no. 1–2, pp. 124–151, 2005.

Frintrop, S., Klodt, M., Rome, E., "A real-time visual attention system using integral images", In Proceedings Of the 5th International Conference on Computer Vision systems, ICVS, 2007.

Hamid, R., Johnson, A., Batta, S., Bobick, A., Isbell, C., Coleman, G., "Detection and explanation of anomalous activities: representing activities as bags of event n-grams", CVPR'05, vol. 1, pp. 1031–1038, Jun 2005.

Harris, C., Stephens, M., “A combined corner and edge detector”, Alvey Vision Conference, pp. 147–152, 1988.

Itti, L., Baldi, P., “A Principled Approach to Detecting Surprising Events in Video”, CVPR’05, 2005, vol. 1, pp. 631–637, 2005.

Itti, L., Baldi, P., “Bayesian surprise attracts human attention”, Vision Research, vol. 49, no. 10, pp. 1295–1306, 2009.

Itti, L., Koch, C., Niebur, E., “A model of saliency-based visual attention for rapid scene analysis”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

James, W., “The principles of psychology”, Cambridge, MAL Harvard UP, 1890/1981.

Kadir, T., Brady, M., Saliency, scale and image description, International Journal of Computer Vision, vol. 45, no. 2, pp. 83–105, 2001.

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

Koch, C., Ullman, S., “Shifts in selective visual attention: towards the underlying neural circuitry”, Human Neurobiology, vol. 4, no. 4, pp. 219–227, 1985.

Koffka, K., Principles of Gestalt Psychology, Harcourt, New York, 1935.

Laptev, I., Lindeberg, T., “Space-Time Interest Points”, in Proceedings of the ICCV’03, Nice, France, pp. 432–443, 2003.

Laptev, I., Caputo, B., Schuldt, C., Lindeberg, T., “Local Velocity-Adapted Motion Events for Spatio-Temporal Recognition”, Computer Vision and Image Understanding, vol. 108, pp. 207–229, 2007.

Lee, K., Buxton, H., Feng, J., “Cue-guided search: A computational model of selective attention”, IEEE Transactions On Neural Networks, vol. 16, no. 4, pp. 910–924, 2005.

Leventhal, A., “The neural basis of visual function: vision and visual dysfunction”, Nature Neuroscience, vol. 4, 1991.

Lindeberg, T., “Feature detection with automatic scale selection”, International Journal of Computer Vision, vol. 30, no. 2, pp. 79–116, 1998.

Lowe, D., “Object recognition from local scale-invariant features”, In Proceedings of ICCV, pp. 1150–1157, 1999.

Mahadevan, V., Vasconcelos, N., “Spatiotemporal Saliency in Dynamic Scenes”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009.

Ma, Y.F., Lu, L. Zhang, H.J., Li, M., “A user attention model for video summarization”, ACM Multimedia Conference, pp. 533–542, 2002.

May, Y., Zhang, H., “Contrast-based image attention analysis by using fuzzy growing”, In Proceedings ACM International Conference on Multimedia, pp. 374–381, 2003.

Mikolajczyk, K., Schmid, C., “An affine invariant interest point detector”, ECCV, pp. 128–142, 2002.

Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L., “A comparison of affine region detectors”, International Journal of Computer Vision, vol. 65, no. 1/2, pp. 43–72, 2005.

Milanese, R., Gil, S., Pun, T., “Attentive mechanisms for dynamic and static scene analysis”, Optical Engineering, vol. 34 no. 8, pp. 2428–2434, 1995.

Navalpakkam, V., Itti, L., “An integrated model of top-down and bottom-up attention for optimal object detection”, Computer Vision and Pattern Recognition (CVPR), pp. 1–7, 2006.

Navalpakkam, V., Itti, L., “Modeling the influence of task on attention”, Vision Research, vol. 45, no. 2, pp. 205–231, 2005.

Niebles, J.C., Wang, H., Fei-Fei, L., “Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words”, British Machine Vision Conference (BMVC), Edinburgh, 2006.

Okamoto, H., Yasugi, Y., Babaguchi, N., Kitahashi, T., “Video clustering using spatiotemporal image with fixed length”, ICME’02, pp. 2002–2008, 2002.

Oliver, N.M., Rosario, B., Pentland, A.P., “A Bayesian computer vision system for modeling human interactions”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, Aug 2000.

Park, S., Shin, J., Lee, M., “Biologically inspired saliency map model for bottom-up visual attention”, Lectrure Notes in Computer Science, pp. 418–426, 2002.

Rapantzikos, K., Avrithis, Y., “An enhanced spatiotemporal visual attention model for sports video analysis”, International Workshop on Content-based Multimedia indexing (CBMI’05), Riga, Latvia, Jun 2005.

Rapantzikos, K., Tsapatsoulis, N., “Enhancing the robustness of skin-based face detection schemes through a visual attention architecture”, Proceedings of the IEEE International Conference on Image Processing (ICIP), Genova, Italy, vol. 2, pp. 1298–1301, 2005.

Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y., Kollias, S., “A Bottom-Up Spatiotemporal Visual Attention Model for Video Analysis”, IET Image Processing, vol. 1, no. 2, pp. 237–248, 2007.

Rapantzikos, K., Avrithis, Y., Kollias, S., “Dense saliency-based spatiotemporal feature points for action recognition”, Conference on Computer Vision and Pattern Recognition (CVPR), 2009.

Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y., Kollias, S., “Spatiotemporal saliency for video classification”, Signal Processing: Image Communication, vol. 24, no. 7, pp. 557–571, 2009.

Rensink, R.A., “Seeing, sensing, and scrutinizing”, Vision Research, vol. 40, no. 10–12, pp. 1469–1487, 2000.

Ristivojević, M., Konrad, J., “Space-time image sequence analysis: object tunnels and occlusion volumes”, IEEE Transactions Of Image Processings, vol. 15, pp. 364–376, Feb. 2006.

Rothenstein, A., Tsotsos, J., “Attention links sensing to recognition”, Image and Vision Computing, vol. 26, no. 1, pp. 114–126, 2008.

Rutishauer, U. Walther, D., Koch, C., Perona, P., “Is bottom-up attention useful for object recognition?”, Computer Vision and Pattern Recognition (CVPR), vol. 2, 2004.

Rybak, I., Gusakova, V., Golovan, A., Podladchikova, L., Shevtsova, N., “A model of attention-guided visual perception and recognition”, Vision Research, vol. 38, no. 15, pp. 2387–2400, 1998.

Shao, L., Kadir, T., Brady, M., “Geometric and photometric invariant distinctive regions detection”, Information Sciences 177, vol. 4, pp. 1088–1122, 2007.

Siagian, C., Itti, L., “Biologically inspired robotics vision monte-carlo localization in the outdoor environment, In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2007.

Sillito, A., Jones, H., “Context-dependent interactions and visual processing in V1”, Journal of Physiology-Paris, vol. 90, no. 3–4, pp. 205–209, 1996.

Stauffer, C., Grimson, E., “Learning Patterns of Activity Using Real-Time Tracking”, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 8, pp. 747–757, Aug 2000.

Sternberg, R., “Cognitive Psychology,” Wadsworth Publishing, 2006.

Sun, Y., Fisher, R., “Object-based visual attention for computer vision”, Artificial Intelligence, vol. 146, no. 1, pp. 77–123, 2003.

Taylor, J.G., “Attentional movement: the control basis for consciousness”, Society for Neuroscience Abstracts, vol. 26, no. 2231, 2000.

Taylor, J.G., “CODAM: A neural network model of consciousness”, Neural Networks, vol. 20, no. 9, pp. 983–992, Nov 2007.

Taylor, J.G., “On the neurodoynamics of the creation of consciousness”, Cognitive Neurodynamics, vol. 1, no. 2, Jun 2007.

Taylor, J.G., “Paying attention to consciousness”, Progress in Neurobiology, vol. 71, pp. 305–335, 2003.

Taylor, J.G., Hartley, M., Taylor, N., Panchev, C., Kasderidis, S., “A hierarchical attention-based neural network architecture, based on human brain guidance, for perception, conceptualisation, action and reasoning”, Image and Vision Computing, vol. 27, no. 11, pp. 1641–1657, 2009.

Torralba, A., “Contextual priming for object detection”, International Journal of Computer Vision, vol. 53, no. 2, pp. 169–191, 2003.

Torralba, A., Oliva, A., Castelahno, M., Henderson, J., “Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search”, Psychological Review, vol. 113, no. 4, pp. 766–786, 2006.

Treisman, A.M., Gelade, G., “A feature integration theory of attention”, Cognitive Psychology, vol. 12, no. 1, pp. 97–136, 1980.

Tsotsos, J.K., Culhane, S.M., Wai, W.Y.K., Lai, Y., Davis, N., Nuflo, F., “Modelling visual attention via selective tuning”, Artifficial Intelligence, vol. 78, pp. 507–545, 1995.

Walther, D., Koch, C., “Modelling attention to salient proto-objects”, Neural Networks, vol. 19, no. 9, pp. 1395–1407, 2006.

Walther, D., Rutishauer, U., Koch, C., Perona, P., “On the uselfuness of attention for object recognition”, In Workshop of Attention for Object Recognition at ECCV, pp. 96–103, 2004.

Walther, D., Rutishauer, U., Koch, C., Perona, P., “Selective visual attention enables learning and recognition of multiple objects in cluttered scenes, Computer Vision and Image Understanding (CVIU), vol. 100, no. 1–2, pp. 41–63, 2005.

Wang, Y., Jiang, H., Drew, M.S., Li, Z., Mori, G., “Unsupervised Discovery of Action Classes”. In Proceedings of CVPR’06, vol. 2, pp. 17–22, 2006.

Wertheimer, M., “Laws of Organization in Perceptual Forms”, First published as “Untersuchungen zur Lehre von der Gestalt II, in Psycologische Forschung, vol. 4, pp. 301–350, 1923.

Wolfe, J.M., “Guided search 2.0: A revised model of visual search”, Psychonomic Bulletin & Review 1, vol. 2, pp. 202–238, 1994.

Wolfe, J.M., “Guided search 4.0: current progress with a model of visual search”, Integrated Models of Cognitive Systems, pp. 99–119, 2007.

Wolfe, J.M., Cave, K.R., Franzel, S.L., “Guided search: an alternative to the feature integration model for visual search”, Journal of Experimental Psychology: Human Perception and Performance, vol. 15, no. 3, pp. 419–433, 1989.

Zhong, H., Shi, J., Visontai, M., “Detecting Unusual Activity in Video”, CVPR’04, Washington, DC, vol. 2, pp. 819–826, Jun 2004.