Cogn Comput (2011) 3:241–263
DOI 10.1007/s12559-010-9094-8

A Spatiotemporal Saliency Model for Video Surveillance

Tong Yubing · Faouzi Alaya Cheikh · Fahad Fazal Elahi Guraya · Hubert Konik · Alain Trémeau

Received: 22 March 2010 / Accepted: 27 December 2010 / Published online: 8 January 2011

Abstract A video sequence is more than a sequence of still images. It contains a strong spatial–temporal correlation between the regions of consecutive frames. The most important characteristic of videos is the perceived motion foreground objects across the frames. The motion of foreground objects dramatically changes the importance of the objects in a scene and leads to a different saliency map of the frame representing the scene. This makes the saliency analysis of videos much more complicated than that of still images. In this paper, we investigate saliency in video sequences and propose a novel spatiotemporal saliency model devoted for video surveillance applications. Compared to classical saliency models based on still images, such as Itti's model, and space–time saliency models, the proposed model is more correlated to visual saliency perception of surveillance videos. Both bottom-up and top-down attention mechanisms are involved in this model. Stationary saliency and motion saliency are, respectively, analyzed. First, a new method for background subtraction and foreground extraction is developed based on content analysis of the scene in the domain of video surveillance. Then, a stationary saliency model is setup based on multiple features computed from the foreground. Every feature is analyzed with a multi-scale Gaussian pyramid, and all the features conspicuity maps are combined using different weights. The stationary model integrates faces as a supplement feature to other low level features such as color, intensity and orientation. Second, a motion saliency map is calculated using the statistics of the motion vectors field. Third, both motion saliency map and stationary saliency map are merged based on center-surround framework defined by an approximated Gaussian function. The video saliency maps computed from our model have been compared to the gaze maps obtained from subjective experiments with SMI eye tracker for surveillance video sequences. The results show strong correlation between the output of the proposed spatiotemporal saliency model and the experimental gaze maps.

Keywords Visual saliency · Motion saliency · Background subtraction · Center-surround saliency · Face detection · Video surveillance

T. Yubing · H. Konik · A. Trémeau (✉)
Laboratoire Hubert Crurien UMR 5516, Université Jean Monnet, 42000 Saint-Etienne, France
e-mail: alain.tremeau@univ-st-etienne.fr

F. A. Cheikh · F. F. E. Guraya
Faculty of Computer Science and Media Technology, Gjøvik University College, Gjøvik, Norway

Introduction

Under natural viewing conditions, humans tend to focus on specific parts of an image or a video which evokes our interests naturally. These regions carry most useful information needed for our interpretation of the scenes. Video contains more information than a single image, and the perception of video is also different from that of single image because of the additional temporal dimension of the sequence. Several saliency models have been proposed in recent years. Itti's model [1] is the most widely used saliency model for stationary image. GAFFE [2], frequency-tuned saliency detection model [3] and the model based on phase spectrum and inverse Fourier transform [4] are other

saliency models for still images. All of them adopted the bottom-up visual attention mechanism. In [3], image saliency map is obtained from the range of frequencies in the image spectrum that represent the important image details. Next, the outputs of several band-pass filters are combined to compute the saliency map via DOG (Difference of Gaussians). Low level image features including intensity, orientation and color or contrast are used to construct feature conspicuity maps, which are then integrated into the final saliency map with WTA (Winner Take All) and IOR (Inhibition of Return) principles inspired from the visual nervous system [1, 2]. Besides the above low level features, face, text and other features have also been considered for saliency analysis [5, 6]. All of them are designed for the saliency analysis of stationary image instead of video. The perception of visual saliency in video is much different from that in still images. For example, the texture feature of an object can be salient in a still image meanwhile may not be perceived when the object moves fast in a video. So the above stationary saliency model is not necessarily relevant to characterize the saliency in a video.

Usually, videos are viewed as frames sequence, with a certain frames rate used to render the video with natural/smooth motion. Through video display, we can get a clear perception of the real scene with some factors such as who, where, what [7]. Video saliency involves more information than that can be found in still images and is more complicated than stationary image saliency. Meanwhile, many papers have contributed to static saliency detection fewer papers purely dealt with spatiotemporal saliency. Many papers devoted to video saliency detection are based on the computation of motion saliency map [8–11], other are based on the computation of space–time saliency map. Thus, Marat et al. proposed in [12] a space–time saliency detection algorithm, which fuses static saliency map and dynamic saliency map. Gao et al. proposed in [13] a dynamic texture model in order to capture the motion patterns even in the case that the scene is itself dynamic. Zhang et al. extended in [14] their SUN framework to a dynamic scene by introducing temporal filter (Difference of Exponential:DoE) and fitting a generalized Gaussian distribution to the estimated distribution for each filter response. Compared with other spatiotemporal saliency models, such as the "surprise" model [15], which lack of a sophisticated unified representation for the spatial and temporal components of saliency, the proposed model is based on a unified framework of the spatial and temporal components of saliency. Furthermore, it does not require many design parameters such as the number of filters, type of filters, choice of the non-linearities, proper normalization scheme, nor to learn a visual saliency model directly from human eye-tracking data using a support vector machine (SVM). Lastly, compared to space–time models, the proposed model is not based on the computation of all local region neighborhoods, such as in [11], nor on the computation of local kernels, such as in [16], but on the computing of local motion vectors of foreground objects.

Motion is an important part in videos; however, videos are more than only motion. Both static saliency map and motion saliency map should be considered. Likewise, other information such as distance, depth and spatial position should also be involved. In [8], raw motion map is described using the difference of neighboring images which is a very rough description of motion. For example, some light intensity change might be viewed as motion. In [9], motion saliency is obtained from the module of motion vector derived from optic flow equation. The magnitude and angle of the motion vectors are two important parameters, but also the direction of the motion. This latter is overlooked in [9]. In [10], a motion attention model is proposed based on motion intensity, spatial coherence inductor and temporal coherence inductor. As for the model proposed in [8], in the latter model some fault motions might be detected due to illuminant changes, such as shadows, in local areas of the background instead of real foreground object movement. In [11], the continuous rank on the eigenvalue of coefficient metric derived from neighborhood optical flow equation is viewed as a measurement for motion saliency. But sometimes the optical flow cannot get the accurate motion especially when there is no enough change of gray depth.

Most of the above saliency map methods are based on the bottom-up attention mechanism. Motion feature and other stationary features including color, orientation and intensity are viewed as low level features computed from the bottom. In all these models, every feature is individually analyzed for feature conspicuity and finally combined with different weights. In fact, human perception is more complicated, both bottom-up and top-down framework should be involved. For example, just after looking few frames in a video, a viewer might unconsciously start searching for similar objects in the following successive frames. Meanwhile the bottom-up process is task independent; the top-down process is task-dependent. The top-down process intervenes both in passive and in active viewing such as visual search, object tracking, scene comprehension. [12]. Thus, the analysis of first frames provides unconsciously some video's semantic information, including foreground/background information, to the viewer which is used to predict gaze for the following frames. Moreover, our visual system is able to detect certain objects more easily than others, especially human faces. Indeed, it has been shown that humans are able to process complex images and to recognize familiar objects

very rapidly [13]. Especially, for surveillance videos, the unconscious searching operation is more focused on human shapes than any other object shapes. Lastly, after watching few frames observers deduce certain information about the scene watched such as the presence of moving objects in front of a still background. Therefore, foreground objects detected in previous frames will attract more the attention of observers in the following frames than the background. Any saliency model based on visual perception devoted to video surveillance should consider all these visual phenomena. For this reason, in this paper, we propose to analyze the content of a scene through a background subtraction and foreground objects extraction. As suggested in [17], the related problem of background subtraction is treated here as the complement of saliency detection, by classifying non-salient (with respect to appearance and motion dynamics) point in the visual field as background. The first step of our approach consists to analyze the scene's semantic content through a bottom-up attention process based on the difference between foreground objects and background. The second step consists to compute features saliency map and motion saliency map based on this information (see synopsis shown in Fig. 1). The first contribution of this paper is to propose a new technique based on the partitioning of the scenes in foreground objects and background to analyze the semantic content of surveillance videos. This technique based on a top-down attention process has never been done in any previous research on saliency detection. The second contribution of this paper is to address the video saliency problem through a unified approach combining bottom-up and a top-down attention models. For the former, low level features such as color, intensity and orientation are used, for the latter, face and foreground objects have been considered. Both stationary saliency and motion saliency maps have been considered in our approach. Next saliency maps are merged based on a center-surround framework approximated by a spatial Gaussian distribution.

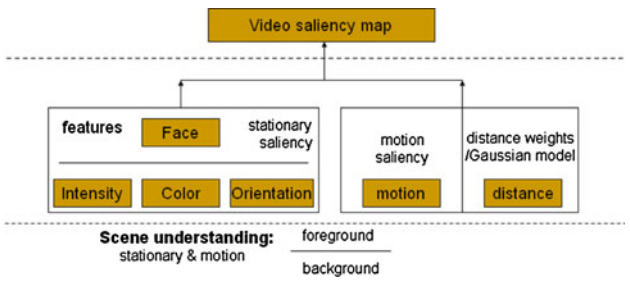

Fig. 1 Synopsis of the spatiotemporal saliency model proposed

The proposed approach is constrained by three assumptions: (a) salient objects are distinct of the background, (b) the number of interesting objects in the scene is limited, and (c) even if the background is not static the information provided by the background is less useful to the observer than foreground moving objects. These assumptions, which are observed especially in surveillance videos in indoor environments, limit the usability of existing methods based on background subtraction [18] but make the foreground object detection easier. Lastly, methods based on background subtraction can be easily extended to any video object detection problem satisfying the same constraints (e.g. an object of interest in a dynamic environment such as a moving car in outdoor environment). In these cases, relevant information can be learned from frames and task contexts in predicting where humans look while performing complex visually guided behavior [12].

The following sections of the paper are arranged as follows: background detection and foreground extraction based on scene understanding are described in "Scene understanding and background extraction in surveillance video". Next, a novel spatiotemporal model for saliency detection is proposed in "Multi-feature model for saliency detection". Lastly, comparative results based on psychophysical experiments and objective metrics are given in "Discussion and Experiments" to evaluate the performance of the proposed method relatively to the performance of Itti's model, frequency-tuned model and phase spectrum model, and GBVS model. Conclusions are given in last section.

Scene Understanding and Background Extraction in Surveillance Video

For scene understanding in videos, three factors are necessarily included, who, where and what. Those factors are usually related to foreground objects, background, motion and events [19]. In video surveillance applications, after a short period of analysis of the semantic content of the video based on an unconscious bottom-up attention process, the observer attention is focused on the moving parts in the foreground. The background becomes useless unless moving objects appear in the background. The analysis of first frames provides to the observer some semantic information on the video, including foreground/background information, which are then used to analysis the following frames. So, if there is no change in the background of the current frame compared to previous frames, then it is not necessary to update the background information as the background of the current frame provides no additional useful information. That is the main reason why

background detection is first processed followed by foreground extraction. This idea was already used in [20, 21].

We have restricted our study to video sequences with static background. This limitation is not very restrictive as different techniques can be used to segment a video into continuous shots, e.g. see [22–24]. For complex dynamic scenes, where local variation in the background (either spatially or temporally) is significant, sophisticated models must be used otherwise this leads to a poor level of performance. The main shortcoming of sophisticated models, such as the DiscSal algorithm proposed in [17], is their computational performance. From the experiments, we conducted, the assumption of a continuous background is valid in the context of video surveillance. In a general way, we consider that changes in background due to photometric effects (e.g. shadows) or slow continuous movements (e.g. camera motion) have little impact on the current frame perception within a video sequence. We consider also that short-term memory has a high impact on the current frame perception meanwhile the impact of previous frames is relatively low [25]. An experiment done for time-varying quality estimation showed that human memory seems to be limited to about 15 s [26].

Many methods have been used for background subtraction. According to different background modeling approaches, these methods can be further classified as parametric and nonparametric methods [17, 20]. For parametric background modeling methods, the most commonly used model is the Mixture of Gaussians (MOG) [27, 28]. Another class of commonly used background modeling methods is based on nonparametric techniques, such as Kernel Density Estimator (KDE) of [29] or the "surprise" model proposed by Itti et al. [15]. Comparing with the parametric background modeling methods, the nonparametric ones have the advantages that they do not need to specify the underlying model and estimate its parameters explicitly [20]. Therefore, they can adapt to arbitrary unknown data distribution. The major drawback of nonparametric methods is their computational cost. The main advantage of nonparametric background techniques is their simplicity [20, 21, 30]. Comparing with background learning techniques (e.g. [27]), which require a training set of "background only" images, the proposed approach does not need a "global background model" or any type of training. Comparing with batch processing techniques (e.g. [31]), which require a large number of video frames, the number of video frames required by the proposed approach is related to the range of variation of the background.

In [21], a sliding window was used to search background pixels frame by frame. The mean shift algorithm was used by Yazhou et al. in [20] to detect background pixels among pixels emerging in video frames. Recently a new algorithm based on quasi-continuous histograms (QCH) had been proposed by Sidibé et al. in [30] to outperform the mean shift algorithm. In the above background extraction methods, searching points are computed in every frame to estimate the background of videos. In the following section, we propose a new scene background extraction algorithm for surveillance videos based on a different searching process. The main idea of this algorithm is to use statistical pixel information to generate background with less searching points.

Background Extraction

In surveillance videos without camera movement, the background is quasi-stable and only foreground objects emerge temporarily in frames [30]. For example, several frames of a surveillance video are shown in Fig. 2. Figure 3 shows the intensity variation at the center of the dark circle shown in the Fig. 2.

Background models try to estimate the most probable intensity and color values for every pixel in a scene. In [20], Yazhou et al. proposed a model based on several features as follows:

$$V_{\text{obsv}} = M_{\text{obj}} + D_{\text{cam}} + C + M_{\text{bgd}} + N_{\text{sys}} + S_{\text{illum}} \quad (1)$$

where V_{obsv} describes the observed values in the scene, M_{obj} the moving objects, D_{cam} the camera displacement, C the ideal background scene, M_{bgd} the moving background, N_{sys} the system noise, and S_{illum} the long-term illumination change.

Considering the two limitations considered above on scene content and time limit for surveillance videos, the camera movement and background movement are omitted

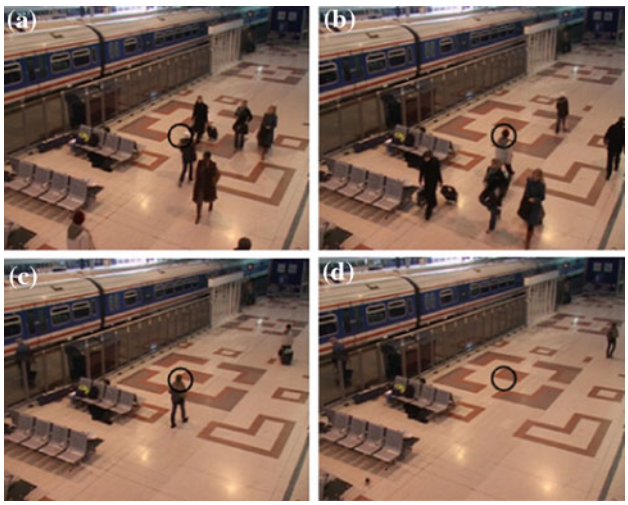

Fig. 2 Frames of a surveillance video

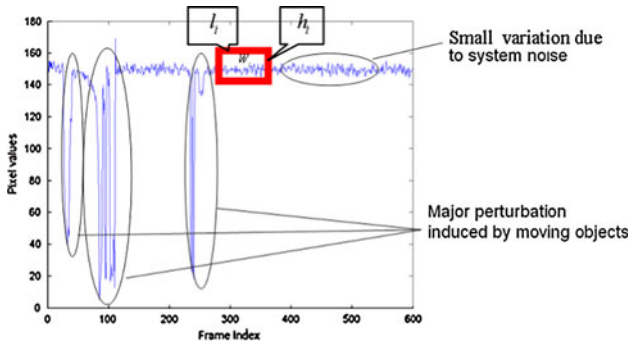

Fig. 3 Pixels intensity variation at the center of the *dark circle*

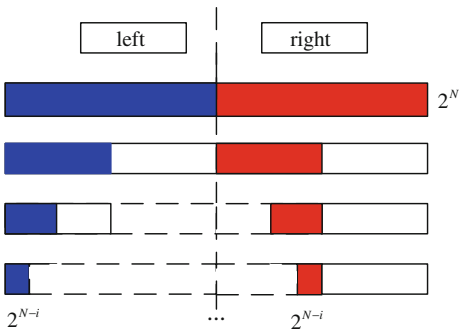

Fig. 4 Binary tree search process

in our background extraction model. Both the noise from image sensor, N_{sys} and long-term illumination change are included in N_{noise}. Then, we propose a simplified model for background extraction and foreground extraction as follows:

$$V_{obsv} = V_{background} + N_{noise} + V_{forground} \tag{2}$$

N_{noise} describes the system noise such as the background can be considered as stable, as shown in Fig. 3.

We propose an algorithm similar to the mean shift algorithm adopted in [20] to search the emerging pixels of highest frequency in a sequence as background pixels. The advantage for our algorithm is its higher performance in term of computational time since it is based on binary tree searching algorithm that can be easily parallelized instead of sequential searching as that in [20]. The emerging pixels are computed from a temporal sliding window defined by the sliding window length (l_i) and height (h_i). For example in Fig. 3, this sliding window corresponds to the red rectangle superimposed on the pixels intensity values curve. The length characterizes the number of successive frames taken into account for background extraction. The height defines the maximal range of variation for the background. This range of variation is related to the range of variation of the noise. In general, in surveillance videos, the distribution of pixel values belonging to the background varies within a small range in consecutive frames.

The use of a temporal sliding window mechanism is related to how background is perceived by the Human Visual System [30, 32]. In a general way, observers make a primary decision on whether the current pixel belongs to background after watching the first frames of a video. Then, they move their eyes onto the following frames just as moving a sliding window on those frames. If there is no change or only small change and that the change lasts very shortly in the following frames, the observers confirm their previous estimation on background. Below we give a clear definition of the temporal sliding window and of the binary tree searching algorithm used to search the pixels with the highest probability of belonging to the background. For every sliding window, the following attributes are computed: the mean value (μ_i) and the standard deviation value (δ_i) of all pixels of the current window in the current frame, and the number (n_i) of pixels emerging in this window. In this study, we have considered that the background should be relatively stable during the video (e.g. see red rectangle in Fig. 3), that means that δ_i should be small and that n_i should be high in Eq. (3).

Then background extraction is equivalent to find out the pixels satisfying the following constraint:

$$\min \frac{\delta_i}{n_i}, \quad i = 1, \ldots, k \qquad \text{s.t.} \begin{cases} \delta_i \leq \delta_0 \\ n_i \geq n_0 \end{cases} \tag{3}$$

where i is the number of the frame under study, k is the number of frames in the sequence, δ_0 and n_0 are constant values.

Since background is viewed to be quasi-stable and is often present in the video sequence, we can make the hypothesis that all or parts of the current frame detected as belonging to the background will definitely appear in the previous frames or in the future frames of the video. A novel window searching method is proposed here, where the search window is moved using a binary tree searching algorithm in 'jumping' mode rather than in 'sliding' mode as in [21]. As Lipton et al. did in [32], some seeding points are first chosen and after that the sliding window is constructed whose center is those seeding points. Compared with the above methods, our method is simpler as shown in the following pseudo code (Fig. 4).

Step 1. choose 2^N frames in surveillance video for background generation;

Step 2. divide the original search range into left and right as shown in Figure 4, each with length 2^{N-1},

scale=N-1;

Step 3.

If (scale is small enough (*scale* <2))

goto Step 5;

Else

goto Step 4;

End

Step 4.

$$T_{k\in\{left,right\}} = \frac{\delta_k}{n_k}$$

If $T_{left} < T_{right}$

left frames are viewed as the total searching frames;

Else

right frames are viewed as the total searching frames;

End

scale = scale -1;

goto Step 3;

Step 5. calculate the average value in the current sliding window;

Step 6. Stop background generation.

After all pixels of background are searched, a median filter is used to reduce the number of points, which should belong to background but emerge in foreground.

Some background pixel values might be estimated with several key frames such as the starting frame, middle frame and final frame in a video sequence for saving calculation. For example, the pixels in top-left corner in Fig. 5 are always stable in the whole video sequence and those pixels might be estimated without binary tree searching. Then, the tree estimation of background pixels could be optimized and improved with statistical data from key frames. Our proposed method can be easily extended to common surveillance video background generation. Figure 5 shows the results of background generation using this approach. The three first images correspond to original frames extracted from a video and the fourth represent the generated backgrounds.

Figure 6 shows some results obtained from different background generation algorithms. We can see ghosts of two persons in Fig. 6b, c. The background computed from mean shift algorithm is only based on the frequency of appearance of the pixel values, the person on the right emerges from #75 until the end of the video sequence, so the final generated background includes some points from this person. Mean value method also produces similar effect. However, the method that we propose with binary tree search and key frames information provides better results, as shown in Fig. 6d.

Foreground Extraction

Foreground can be extracted by comparing the background and the current frame. Considering the observed model given by Eq. (2), foreground is extracted according the following equation:

$$V_{\text{forground}} = V_{\text{obsv}} - V_{\text{background}} - N_{\text{noise}} \tag{4}$$

Some results from foreground extraction are shown in Fig. 7. We can see that the objects in foreground are extracted except for only few missed points. Those points have been wrongly classified as background because their intensity and color is almost equal to that of background. Some lost points in the foreground objects are shown in the blue circle in Fig. 7c. Motion vectors can be used to

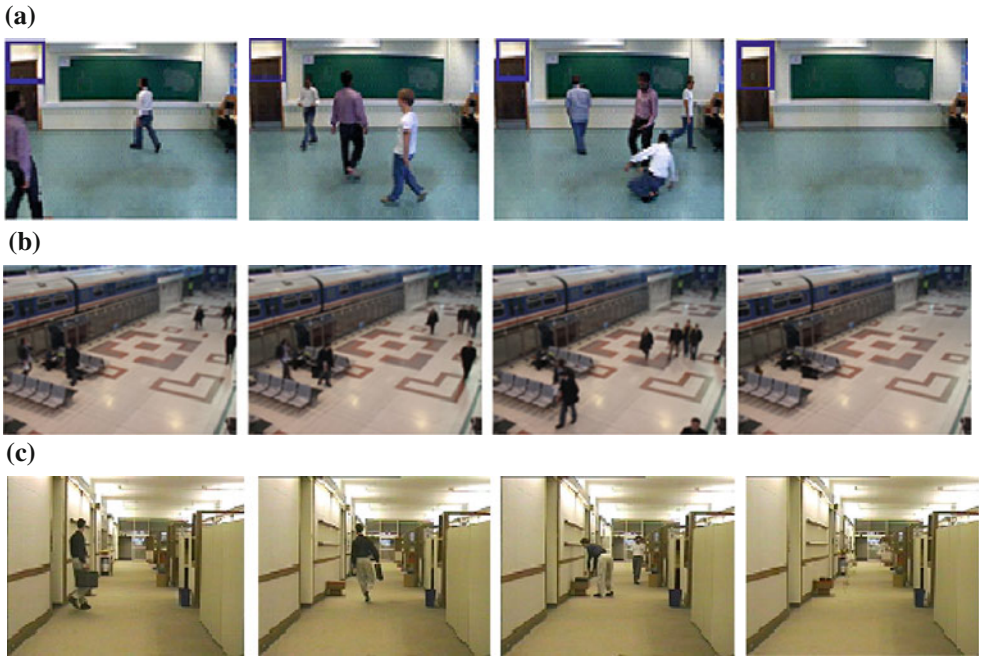

Fig. 5 Examples of background generation. **a** Background generation: example 1. **b** Background generation: example 2. **c** Background generation: example 3

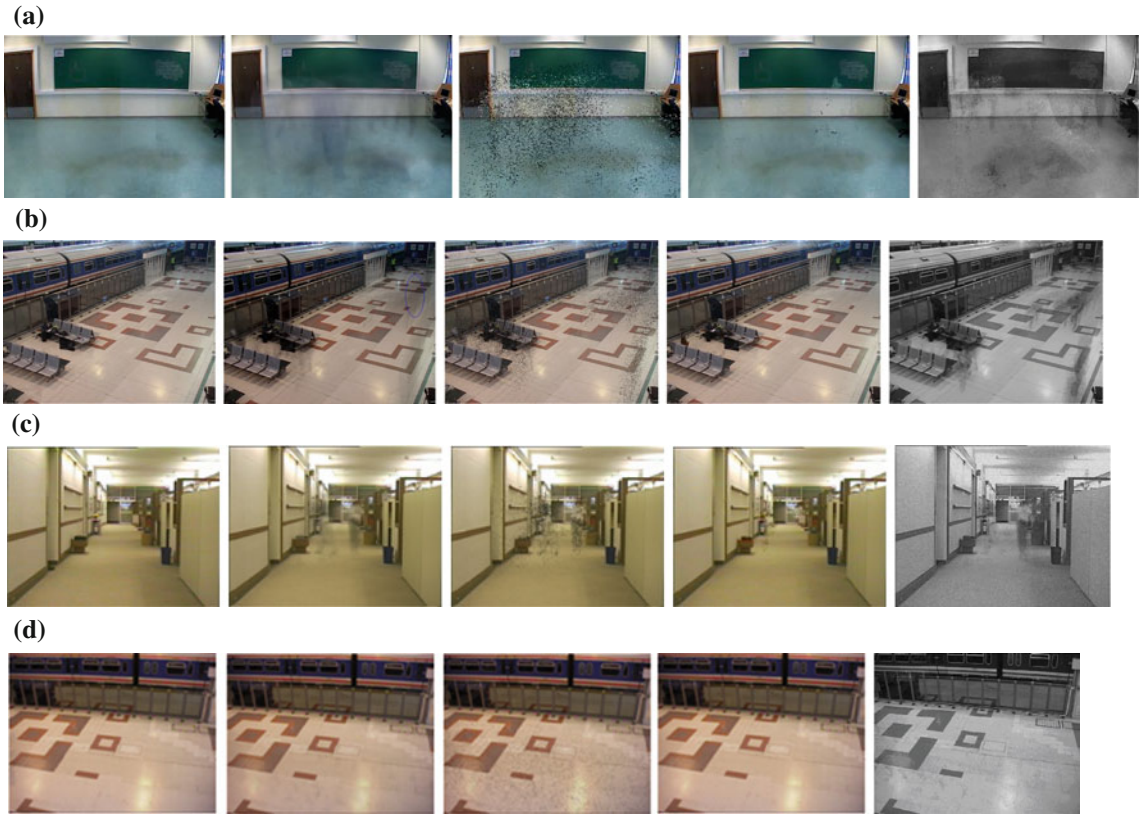

Fig. 6 Comparison of generated backgrounds: (1st column) Ideal background, (2nd column) mean value, (3rd column) mean shift (4th column) with our method, and (5th column) with Mixture of Gaussians (MOG) [28]. **a** Background generation: example 1. **b** Background generation: example 2. **c** Background generation: example 3. **d** Background generation: example 4

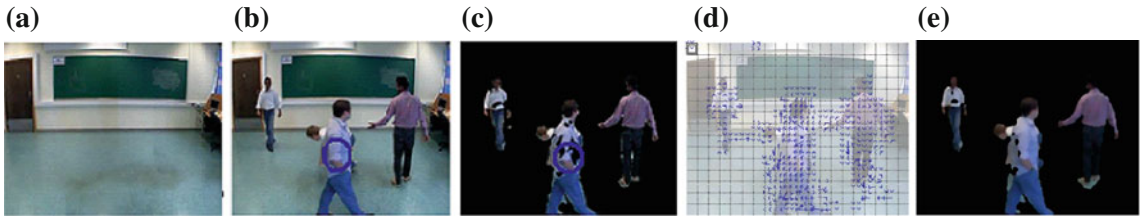

Fig. 7 Foreground extraction with or without taking into account the motion vector field. **a** Background image. **b** Current frame image. **c** Foreground objects. **d** #173 motion vector. **e** Improved foreground

improve foreground extraction. Figure 8a shows the motion vector field. Here, the regions in the extracted foreground are considered to be parts of the same object if the motion vectors in the neighborhood of the current block have similar magnitude. Using motion vector field information, the region highlighted by the blue circle in Fig. 7 is significantly improved as shown in Fig. 7e. In Fig. 8, we show that the extracted foregrounds are more relevant than those obtained using the Mixture of Gaussians method based on background subtraction, such as in [28], especially for the continuous region highlighted by the red circles in Fig. 8a–c.

Multi-feature Model for Saliency Detection

Figure 1 shows the framework that we propose for saliency detection in surveillance videos. Based on the results from background generation and foreground extraction, stationary saliency is computed via multi-feature conspicuity maps including face and low level features such as color, intensity and orientation. Motion saliency is calculated based on motion vectors analysis and on the spatial position of moving objects.

In the static saliency model proposed by Itti, there can be several saliency regions with different priorities in one image. This model can be extended to dynamic scenes if we take into account motion, as motion detection contribute to focus attention on moving objects in neighboring frames. To extend the static saliency model proposed by Itti to dynamic scenes, such as surveillance videos, we propose to weight salient regions computed from the static saliency model by motion information. Thus, when there is more than one object in a scene, the higher priorities are given to salient regions having a significant motion in neighboring frames. Furthermore, we detect more easily moving objects coming into the center field of our visual field instead of those moving off the center field. Thus, the closer the distance to the visual field center, the more salient the object. In order to take into account this effect related to the visual perception (e.g. see [12, 17, 33–36]) we have defined a distance to the center field to weight the motion saliency inversely proportional to this distance. Accordingly, moving objects with higher priorities and closer distance to the visual field center will be noticed first.

Each step of the proposed framework is detailed in the following sections. In "Multi-feature stationary saliency", we present a stationary saliency model based on face detection as a high level feature and low level features related to intensity, color and orientation. Then in "Motion saliency map and a linear combination model weighted by a Gaussian function", we present a motion saliency model

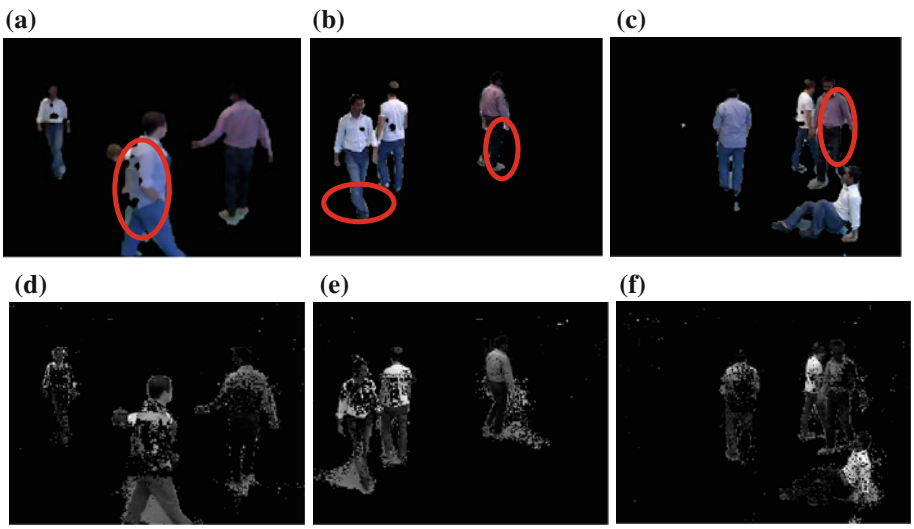

Fig. 8 Comparison of foreground extractions: **a–c** results obtained with the proposed method, **d–f** results obtained after background subtraction by the Mixture of Gaussians method [28]. **a** Frame #173 with our method. **b** Frame #113 with our method. **c** Frame # 231 with our method. **d** Frame #173 with MOG. **e** Frame #113 with MOG. **f** Frame #231 with MOG

based on motion vector field measurement and on distance weights computed according to an exponential function. Lastly in "Merging model of stationary and motion saliency maps", we present a method to merge stationary saliency map and motion saliency map.

Multi-feature Stationary Saliency

Several studies showed that low level features such as intensity, color and orientation features contribute much to our attention than other features and that the visual perception is based on a bottom-up attention framework. In the well-known model of Itti, every feature is analyzed using Gaussian pyramids and multi-scales [1]. Seven feature maps are generated including one intensity, four orientations (at 0, 45, 90, 135 degrees) and two color components (red/green and blue/yellow). Next after a normalization step, all those feature maps are combined into three conspicuity maps including intensity conspicuous map C_i, color conspicuous map C_c and orientation conspicuous map C_o. Finally, these conspicuity maps are combined together to define a single saliency map according to the following equation:

$$S_{\text{Itti}} = \frac{1}{3}\sum_{k\in i,c,o} C_k \tag{5}$$

Besides the above low level features, faces have also been considered for saliency analysis [5, 6]. Cerf et al. showed that faces are features, which focus more attention than other features in many images. Psychological tests have proven that face, head or hands can be perceived by observers prior to any other details [37]. So faces can be used as high level feature for saliency map. One drawback of Itti's visual attention mechanism model is that its saliency map model is not well adapted for images with faces. Several studies in face recognition have shown that skin hue features could be used to extract the face information. To detect face, Cerf et al. proposed in [6] to use a learning approach based on adaboost algorithm but it requires many iterations. To detect heads and hands in images, we propose instead to use the face recognition and location algorithm proposed by Koch in [38]. This algorithm is based on a Gaussian model of the skin hue distribution in the (r', g') color space which is considered as a color invariant space. For a given color pixel of values (r', g'), the model's hue response is then defined by the following equation:

$$h(r',g') = \exp\left(-\frac{1}{2}\left(\frac{(r'-\mu_r)^2}{\sigma_r^2}+\frac{(g'-\mu_g)^2}{\sigma_g^2}+\frac{\rho(r'-\mu_r)(g'-\mu_g)}{\sigma_r\sigma_g}\right)\right) \tag{6}$$

$$r' = \frac{r}{r+g+b} \quad \text{and} \quad g' = \frac{g}{r+g+b} \tag{7}$$

where (μ_r, μ_g) is the average of the skin hue distributions, σ_r^2 and σ_g^2 are the variances of the r' and g' components, and ρ is the correlation between the components r' and g'. These parameters have been statistically estimated from 1153 photographs containing faces. The function $h(r', g')$ can be considered as a color variability function around a given hue.

Next, a Gaussian Pyramid (GP) based on a multi-scale sub-sampling operation and a Gaussian smoothing was computed from $h(r', g')$. Then, the center-surround (CS) map was calculated from the pyramid, in the same way as in the Itti's model. Thus, center-surround is implemented as the difference between fine and coarse scales [1]. Lastly, the results were normalized (Norm) to obtain the saliency map S_{face} defined as follows:

$$S_{\text{face}} = \text{Norm}(\text{CS}\{\text{GP}(h(r', g'))\}) \tag{8}$$

Then, stationary saliency based on multi-features conspicuity is defined as follow:

$$S_S = f(S_{\text{Itti}}, S_{\text{Face}}) \tag{9}$$

We use a linear model with estimated weights defined as follow:

$$S_S = \frac{1}{8}(2C_i + 2C_c + C_o + 3C_F) \tag{10}$$

The linear model defined by Eq. (10) is the best combination as possible that we can obtain by an optimization-based approach in regards to the dataset considered. In order to illustrate the effect of the face feature C_F on the saliency map see Fig. 9. Here C_i, C_c and C_o features are ineffective to detect the face of the man at the center of the image. This optimization was obtained via an exhaustive process. While a heuristic-driven approach could serve very well to implement the main ideas set out here, we found that the optimization-based approach produces equal or better combinations of features than the heuristic-driven method. Note that results shown below are meant to illustrate the beneficial effect of a stationary saliency based on multi-features, not to generate the best combination that could possibly be created. That is the reason why we have not considered other combinations than the linear model usually used by other papers. Note also that the face in the right corner of Fig. 9 is not detected as this face is out of the center field so it is not considered as a salient region in regards to the central-surround vision model.

For most images containing faces, heads or hands, this model based on skin hue detection gives better results than Itti's model, i.e. gives more accurate saliency maps. The example shown in Fig. 10 illustrates the difference between Itti's model and the stationary model proposed

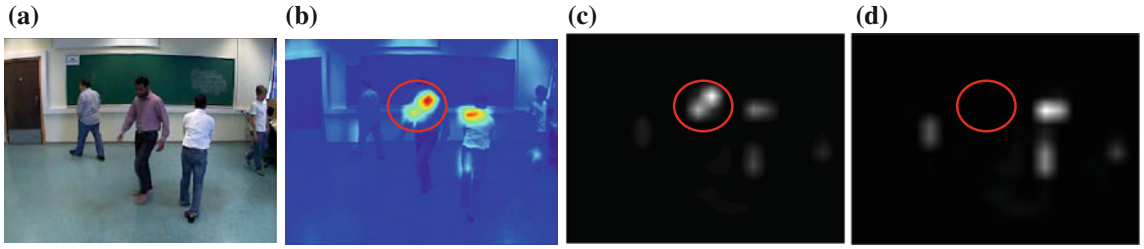

Fig. 9 Saliency map computed with different weights for merging C_F with other features. When the weight of $C_F = 0$ that means that we do not take into account the face feature C_F in Eq. (9). The main difference between these saliency maps is surrounded in *red*. **a** Original frame. **b** Video Saliency map. **c** Saliency map computed. **d** Saliency map computed superimposed to the image with a weight of 3/8 for C_F with a weight of 0 for C_F. (Color figure online)

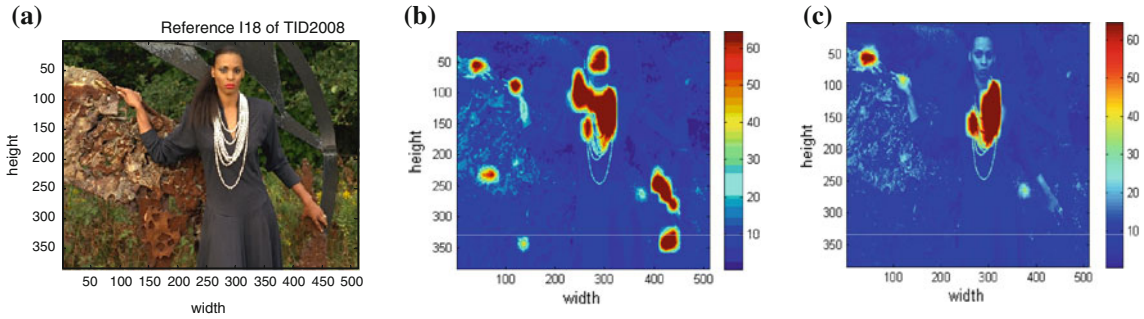

Fig. 10 Saliency region from stationary model and Itti's model. **a** I18 image in TID2008. **b** Saliency map from stationary model. **c** Saliency map from Itti's model

when we analyze images containing faces. The image 'I18' shown in Fig. 10a is a reference image including face, eyes and hands of the Tampere Image Database (see [39]). Figure 10b is the saliency map computed from the stationary model and Fig. 10c is the saliency map computed from Itti's model. The image computed from stationary model seems more reliable in terms of visual perception than those computed from Itti's model as in a general way observers focus on the neighborhood around eyes and tend to observe, find and understand the expression on faces.

Motion Saliency Map and a Linear Combination Model Weighted by a Gaussian Function

Motion feature is also involved in our video saliency map model as it carries very important information about the objects in a video and their actions within the scene. Thanks to the motion information we know what happens in a video. We can also know that some regions or objects may be much less salient in video than in images. For example, some texture of objects in images might be omitted in videos with fast motion.

In this paper, the motion information computed in videos is based on motion estimation from motion vector field computed with more than one reference frame. We use the full searching and block matching algorithms to find the most relevant motion vectors. These two algorithms are normally used in video compressing such as mpeg-4AVC/H.264. Motion vector field computed from motion estimation is shown in Fig. 11b. Frame #62 is viewed as the current frame, meanwhile the previous frame #61 is shown in Fig. 11a. We can see that there seems to be some motion in blue circles due to light flickering or other noise although there is no movement in these areas. Fortunately, the effect of those pseudo motions can be eliminated by the above foreground extraction.

The motion saliency map is computed based on the motion vector field, the intensity of motion vectors, spatial coherence and temporal coherence of the motion as in [8]. The intensity of motion vector is defined by:

$$I = \sqrt{(mv_x)^2 + (mv_y)^2} \tag{11}$$

Besides the intensity of the current motion vector (mv), the phase of motion vector, θ angle, is defined by.

$$\theta = \left| \arctan\left(\frac{mv_y}{mv_x}\right) \right| \tag{12}$$

θ is distributed in [0, 360°] after normalization.

Besides the current motion vector, the motion vectors of the neighboring blocks in the same frame are also

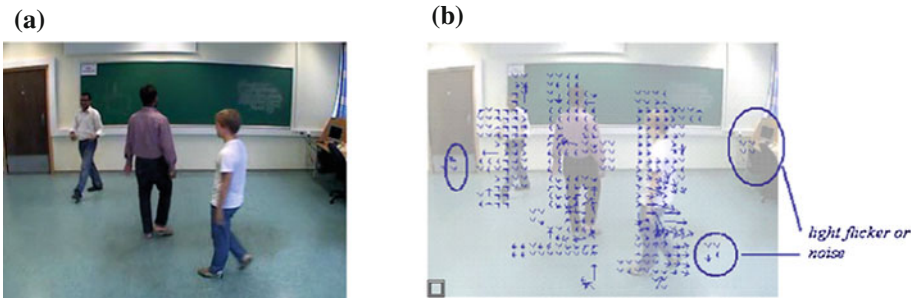

Fig. 11 Motion vectors field. **a** #61 frame. **b** Motion vectors of #62 frame

analyzed. If we consider the angle value of a motion vector as a stochastic variable and compute its probability distribution function, then the consistency of angles of motion vectors in the neighborhood of the current block can be measured by the entropy [40]. The higher the entropy is, the poorer the consistency of angles is, and therefore the motion is less salient in the current block. The spatial motion saliency is therefore described by the entropy derived from those motion vectors angles in the spatial neighborhood. The distribution probability density ρ_i of angles variations is computed using the histogram distribution of θ values within the overlapped neighboring fields as follows:

$$\rho_i = F_i \Big/ \sum_{i=1}^{N} F_i \tag{13}$$

where F_i is the frequency of the ith bin of phases histogram and N is the number of histogram bins of θ values in a field of k x k pixels.

Figure 12 shows an example of motion vectors computed from blocks of size 16 × 16 pixels and of neighborhood blocks of size 7 × 7 blocks and the corresponding phase values histogram distribution.

Then the spatial motion saliency Cs is computed from the entropy as follows:

$$C_s = -\sum_{i=1}^{N} \rho_i \times \lg \rho_i \tag{14}$$

When the phases of motion vectors of a moving object are consistent then C_s is small. The more the phases of motion vectors are disordered, the higher C_s is and in this case the motion information is not reliable. In a general way the motions of higher intensity are more consistent than motions of lower intensity, consequently C_s is more sensitive to motions with lower intensity.

Extensive psychophysics experiments have shown that motion saliency do not depend on absolute quantities, such as the direction of the motion vectors fields, how coherent their motion is, or the type of background motion [20]. Instead, the coherent perception of moving objects, even when the vertex motions are incoherent and the background motion cannot be easily explained by a physical geometric transformation, suggests that both motion saliency and perceptual organization are driven by measurements of local motion contrast [20, 41]. To account for the variability between the state at time t and the sequence of past states and to make the spatiotemporal features robust enough to handle complex dynamic backgrounds we propose to analyze the temporal consistency of motion vectors. Besides the above spatial saliency of motion vectors, the temporal consistency of motion vectors of the current frame with neighborhood frames is measured by the temporal motion vectors entropy computed from the angles histogram of motion vectors of the current block in previous $\boldsymbol{L}$ frames. When the phases of motion vectors of a moving object are inconsistent in successive frames then the motion information between blocks of same spatial position in the neighboring frames is not reliable and C_t is high. The more the phases of motion vectors of neighboring frames are consistent, the lower C_t is. In a generally way, C_t is very sensitive to object motion.

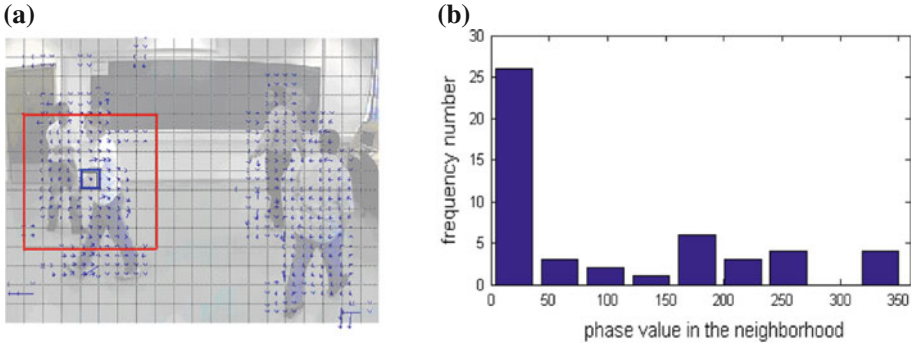

Fig. 12 Motion vector field and phase values histogram. In **a**: the *central blue rectangle* represents the current neighborhood block and the red rectangle represents the neighborhood blocks taken into account. **a** Motion vector field of the 7th frame. **b** Histogram of the 145th neighborhood block. (Color figure online)

Extensive psychophysics experiments have shown that local motion contrast attracts attention causing a pop out effect. This explains why search for a moving target among stationary distractors is easier than in the opposite case or than searching for a faster moving target among slow moving distractors [42]. The more the velocity of an element differs from that of the surrounding the more the element is salient. Here, the more the motion vectors are temporally consistent the smaller the probability is that the temporal motion is salient.

Finally, motion saliency map is computed based on the intensity of motion vector I, spatial motion saliency C_s and temporal motion saliency C_t, as suggested in [10], as follows:

$$S_M = I \cdot C_t(1 - I \cdot C_s) \tag{15}$$

This formula is justified by the fact that C_s is more sensitive to motions with lower intensity, C_t is very sensitive to object motion and that in a generally way, motions of high intensity attract much more the attention of observers than those of lower intensity.

Merging Model of Stationary and Motion Saliency Maps

The stationary saliency map S_S and motion saliency map S_M of every frame are then merged with different weights.

Some widely used methods such as Itti's model assumed that the Human Visual System can catch 3 or 5 salient objects at the same time [1, 43]. This assumption is contradicted by motion saliency models, especially for real surveillance videos in CIF size. Though not enough researches have been conducted by physiologists to support this hypothesis, we believe that the Human Visual System focuses mainly on only one moving object when there are several objects moving simultaneously. This object corresponds in general to the most salient moving object that is coming into the center region of our visual field instead of other objects moving outside our visual field. This hypothesis is supported by the various gaze maps and experiments that we conducted on surveillance videos.

Considering that the Human Visual System focus more easily his attention on the moving object in the center of observing window than those that are far away from the center, we propose to weight the motion saliency according to the following distance, next to merge the motion saliency map with the stationary saliency map as follows:

$$S_{V_G} = (\alpha \cdot S_M + (1 - \alpha) \cdot S_s) \cdot w_i \tag{16}$$

$$w_i = e^{-d} \tag{17}$$

$$d = \frac{\sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}}{8} \tag{18}$$

$$x_c = \frac{\text{width}_{\text{mb}}}{2}; \quad y_c = \frac{\text{height}_{\text{mb}}}{2} \tag{19}$$

where width_{mb} and $\text{height}_{\text{mb}}$ are the height and width from the center of the frame and (x_c, y_c) are the coordinates of the center. w_i is the weight of the block number i centered on pixel of coordinates (x_i, y_i) of size (16×16) located at a distance descriptor d of pixel (x_c, y_c). w_i is normalized into [0,1] as shown in Fig. 13b and used to describe the spatial position effect on motion saliency map. The more a block is closed to the center point, the higher its weight is.

Beside the above merging method based on a 2-D approximated Gaussian distance model, we have also tested other merging models for comparison purpose including Mean, Max and Multiplication merging models as follows:

$$S_{V_{\text{mean}}} = \frac{(S_M + S_S)}{2} \tag{20}$$

$$S_{V_{\text{max}}} = \text{Max}(S_M, S_S) \tag{21}$$

$$S_{V_{\text{multi}}} = S_M \times S_S \tag{22}$$

Each salience map is linearly normalized to have zero mean and unit standard deviation. The best experimental combination that we obtained for Eq. (16) have been achieved for $\alpha = 3/7$. This combination based on a linear

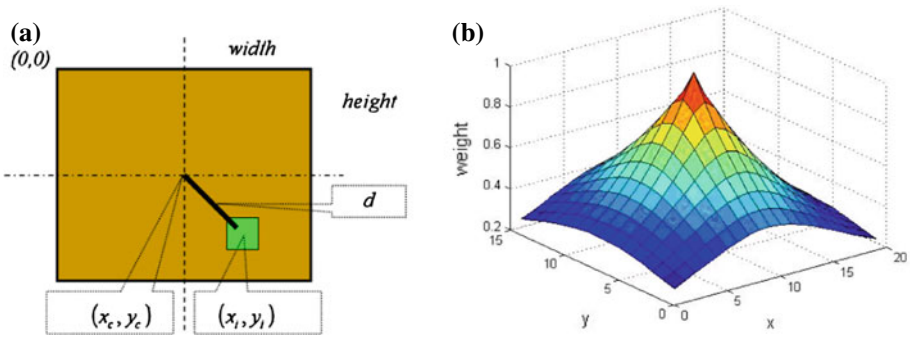

Fig. 13 The motion saliency computed in the current block is weighted based on the distance of the block to the center. **a** Block position. **b** 2-D distance model to weight blocks

combination had been obtained from a database which includes the TID2008 database [39] which contains 612 photographs with faces and head images, and our own experimental database of surveillance videos, which includes more than 1,000 frames.

Discussion and Experiments

In order to analyze the performance of the video saliency maps computed from the model implemented, we compute in "Comparisons between gaze maps and saliency maps" the gaze maps of several surveillance video sequences next we compared our results with these gaze maps. Figure 14 show three examples of indoor surveillance videos used for our analysis. Next, in "Quantitative comparison of saliency maps", we make quantitative comparisons based on NSS values.

Comparisons Between Gaze Maps and Saliency Maps

In our experiments we mainly used indoor surveillance videos with people moving inside a static background. As example in Figs. 15a, b there is two people with normal walking action, meanwhile in Fig. 15c there is four people with sudden actions. In the following, we mean by simple action a video sequence with a continuous movement, oppositely to sudden actions. Simple actions are differentiated from sudden actions thanks to the motion vector field. Simple motion are characterized by motion vectors consistent both spatially and temporally, meanwhile sudden actions are characterized by more incoherent motion vectors, as example see Fig. 15d, e. Let us note that in these two figures some fault motions have been detected due to local light changes in the background. These motions do not affect much the motion saliency detection as there is no motion in neighboring blocks. Figure 15b, c show also that sudden actions attract more the attention than continuous actions. This example 15 illustrates the interest to analyze the coherence of motion vectors in function of neighboring blocks. Furthermore, Fig. 15a, b show that people moving within the center of the frame attract more the attention than actions within the surround.

In this study we have used 16 video sequences shot by ourselves. Large video sequences have been divided into smaller video shots. The duration of the video shots is from 6 to 21 s. Then we have computed gaze maps from subjective experiments done on those video shots. All the experiments and parameter estimations outlined in this paper are based on it. In our experiments, we have not used outside videos or inside videos with camera motion as either the motion vector fields are irrelevant or request to compute the camera motion before extracting the background. As our model is based on a face features approach it is not adapted to detect other objects defined by other features. To explore this kind of videos we should consider more complex models than the Gaussian model and other features than the skin hue distribution. As example see Fig. 16. Here, C_F and w_i features are effective to detect the butterfly as it is positioned at the center of the image and its hue distribution is quite similar to the skin hue distribution. The C_F feature is more effective than the hue distribution of the background is different of the hue distribution. That means that our saliency map algorithm can be extended to any moving object provided, however, to implement appropriate descriptors. To reach this aim, we could consider a learning strategy to learn which features best characterize/distinguish moving objects in the scene [44–47].

The more video saliency maps are closed to the gaze maps of videos the more the model used is performing in terms on visual detection. To compute the gaze maps, we did subjective experiments with an eye tracker. Twenty observers, aged between 25 and 42, participated to the experiments done with a 50 Hz infra-red SMI eye tracker. During the experiments, observers were asked to watch surveillance videos on a 17 inch CRT display as they normally would do under normal viewing conditions. The resolution of the display was of 1,024 × 768 pixels. The distance between the monitor and the observer was

Fig. 14 Examples of indoor surveillance videos where only people move in the scene. In these three video the people are the same but their motion is different. **a** Video 1: 284 frames. **b** Video 2: 194 frames. **c** Video 3: 526 frames with 4 moving people with 4 moving people with 4 moving people

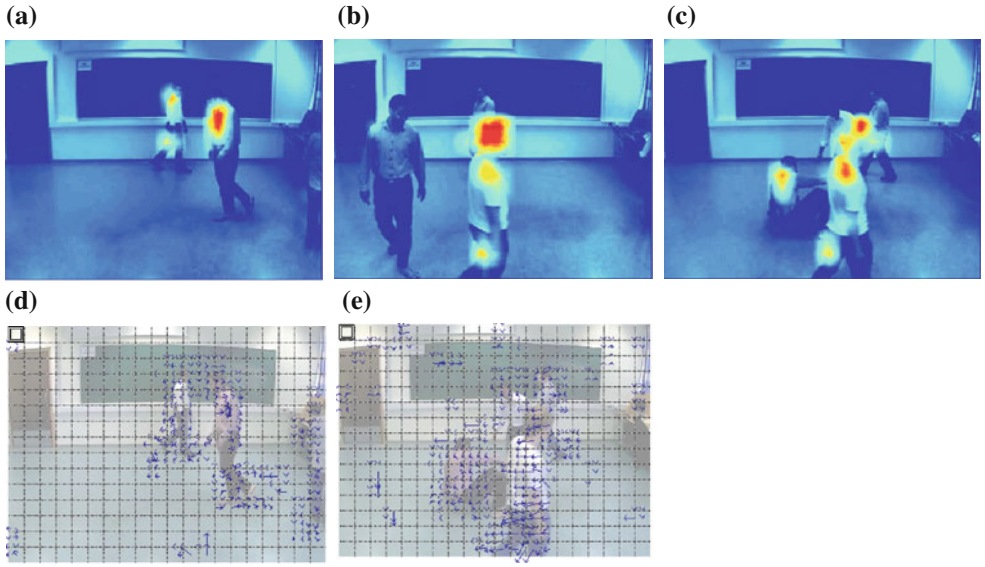

Fig. 15 Gaze map of different frame of the video 3 with different actions. **a** People with normal walking action within the center (frame #389). **b** People with normal walking action anywhere in the scene (frame #155). **c** People with sudden actions anywhere in the scene (frame #476). **d** Motion vector field of frame #389. **e** Motion vector field of frame #476

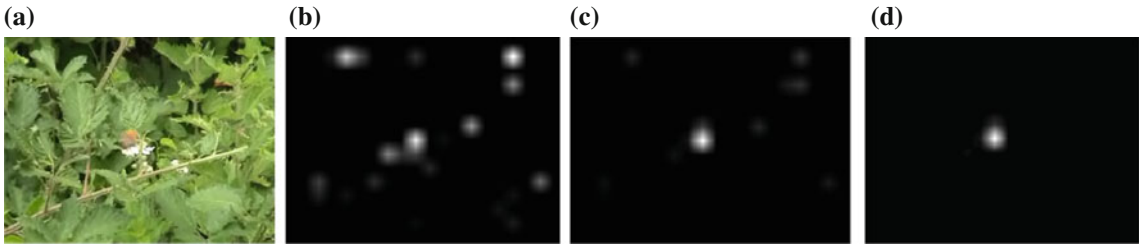

Fig. 16 Example of saliency map computed for a frame without face. When the weight of $C_F = 3/8$ and w_i = center-surround filter, the stationary saliency map better detects the butterfly at the center of the image. **a** Original frame. **b** Stationary saliency map with a weight of 0 for C_F. **c** Stationary saliency map with a weight of 3/8 for C_F without center-surround weight (i.e. $w_i = 1$). **d** Stationary saliency map with a weight of 3/8 for C_F with center-surround weight

between 60 and 70 cm. Before each experiment, a test was performed to detect the dominant eye of the observer. During experiments, observers' dominant eye was tracked and tracking data were saved with a system processing with the SMI IView software. Gaze maps were computed from fixation points of the dominant eye. First, a fixation frequency map was computed for each frame of each video by adding up all the fixation positions of each observer. As with the Human Visual System, the fixation frequency map was next filtered by a spatial Gaussian filter. These frequency maps were filtered by a spatial Gaussian filter of $\sigma = 37$, which was chosen to approximate the size of the viewing field corresponding to the fovea in the gaze map [48]. The size of the Gaussian window was of 40×40 pixels. Next, the average of these Gaussian maps for all observers was computed, then normalized and superimposed to the original frame with a colormap of 64 color values, where blue colors correspond to lowest gaze map values and red colors correspond to the highest gaze map values, i.e. the most salient regions of a video frame. Figure 17 shows the SMI device and the gaze map computed for a frame superimposed to the original image.

Figure 18 shows two other frames of the same surveillance video and the corresponding gaze map and saliency map superimposed to the original image. Figure 18a, d are two original frames of a same video. Figure 18b, e show the corresponding gaze maps superimposed to the original images and Fig. 18c, f show the video saliency maps computed from the model that we propose superimposed to the original images. In surveillance videos, the attention of observers is usually attracted by moving objects, especially those entering into the center area of the observed image. That why we have proposed above to weight our video saliency map model by a Gaussian distance. The results of Fig. 18c show the effectiveness of this weighting function and of the video saliency map model that we propose. As

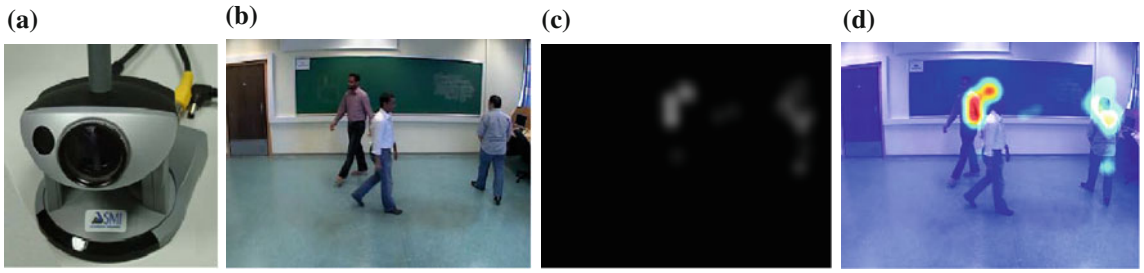

Fig. 17 Gaze map based on eye tracking and Gaussian filtering. **a** SMI device. **b** Original image. **c** Gaze map for #65. **d** Gaze map superimposed to the image

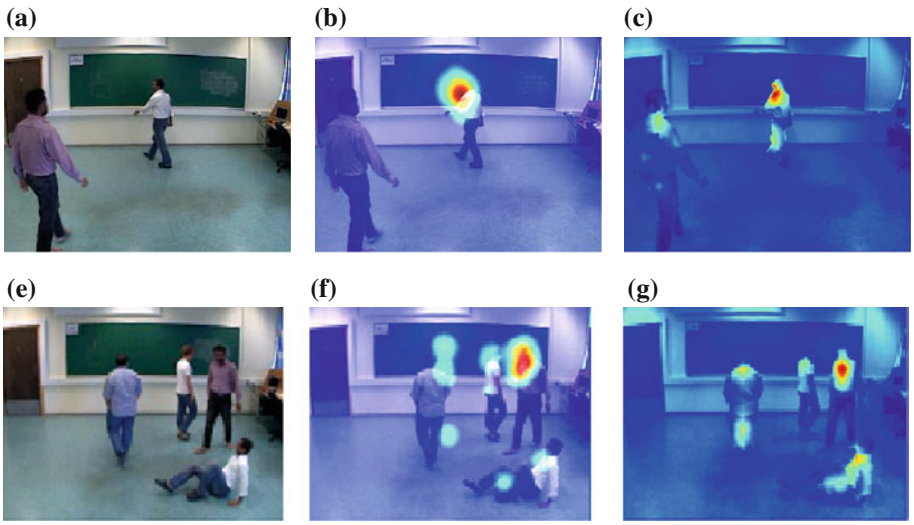

Fig. 18 Examples of gaze maps computed from visual experiments and of corresponding video saliency maps computed from our video saliency model. **a** Original image. **b** Gaze map superimposed to the image. **c** Video Saliency map superimposed to the image. **d** Original image. **e** Gaze map superimposed to the image. **f** Video Saliency map superimposed to the image

we can see in Fig. 18c the moving object, i.e. the person, which is in the center area of the image is detected both by the gaze map computed from visual experiments and by the video saliency map model that we propose.

Additionally to the above video saliency maps computed merging stationary and motion maps with Gaussian weights, other merging modes such as Mean, Max or Multiplication have been also used in our experiments. Figures 19 and 20 show two saliency maps computed with different merging modes. Figures 19b and 20b represent the gaze map images with the subjective gaze map superimposed on the original image. Figures 19c and 20c represent the corresponding video saliency maps image computed with Gaussian weights superimposed on the original image.

Among the four merging modes tested including Mean, Max, Multiplication and Gaussian weights, the one which gives the best results, i.e. the one for which the video saliency maps are the closest of the subjective gaze maps, is the mode computed with the linear combination defined by Eq. (15) weighted by a center-surround function. This shows that besides the stationary and motion features other information such as distance or depth might also affect our visual perception. The example of Fig. 21 shows the impact of the center-surround weight on the saliency map. Differences between Fig. 21c, d are subtle since both of them are based on the same foreground and motion vector field. But as with the center-surround weight the saliency of boundary regions is reduced then the saliency map better approximates the gaze map illustrated by Fig. 21b. Inversely, there is no difference between Fig. 21g, h as all people are outside the center of the image. The example of Fig. 22 shows the impact of the motion saliency on the saliency map. $\alpha = 0$ means that we only take into account the S_S stationary saliency map to compute the saliency map without any motion information so only low level features and face feature are considered in the saliency map, as shown in Fig. 22d. $\alpha = 1$ means that we only take into account the S_M motion saliency map without any low level features and face feature in the saliency map, as shown in Fig. 22c. If we compare the saliency maps predicted by the proposed model with gaze maps obtained by subjective experiments, here the best results are obtained with $\alpha = 3/7$. Let us note that even if differences between Fig. 22b, d, and between Fig. 22f, h are subtle they are nevertheless noticeable. Some of these differences are surrounded in red.

In order to compare the relevance of the video saliency model that we proposed with other saliency models such as

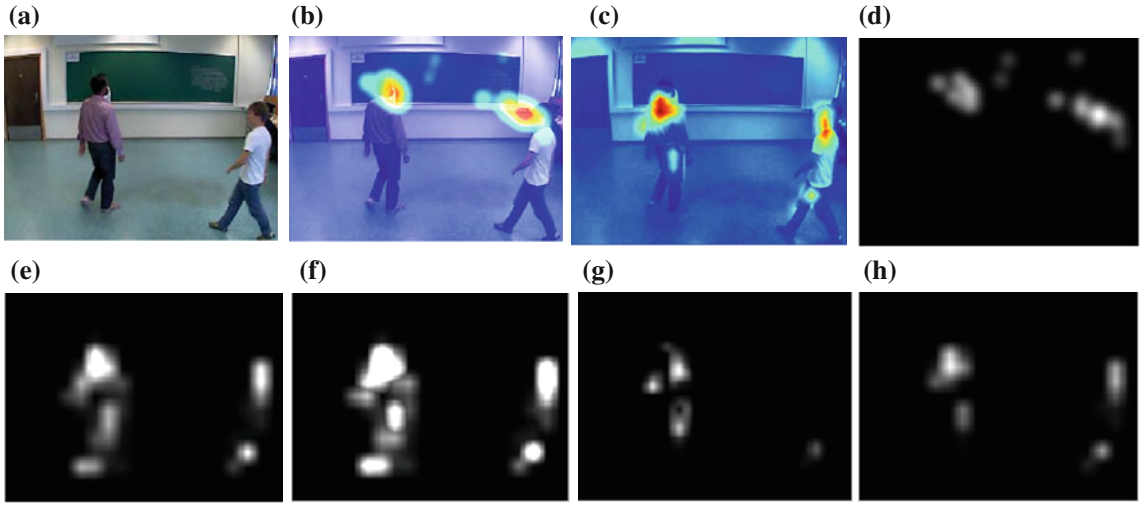

Fig. 19 Example of gaze maps computed with different kinds of stationary and motion merging modes. **a** Original #41. **b** Gaze map superimposed to the image. **c** Video saliency map superimposed to the image. **d** Subjective gaze map. **e** Mean. **f** Max. **g** Multiplication. **h** Linear combination

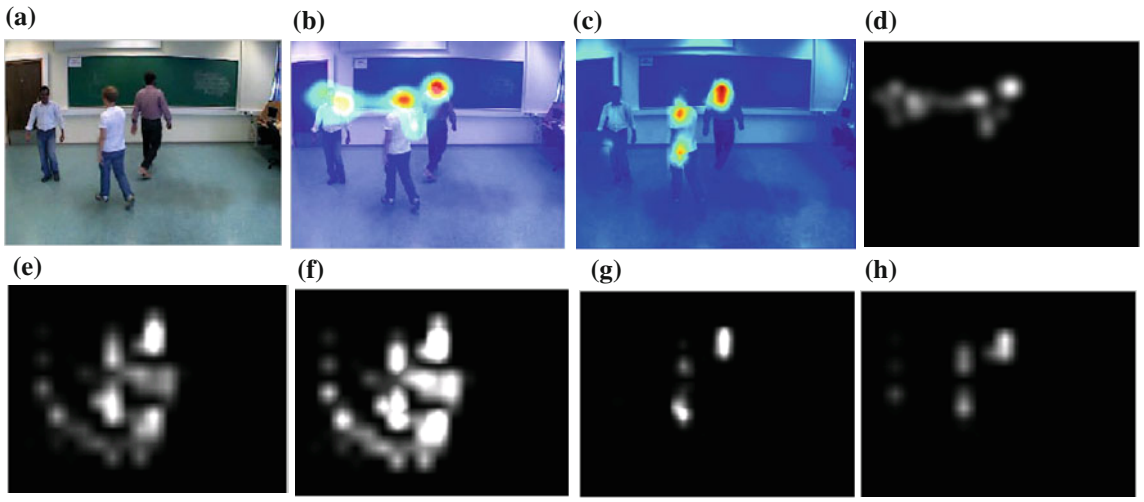

Fig. 20 Example of gaze maps computed with different kinds of stationary and motion merging modes. **a** Original #91. **b** Gaze map superimposed to the image. **c** Video saliency map superimposed to the image. **d** Subjective gaze map. **e** Mean. **f** Max. **g** Multiplication. **h** Linear combination

Itti's model [1], frequency-tuned saliency detection [3] and phase spectrum saliency model [4, 43], we propose to study their closeness to the corresponding subjective gaze maps. Itti's model is considered as a reference model for stationary saliency map detection. Itti's model can be extended to videos, based on a frame by frame approach, but in that case the inter-frame information and motion information are not taken into account. Then for a fair comparison, we compare also our model with GBVS model which is an improved saliency detection model of the Itti's model [49]. For computing, the saliency map of the current frame the GBVS model uses information computed from previous frames. As example see Figs. 23 and 24. Among the five saliency models tested the one which gives the best results, i.e. the one for which the video saliency maps are the closest of the subjective gaze maps, is the video saliency model computed with center-surround weights. The main errors of detection obtained with the other models tested are linked to the detection of salient regions in stationary frame which are not salient in regards to motion. Other errors are linked to the detection of salient regions in stationary frame, such as the chair and the curtain at right in the Fig. 23, which are outside the center area. Other differences, more subtle, can also be seen as example on moving people. Although GBVS is effectively relevant to detect salient regions, such as moving people, GVBS gives too much importance on large moving regions, such as the legs and foot in Figs. 23 and 26. Compared to our model

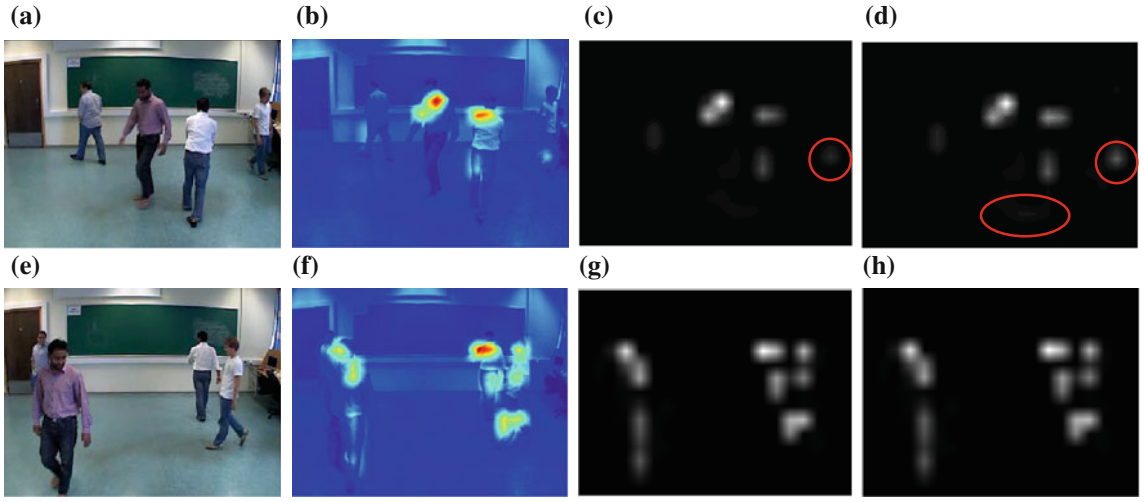

Fig. 21 Effect of the Gaussian weight in Eq. (15). **a** Original frame. **b** Video saliency map superimposed to the image. **c** Saliency map computed with center-surround weight. **d** Saliency map computed without center-surround weight (i.e. $w_i = 1$). **e** original frame. **f** Video saliency map superimposed to the image. **g** Saliency map computed with center-surround weight. **h** Saliency map computed without center-surround weight (i.e. $w_i = 1$)

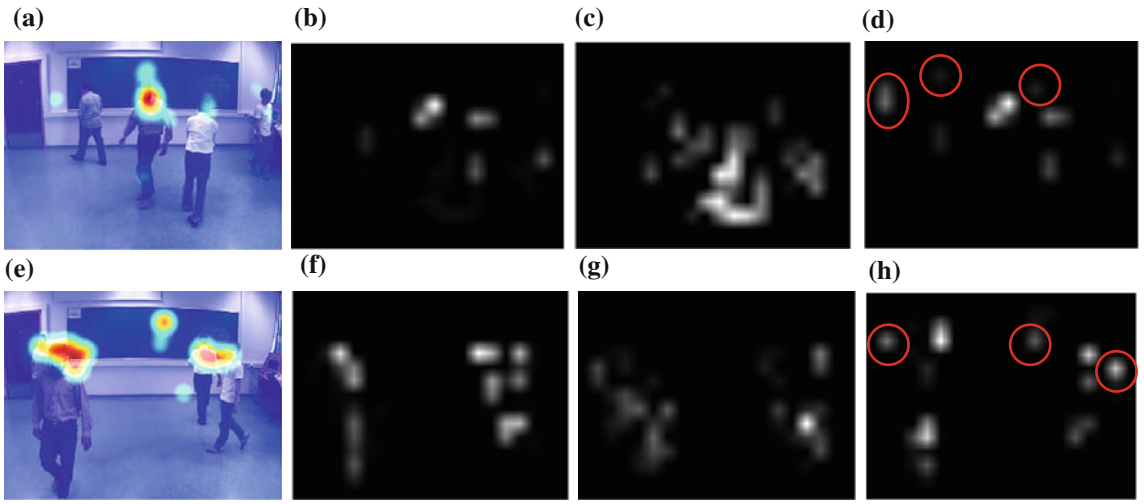

Fig. 22 Effect of parameter α in Eq. (15). **a** Gaze map of Fig. 21a. **b** Saliency map computed with $\alpha = 3/7$. **c** Saliency map computed with $\alpha = 1$. **d** Saliency map computed with $\alpha = 0$. **e** Gaze map of Fig. 21e. **f** Saliency map computed with $\alpha = 3/7$. **g** Saliency map computed with $\alpha = 1$. **h** Saliency map computed with $\alpha = 0$

the two main shortcomings of the GVBS model is that face or background are not considered in this model. However, gaze maps computed during our subjective experiment show that it is important to pay more attention on face or head instead of all the body of people. Figure 25 shows effectively that the saliency model proposed in this paper gives better results than the GBVS model and that the saliency maps computed with our model are closer to gaze maps. This is not surprising, as strong motion cues are not present in our study. During periods of rather still video content, color, intensity and orientation are better predictors of saliency than motion, which essentially yielded no output during these periods [50]. Likewise, color, intensity, and orientation are better predictors of saliency than motion for quasi-stable background regions in surveillance videos. We did not do a systematic comparison of our model with other saliency models because as indicated in the introduction these models are not comparable to our model from a theoretical point of view. Thus, human eye fixation data used by Itti et al. [15, 50] for dynamic scenes cannot be employed to analyze the performance of our model as the proposed model has been developed for

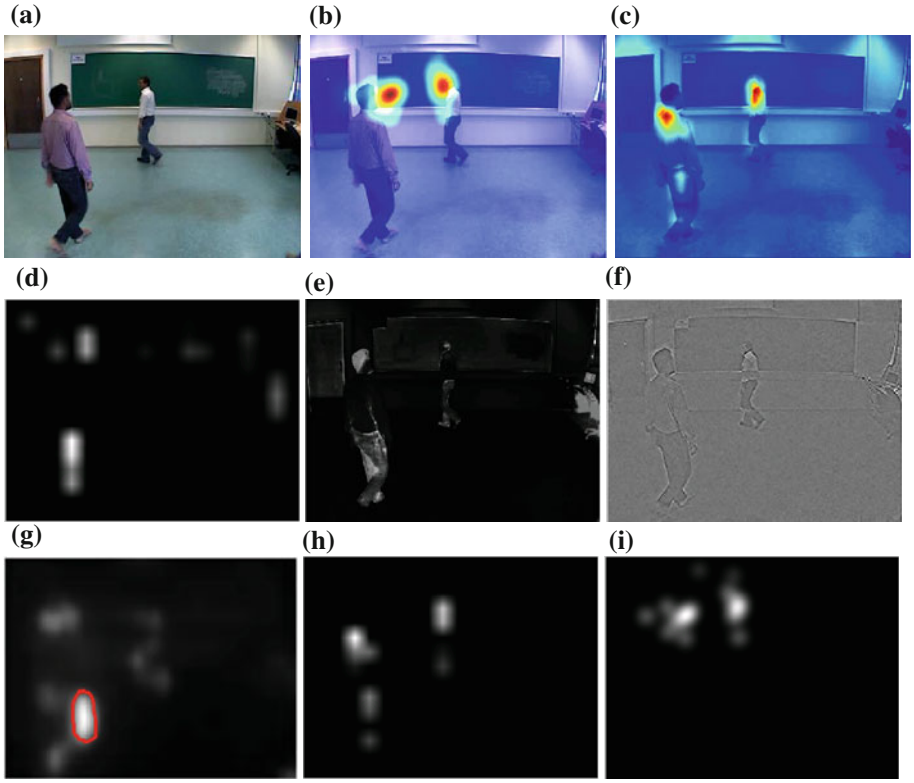

Fig. 23 Saliency maps comparison between the saliency model that we proposed and the Itti's, frequency tuned, phase spectrum and the GBVS models. **a** Original #21. **b** Gaze map superimposed to the image. **c** Video Saliency map superimposed to the image. **d** Itti's model. **e** Frequency-tuned model. **f** Phase spectrum model. **g** GBVS model. **h** Our model. **i** Subjective gaze map

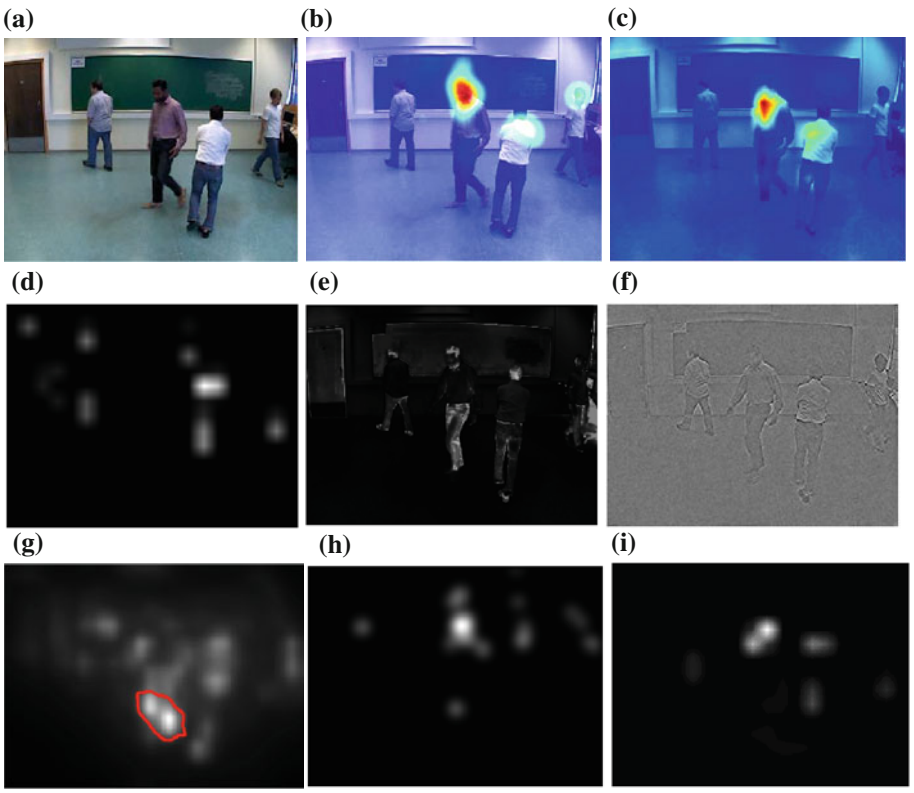

Fig. 24 Another example of comparison between different saliency metrics. **a** Original #273. **b** Gaze map superimposed to the image. **c** Video saliency map superimposed to the image. **d** Itti's model. **e** Frequency-tuned model. **f** Phase spectrum model. **g** GBVS model. **h** Our model. **i** Subjective gaze map

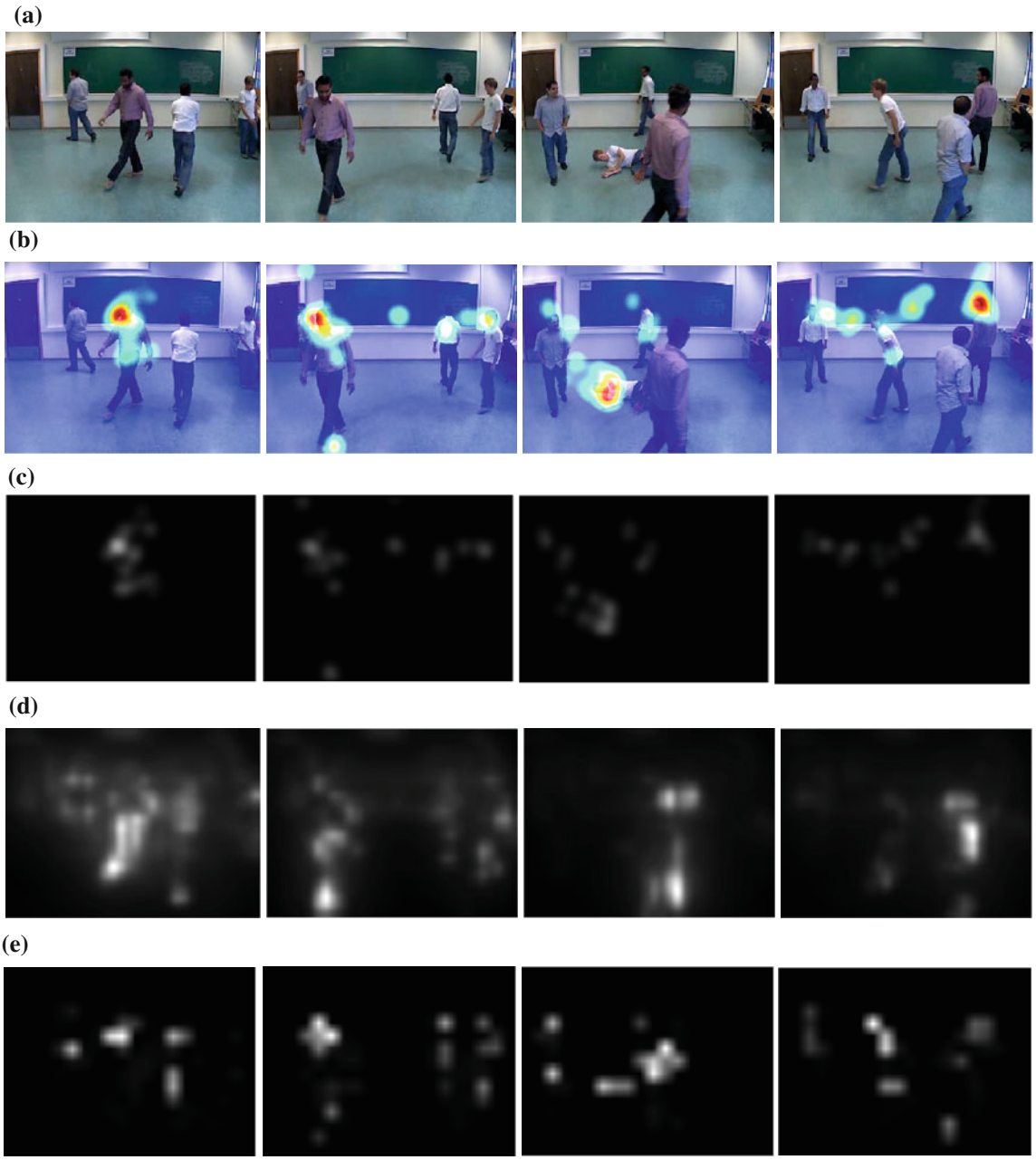

Fig. 25 Comparison of saliency maps computed from the GBVS video saliency detecting model and our proposed model. **a** Original frames: #7, #42, #108 and #190. **b** Gaze map superimposed onto the original frames: #7, #42, #108 and #190. **c** Subjective gaze maps for frames: #7, #42, #108 and #190. **d** Saliency maps computed with the GBVS Model. **e** Saliency maps computed with the proposed saliency map model

surveillance videos where the background is quasi-static and salient moving objects moves continuously.

Quantitative Comparison of Saliency Maps

Besides subjective comparison, various objective criteria can be used for the comparison of saliency maps such as distance metrics and ROC [49, 51]. We use the Normalized Scan path Saliency (NSS) to estimate the overlapping rate between gaze map and saliency map as in [9, 52]. The NSS of kth frame is defined as follows:

$$\mathrm{NSS}(k) = \sum\nolimits_{x,y} \left(\frac{\left(\overline{G_V(x,y,k) \times S_{Vm}(x,y,k)} - \overline{S_{Vm}(x,y,k)} \right)}{\delta_{S_{Vm}(x,y,k)}} \right) \tag{23}$$

where $G_V(x, y, k)$, the value of the subjective gaze map of the frame k normalized to obtain unit mean and

$S_{Vm}(x, y, k)$, the value of the video saliency map, computed at pixel of coordinates (x,y) for the frame k, normalized according a scale ranked from 0 (no saliency) to 1 (highest saliency).$\delta_{S_{Vm}(x,y,k)}$ represents the standard square error of saliency maps computed from every frame.

The NSS has been computed for each frame of each videos used for this study. The mean NSS value is a standard score criteria which expresses the divergence of the subjective gaze maps from the mean saliency maps in function of the standard deviations of the video saliency model. This criterion was especially designed to study eye movement data and so, the corresponding results can be easily interpreted [12]. The greater the value of the score is, the greater the correspondence than would be expected by chance between fixation locations and the salient points predicted by the model is.

Besides the above subjective gaze map, another randomized eye movement gaze map is also used to compare the saliency map predicted by our model with different saliency models. The randomized gaze map associates to each frame of a video the fixation locations of observers when they were looking at another video clip. If a model can correctly predict the fixation locations of eyes, the NSS of subjective gaze maps and saliency maps should be high meanwhile the NSS of randomized gaze maps and saliency maps should be low at the same time.

As we have not real subjective randomized gaze map, that would mean we should observe two video sequences at the same time in subjective experiments, one way is to use a random function to generate randomized gaze map; another way is to use the useless or irrelevant gaze map as randomized gaze map instead of random array generated by random functions as in [53]. Examples of unacceptable gaze map are illustrated in Fig. 26. The content of frames shown in Fig. 26 is considered as unacceptable to compare the saliency map predicted by our model with different saliency models as: in Fig. 26a the image is too blur and the motion of the butterfly is not salient consequently moving vector fields are useless, in Fig. 26c some people are seen from behind so features based on skin hue are useless. The reader might think that we have implemented a rather extreme operation. But a little thought illustrates this is not the case. As we say above, the NSS of subjective gaze maps and saliency maps should be high meanwhile the NSS of randomized gaze maps and saliency maps should be low at the same time. Being more restrictive on the number of acceptable gaze maps, we penalize more the NSS score of our saliency model but we strengthen the accuracy of our saliency model.

Normally, NSS value is higher for real gaze maps than for randomized generated gaze maps or irrelevant gaze maps. NSS value on randomized generated gaze maps should be the smallest. The closer saliency map and gaze map are, the better the performance of the saliency model is. The higher the difference between NSS value computed from real gaze maps and NSS value computed from irrelevant gaze maps or randomized generated gaze maps is, the better the performance of the saliency model is.

Table 1 gives out some data about NSS with real gaze maps or randomized gaze maps. Once again for comparison purpose, we have considered five saliency maps derived from different weights merging methods for stationary saliency map and motion saliency map. We found similar results using a variety of different metrics (ROC, Earth Mover's Distance, Kullback–Leibler Distance, etc.). The method with center-surround weight (S_{V_G}) gets the best performance compared with other merging modes.

Table 1 also shows that HVS trends to focus on the object moving into the center of insight window instead of those far away from the center.

Table 1 Gaze map and saliency map comparison

Criteria	Fuse mode			
	$S_{V_{multi}}$	$S_{V_{\max}}$	$S_{V_{mean}}$	S_{V_G}
NSS on real gaze maps	0.717	1.0256	1.045	1.1815
NSS on randomized generated gaze maps	0.0002	0.0002	0.0002	0.0002
NSS on irrelevant gaze maps	0.3998	0.5972	0.7587	0.5069

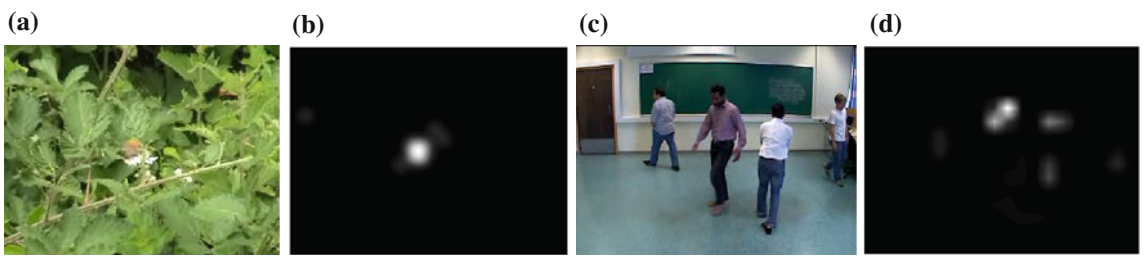

Fig. 26 Examples of unacceptable gaze map to compare the saliency map predicted by our model with different saliency models. As the content of the original frames cannot be exploited by our approach the gaze maps of these frames has been used as randomized gaze map. **a** Original frame. **b** Corresponding gaze map with a weight of 0 for C_F (see Eq. 10). **c** Original frame. **d** Corresponding gaze map

Table 2 Comparison with different saliency models

Criteria	Model					
	Phase spectrum	Itti	FT (frequency tune)	Motion Saliency [10]	GBVS (video saliency [49])	Proposed method
NSS on real gaze maps	0.0018	0.0657	0.1598	0.453	0.7115	1.1815
NSS on randomized generated gaze maps	0	0.0001	0.0002	0.0001	0.0001	0.0002
NSS on irrelevant gaze maps	0.0079	0.1563	0.1276	0.6127	0.6211	0.5069

We have also compared our result with other saliency detection algorithms including Itti's model, frequency-tuned saliency detection and phase spectrum saliency as shown in Table 2. As there is no standard visual saliency model for surveillance video, here the motion saliency map used in [10] and GBVS [49] were used for results comparison.

If the saliency maps computed with the proposed model were much close to gaze maps, NSS on real gaze maps should be higher than the two other NSS values and the differences between NSS on real gaze maps and the two other NSS values should be also higher than those computed with other models. From data in Table 2, we show that NSS on real gaze maps computed with our proposed model is far higher than that of other stationary models such as Itti's, frequency tuned and phase spectrum, and that NSS values are quit similar on randomized generated gaze map. The models considering inter-frame information, such as the motion saliency model [10], the GBVS model [49] and our model, show higher NSS values on real gaze maps than the stationary models. That confirms that the motion information and the inter-frame information play important role for detecting saliency in video. Let us also note that NSS on real gaze maps computed with the motion saliency model [10] are much lower than that computed with the GBVS model or our model. The reason for that may be due to video sequences used in this study and to the motion saliency model used, as this latter is based on the computation of motion vectors from block match motion estimation in mpeg2 without considering background or foreground. Indeed low quality recorded video can definitely decrease the precision of motion vectors and then further decrease the precision of the saliency map.

Furthermore, we can also note that the NSS values in Table 2 on irrelevant gaze maps are higher with the motion saliency model, the GBVS model and our model than with stationary methods. At the same time, we can note that all the NSS values on irrelevant gaze maps in Table 1 are high meanwhile the corresponding NSS values on randomized generated gaze maps are very low. Here, we can do some interpretations with reference to Fig. 26a, c which have been considered irrelevant and which have a very different content. If we look at their gaze map, we can see that observers focus, as for any video, on the center part of video frames no matter what kinds of video they are looking for. Therefore, the probability that a gaze map emerges in the center of a frame is very high. This should not happen with randomized generated gaze maps since these latter are based on a random function. That explains why NSS values are higher on irrelevant gaze maps.

Conclusion

In this paper, a new spatiotemporal saliency detection algorithm for video surveillance is proposed. With the knowledge of scene content, background generation and foreground objects extraction are analyzed, and then multi-features including high level feature such as face and other low level feature including color, orientation and intensity have been used to compute stationary feature conspicuity maps. Motion saliency map is based on the motion vector analysis. Motion saliency map and stationary saliency map are then merged with Gaussian distance weights. We have compared saliency maps predicted by the proposed model with gaze maps of surveillance videos obtained by subjective experiments. Comparing to previous work, we show that our multi-feature-based video saliency detection model gives a closer correspondence to gaze map. The objective of this paper was to further investigate the effect of several spatiotemporal saliency features which are much correlated to the human visual perception of saliency instead of generating a new saliency detection model based only on a computer vision approach without taking into account cognitive computation. Under this objective perspective, our results are very encouraging and show substantial improvements.

It is also interesting to note the effect of the bottom-up and top-down process set out here, based on the merging of spatiotemporal saliency features, on the saliency detection. Two embodiments of the main idea were presented or suggested. The first, which serves to motivate the discussion, is a reasonable heuristic approach that drives the merging of spatiotemporal saliency features. The second,

an optimization-based approach, casts the foundations to extend the saliency detection approach to other categories of targets than people and to develop an online saliency detection process based on a multiple instance learning approach. That is, with many possible targets, different observers may orient toward different locations, making saliency model more difficult for a simple metric to accurately predict all observers [26]. In that case, as suggested in [26], dynamic metrics should be used to improve more steeply, indicating that stimuli which more reliably attracted all observers carried more saliency.

In this paper, we have mainly considered surveillance videos with quasi-stable background. In the next step, we will focus on more complicated scene where background and foreground objects are both moving. More refined algorithm should be necessary to get the suitable foreground objects for saliency analysis. Unfortunately, the current binary tree search actually used for extracting background pixel could not be used as it requires too much computational power. Therefore, both neighborhood information and multi-scale technique will be explored for optimization. Lastly, the merging mode of stationary saliency and motion saliency might be further improved by considering other information such as saliency history. Our model opens new perspectives for more sophisticated models and experimental scenarios. which are enough simple to ensure that the bottom-up saliency map may be used as a mask, highlighting a set of potentially interesting locations in the scene, with top-down influences mainly responsible for deciding upon one specific location among saccade target locations [50].

Acknowledgments This work was supported by the Région Rhône-Alpes via the LIMA project in the context of the cluster ISLE (see http://cluster-isle.grenoble-inp.fr/).

References

1. Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans PAMI. 1998; 20(11):1254–9.
2. Rajashekar U, van der Linde I, Bovik AC, Cormack LK. GAFFE: a gaze-attentive fixation finding engine. IEEE Trans Image Process. 2008; 17(4):564–73.
3. Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned saliency detection model. CVPR: Proc IEEE; 2009. p. 1597–604.
4. Guo CL, Ma Q, Zhang LM. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. Proceedings of IEEE, CVPR, 2008; pp 1–8.
5. Cerf M, Paxon Frady E, Koch C. Faces and text attract gaze independent of the task: Experimental data and computer model. J Vis. 2009;9(12):1–15.
6. Cerf M, Harel J, Einhäuser W, Koch C. Predicting human gaze using low-level saliency combined with face detection. In Platt JC, Koller D, Singer Y, Roweis S, editors. Adv Neural Inf Process Syst 2007;20.
7. Li L-J, Fei-Fei L. What, where and who? Classifying event by scene and object recognition**.** IEEE Int Conf Comput Vis (ICCV); 2007.
8. Scassellati B. Theory of mind for a humanoid robot. Autonom Robots. 2002;12(1):13–24.
9. Marat S, Ho Phuoc T. Spatio-temporal saliency model to predict eye movements in video free viewing. 16th European Signal Processing Conference EUSIPCO-2008, Lausanne: Suisse; 2008.
10. Ma Y, Zhang H. A model of motion attention for video skimming. Proceedings of IEEE, ICIP, Vol. 1, pp. 22–25; 2002.
11. Shan L, Lee MC. Fast visual tracking using motion saliency in video. Proceedings of IEEE, ICASSP. Vol. 1, pp. 1073–1076; 2007.
12. Peters RJ, Itti L. Beyond bottom-up: incorporating task-dependent influences into a computational model of spatial attention. In: Proceedings of IEEE, CVPR; 2007, p. 1–8.
13. Schütz AC, Braun DI, Gegenfurtner KR. Object recognition during foveating eye movement. Vis Res. 2009;49:2241–53.
14. Zhang L, Tong M, Cottrell G. SUNDAy: Saliency using natural statistics for dynamic analysis of scenes. In: Proceedings of the 31st annual cognitive science conference. Netherlands: Amsterdam; 2009.
15. Itti L, Baldi P. Bayesian surprise attracts human attention. Vis Res. 2009;49(10):1295–306.
16. Seo HJ, Milanfar P. Static and space-time visual saliency detection by self-resemblance. J Vis. 2009;9(12):1–27.
17. Mahadevan V, Vasconcelos N. Spatiotemporal saliency in dynamic scenes. IEEE Trans Pattern Anal Mach Intell. 2010;32(1): 171–7.
18. Sevilmis T, Bastan M, Gudukbay U, Ulusoy O. Automatic detection of salient objects and spatial relations in videos for a video database system. Image Vis Comput. 2008;26(10):1384–96.
19. Li LJ, Socher R, Fei-Fei L. Towards total scene understanding: classification, annotation and segmentation in an automatic framework. Comput Vis Pattern Recogn (CVPR); 2009.
20. Yazhou L, Yao H, Wengao, Chen X, Zhao D. Non parametric background generation. J Vis Commun Image Represent. 2007;18:253–63.
21. Wang H, Suter D. A novel robust statistical method for background initialization, visual surveillance. ACCV 2006, LNCS. 2006;3851:328–37.
22. Cotsaces C, Nikolaidis N, Pitas I. Video shot boundary detection and condensed representation: a review. IEEE Signal Process Mag 2006;23(2):28–37.
23. Lu S, King I, Lyu MR. Video summarization by video structure analysis, graph optimization", IEEE International Conference on Multimedia, Expo, 2004. ICME '04. 2004;3:1959–62.
24. Money AG, Agius H. Video summarization: a conceptual framework and survey of the state of the art. J Vis Commun Image R 2008;19:121–43.
25. Carmi R, Itti L. The role of memory in guiding attention during natural vision. J Vis. 2006;6(9):898–914.
26. Pinson M, Wolf S. Comparing subjective video quality testing methodologies. In: Proceedings of SPIE, VCIP, Lugano, Switzerland; 2003.
27. Zivkovic Z. Improved adaptive Gaussian mixture model for background subtraction. Proc Int Conf Pattern Recogn; 2004.
28. Stauffer C, Grimson WEL. Adaptive background mixture models for real-time tracking. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR'99), vol 2. 1999; p. 2246.
29. Elgammal A, Duraiswami R, Harwood D, Davis LS. Background and foreground modeling using nonparametric kernel density for visual surveillance. Proc IEEE. 2002;90(7):1151–63.
30. Sidibé D, Strauss O. A fast and automatic background generation method from a video based on QCH. J Visual Commun Image Represent; 2009.

31. Cucchiara R, Grana C, Piccardi M, Prati A. Detecting moving objects, ghosts, and shadows in video streams. IEEE Trans Pattern Anal Mach Intell. 2003;25(10):1337–42.
32. Lipton AJ, Haering N, Almen MC, Venetianer PL, Slowe TE, Zhang Z. Video scene background maintenance using statistical pixel modeling. United States Patent Application Publication. Pub. No.: US 2004/0126014 A1; 2004.
33. Liu T, Sun J, Zheng NN, Tang X, Shum HY. Learning to detect a salient object. Proc IEEE, CVPR. 2007; p. 1–8.
34. Rutishauser U, Walther D, Koch C, Perona P. Is bottom-up attention useful for object recognition?. Proc IEEE, CVPR. 2004; p. 37–44.
35. Tseng PH, Carmi R, Cameron IGM, Munoz DP, Itti L. Quantifying center bias of observers in free viewing of dynamic natural scenes. J Vis. 2009;9(7):1–16.
36. Gao D, Mahadevan V, Vasconcelos N. On the plausibility of the discriminant center-surround hypothesis for visual saliency. J Vis. 2008;8(7):1–18.
37. Desimone R, Albright TD, Gross CG, Bruce C. Stimulus selective properties of inferior temporal neurons in the macaque. J Neurosci. 1984;4:2051–62.
38. Koch WD. Modeling attention to salient proto-objects. Neural Netw. 2006;19:1395–407.
39. Tampere Image Database (TID). 2008. Page: http://www.ponomarenko.info/tid2008.htm.
40. Ma YF, Zhang HJ. A new perceived motion based shot content representation. Proc IEEE, ICIP. 2001;3:426–9.
41. Jacobson N, Lee YL, Mahadevan V, Vasconcelos N, Nguyen TQ. Motion vector refinement for FRUC using saliency and segmentation. IEEE Trans Image Process (in print); 2010.
42. Belardinelli A, Pirri F, Carbone A. Motion Saliency maps from spatiotemporal filtering. Lecture Notes In: Artificial intelligence, attention in cognitive systems: 5th international workshop on attention in cognitive systems, WAPCV 2008 Fira, Santorini, Greece, p. 112–23; 2008.
43. Ma Q, Zhang L. Saliency-based image quality assessment criterion. Proceedings of ICIC 2008, LNCS 5226, p. 1124–33; 2008.
44. Stalder S, Grabner H, Van Gool L. Beyond semi-supervised tracking: tracking should be as simple as detection, but not simpler than recognition. In: Proceedings ICCV'09 WS on online learning for computer vision; 2009.
45. Babenko B, Yang M, Belongie S. Visual tracking with online multiple instance learning. IEEE conference on computer vision and pattern recognition (CVPR), Miami; 2009.
46. Avidan S. Ensemble tracking. PAMI. 2007;29(2):261–71.
47. Collins R, Liu Y, Leordeanu M. Online selection of discriminative tracking features. PAMI. 2005;27(10):1631–43.
48. Jost T, Ouerhani N, von Wartburg R, Müri R, Hügli H. Assessing the contribution of color in visual attention. Comput Vis Image Underst. 2005;100(1):107–23.
49. Harel J, Koch C, Perona P. Graph-based visual saliency. Proceedings of NIPS, p. 545–52; 2007.
50. Itti L. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Vis Cogn. 2005;12(6):1093–123.
51. Peters RJ, Iyer A, Itti L, Koch C. Components of bottom-up gaze map allocation in natural images. Vis Res. 2005; 45:2397–416.
52. Marat S, Phuoc T, Granjon L, Guyader N, Pellerin D, Guerin-Dugue A. Modelling spatiotemporal saliency to predict gaze direction for short videos. Int J Comput Vis. 2009;82:231–43.
53. Tong Y, Konik H, Cheikh FA, Tremeau A. Multi-feature based visual saliency detection in surveillance video. IEEE, VCIP; 2010 (accepted).