Visual Attention Detection in Video Sequences Using Spatiotemporal Cues

Yun Zhai
University of Central Florida
Orlando, Florida 32816
yzhai@cs.ucf.edu

Mubarak Shah
University of Central Florida
Orlando, Florida 32816
shah@cs.ucf.edu

ABSTRACT

Human vision system actively seeks interesting regions in images to reduce the search effort in tasks, such as object detection and recognition. Similarly, prominent actions in video sequences are more likely to attract human's first sight than their surrounding neighbors. In this paper, we propose a spatiotemporal video attention detection technique for detecting the attended regions that correspond to both interesting objects and actions in video sequences. Both spatial and temporal saliency maps are constructed and further fused in a dynamic fashion to produce the overall spatiotemporal attention model. In the temporal attention model, motion contrast is computed based on the planar motions (homography) between images, which is estimated by applying RANSAC on point correspondences in the scene. To compensate the non-uniformity of spatial distribution of interest-points, spanning areas of motion segments are incorporated in the motion contrast computation. In the spatial attention model, we have developed a fast method for computing pixel-level saliency maps using color histograms of images. A hierarchical spatial attention representation is established to reveal the interesting points in images as well as the interesting regions. Finally, a dynamic fusion technique is applied to combine both the temporal and spatial saliency maps, where temporal attention is dominant over the spatial model when large motion contrast exists, and vice versa. The proposed spatiotemporal attention framework has been extensively applied on several video sequences, and attended regions are detected to highlight interesting objects and motions present in the sequences with very high user satisfaction rate.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding, Perceptual Reasoning.; I.4.10 [**Image Processing and Computer Vision**]: Image Representation.

General Terms

Algorithms.

Keywords

Video Attention Detection, Spatiotemporal Saliency Map.

1. INTRODUCTION

How to achieve a meaningful video representation becomes an interesting problem in various research communities, such as multimedia processing, computer vision and content-based image and video retrieval. The effectiveness of the representation is determined by how well it fits to human perception and reaction to external visual signals. Human perception tends to firstly pick the regions in the imagery that stimulate the vision nerves the most before continuing to interpret the rest of the scene. These attended regions could correspond to either prominent objects in the image or interesting actions in video sequences. Visual attention analysis simulates this human vision system behavior by automatically producing saliency maps of the target image or video sequence. It has a wide range of applications in tasks of image/video representation, object detection and classification, activity analysis, small-display device control and robotics controls. Visual attention deals with detecting the regions of interest (ROI) in images and interesting actions in video sequences that are the most attractive to viewers. For example, in the task of object/action detection, visual attention detection significantly narrows the search range by giving a hierarchical priority structure of the target image or sequence. Consider the following scenario, a video sequence is captured by a camera that is looking at a classroom entrance. At the time the class is dismissed, the majority of the students will be going out of the classroom. In this situation, if two people are trying to walk back into the room, their actions would be considered "irregular" compared to the rest of the students. Attention analysis is able to quickly highlight the abnormal regions and perform further activity analysis on these regions.

1.1 Related Work

Visual attention detection in still images has been long studied, while there is not much work on the spatiotemporal attention analysis. Psychology studies suggest that human vision system perceives external features separately (Treisman and Gelade [25]) and is sensitive to the difference between the target region and its neighborhood (Duncan and Humphreys [6]). Following this suggestion, many works have focused on the detection of feature contrasts to trigger human vision nerves. This is usually referred as the "stimuli-driven" mechanism. Itti *et al.* [10] proposed one of the earliest works in visual attention detection by utilizing

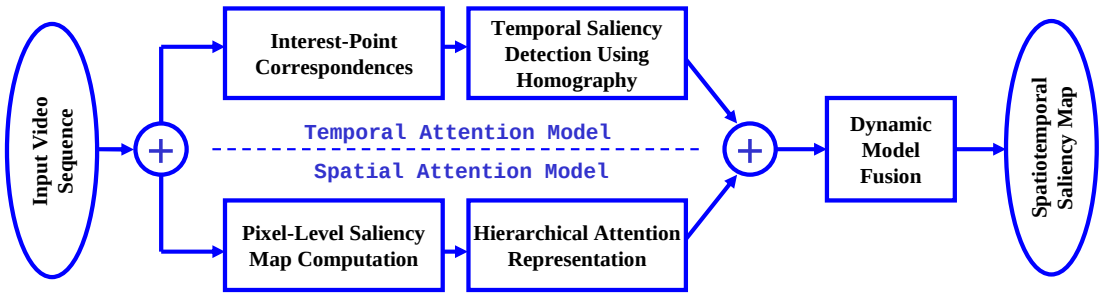


Figure 1: Work flow of the proposed spatiotemporal attention detection framework. It consists of two components, temporal attention model and spatial attention model. These two models are combined using a dynamic fusion technique to produce the overall spatiotemporal saliency maps.

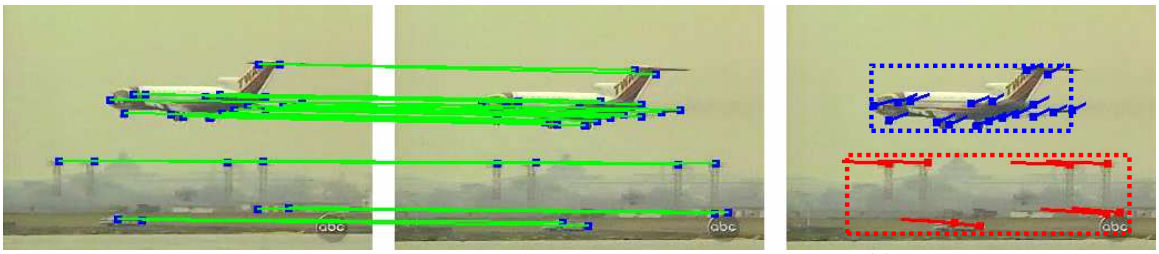
the contrasts in color, intensity and orientation of images. Han *et al.* [8] formulated the attended object detection using the Markov random field with the use of visual attention and object growing. Ma and Zhang [15] incorporated a fuzzy growing technique in the saliency model for detecting different levels of attention. Lu *et al.* [14] used the low-level features, including color, texture and motion, as well as cognitive features, such as skin color and faces, in their attention model. Different types of images have also been exploited. Ouerhani and Hugli [19] has proposed an attention model for range images using the depth information. Besides the heavy investigation using the stimuli-driven approach, some methods utilize the prior knowledge on what the user is looking for. Milanese *et al.* [16] constructed the saliency map based on both low-level feature maps and object detection outputs. Oliva *et al.* [18] analyzed the global distributions of low-level features to detect the potential locations of target objects. A few researchers have extended the spatial attention to video sequences where motion plays an important role. Cheng *et al.* [4] has incorporated the motion information in the attention model. The motion attention model analyzes the magnitudes of image pixel motion in horizontal and vertical directions. Bioman and Irani [2] have proposed a spatiotemporal irregularity detection in videos. In this work, instead of using read motion information, textures of 2D and 3D video patches are compared with the training database to detect the abnormal actions present in the video. Meur *et al.* [12] proposed a spatiotemporal model for visual attention detection. Affine parameters were analyzed to produce the motion saliency map.

Visual attention modelling has been applied in many fields. Baccon *et al.* [1] has proposed an attention detection technique to select spatially relevant visual information to control the orientation of a mobile robot. Driscoll *et al.* [5] has built a pyramidal artificial neural network to control the fixation point of a camera head by computing the 2D saliency map of the environment. Chen *et al.* [3] has applied the visual attention detection technique in devices with small displays. Interesting regions with high saliency values have higher priority to be displayed comparing to the rest of the image. Attention models were used in image compression tasks by Ouerhani *et al.* [20] and Stentiford [24], where regions with higher attention values were compressed with higher reconstruction quality. Peters and Sullivan [22] has applied visual attention in computer graphics to generate the gaze direction of virtual humans.

1.2 Proposed Framework

Video attention methods are generally classified into two categories: *top-down approaches* and *bottom-up approaches*. Methods in the first category, *top-down approaches*, are task-driven, where prior knowledge of the target is known before the detection process. This is based on the cognitive knowledge of the human brain, and it is a spontaneous and voluntary process. Traditional rule-based or training-based object detection methods are the examples in this category. On the other hand, the second category, *bottom-up approaches*, are usually referred as the stimuli-driven techniques. This is based on the human reaction to external stimuli, such as bright color, distinctive shape or unusual motion, and it is a compulsory process.

In this paper, we propose a bottom-up approach for modelling the spatiotemporal attention in video sequences. The proposed technique is able to detect the attended regions as well as attended actions in video sequences. Different from the previous methods, most of which are based on the dense optical flow fields, our proposed temporal attention model utilizes the interest point correspondences and the geometric transformations between images. In this model, feature points are firstly detected in consecutive video images, and correspondences are established between the interest-points using the Scale Invariant Feature Transformation (SIFT [13]). RANSAC algorithm is then applied on the point correspondences to find the moving planes in the sequence by estimating their homographies. Projection errors of the interest points by the estimated homographies are incorporated in the motion contrast computation. In the spatial attention model, we have constructed a hierarchical saliency representation. A linear time algorithm is developed to compute pixel-level saliency maps. In this algorithm, color statistics of the images are used to reveal the color contrast information in the scene. Given the pixel-level saliency map, attended points are detected by finding the pixels with the local maxima saliency values. The region-level attention is constructed based upon the attended points. Given an attended point, a unit region is created with its center to be the point. This region is then iteratively expanded by computing the expansion potentials on the sides of the region. Rectangular attended regions are finally achieved. The temporal and spatial attention models are finally combined in a dynamic fashion. Higher weights are assigned to the temporal model if large motion contrast is present in the sequence. Otherwise, higher weights are assigned to the spatial model



(a) Image 1 (b) Image 2 (c) Motion Regions

Figure 2: One example of the point matching and motion segmentation results. Figure (a) and figure (b) show two consecutive images. The interest points in both images and their correspondences are presented. The motion regions are shown in figure (c).

if less motion exists. The work flow of the proposed attention detection framework is described in Figure 1. To demonstrate the effectiveness of the proposed spatiotemporal attention framework, we have extensively applied it to many video sequences, which contain both sequences with moving objects and sequences with uniform global motion. Very satisfactory results have been obtained and presented in the paper.

The remainder of this paper is organized as follows: The temporal and spatial attention models are presented in Section 2 and Section 3, respectively. Corresponding intermediate results are also presented. Section 4 describes the proposed dynamic fusion method to combine the two individual attention models. Section 5 presents the experimental results and the performance evaluations. Finally, Section 6 concludes our work.

2. TEMPORAL ATTENTION MODEL

In the temporal attention detection, saliency maps are often constructed by computing the motion contrast between image pixels. Most of the previously developed methods generate dense saliency maps based on pixel-wise computations, mostly dense optical flow fields. However, it is well known that optical-flows at edge pixels are noisy if multiple motion layers exist in the scene. Furthermore, dense optical flows maybe erroneous in regions with less texture. In contrast, point correspondences (also known as the sparse optical flows) between images are comparatively accurate and stable. In this section, we propose a novel approach for computing the temporal saliency map using the point correspondences in video sequences. The proposed temporal saliency computation utilizes the geometric transformations between images, which model the planar motions of the moving segments.

Given images in a video sequence, feature points are localized in each image using the point detection method. Correspondences between the matching points in consecutive frames are further established by analyzing the properties of image regions around the points. In our framework, we have applied the Scale Invariant Feature Transformation (SIFT [13]) operator to find the interest points and compute the correspondences between points in video frames. Let $\mathbf{p}_m = (x_m, y_m)$ be the m-th point in the first image and $\mathbf{p}'_m = (x'_m, y'_m)$ be its correspondence in the second image. One example of the interesting point matching is shown in Figure 2. Given the point correspondences, the temporal saliency value $SalT(\mathbf{p}_i)$ of point $\mathbf{p}_i$ is computed by modelling the motion contrast between the target point and other points,

$$SalT(\mathbf{p}_i) = \sum_{j=1}^{n} DistT(\mathbf{p}_i, \mathbf{p}_j), \tag{1}$$

where n is the total number of correspondences. $DistT(\mathbf{p}_i, \mathbf{p}_j)$ is some distance function between $\mathbf{p}_i$ and $\mathbf{p}_j$. In our formulation, we analyze the geometric transformations between images. The motion model used is homography. Homography is used for modelling the planar transformations. The interesting point $\mathbf{p} = [x, y, 1]^T$ and its correspondence $\mathbf{p}' = [x', y', 1]^T$ can be associated by,

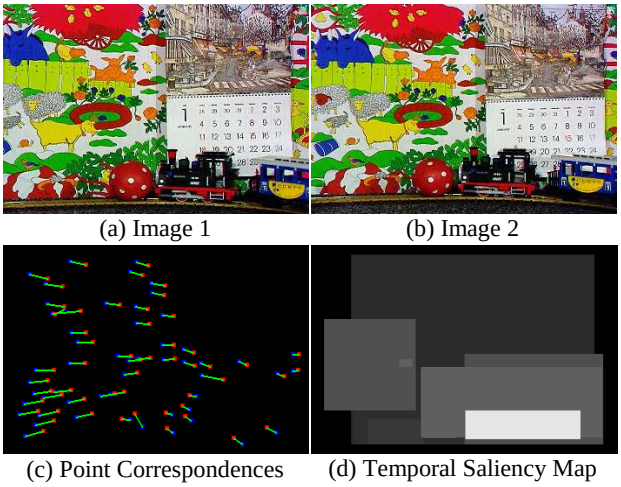
$$\begin{bmatrix} \hat{x}' \\ \hat{y}' \\ \hat{t}' \end{bmatrix} = \begin{bmatrix} a_1 & a_2 & a_3 \\ a_4 & a_5 & a_6 \\ a_7 & a_8 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{2}$$

Here, $\hat{\mathbf{p}}' = [\hat{x}', \hat{y}', \hat{t}']^T$ is the projection of $\mathbf{p}$ in the form of homogeneous coordinates. Parameters $\{a_i, i = 1, \cdots, 8\}$ capture the transformation between two matching planes, and they can be estimated by providing at least four pairs of correspondences. For simplicity, we use $\mathbf{H}$ to represent the transformation matrix in the rest of the text. Also, we normalize $\hat{\mathbf{p}}_i$, such that its their element is 1. Ideally, $\hat{\mathbf{p}}'$ should be the same as $\mathbf{p}'$. With noise present in the imagery, a point $\hat{\mathbf{p}}'$ matches with $\mathbf{p}'$ with an error computed after applying $\mathbf{H}$, as,

$$\epsilon(\mathbf{p}_i, \mathbf{H}) = \| \hat{\mathbf{p}}'_i - \mathbf{p}'_i \| . \tag{3}$$

Motions of objects are only meaningful when certain reference is defined. For instance, a car is said "moving" only if visible background is present in the scene and disagrees with the car in terms of the motion direction. This fact indicates that multiple moving objects are in the scene to indicate local motion existence. In these types of situations, a single homography is insufficient to model all the correspondences in the imagery. To overcome this problem, we apply RANSAC algorithm on the point correspondences to estimate multiple homographies that model different motion segments in the scene. The homographies are later used in the temporal saliency computation process.

For each homography $\mathbf{H}_m$ estimated by RANSAC, a list of points $\mathbf{L}_m = \{\mathbf{p}_1^m, \cdots, \mathbf{p}_{n_m}^m\}$ are considered as its inliers, where n_m is the number of inliers for $\mathbf{H}_m$. Given the homographies and the projection error definition in Eqn.3, we can define the motion contrast function in Eqn.1 as,



(a) Image 1 (b) Image 2

(c) Point Correspondences (d) Temporal Saliency Map

Figure 3: Example of the proposed temporal attention model. Figures (a) and (b) show two consecutive images of the input sequence. Figure (c) shows the interest-point correspondences. Figure (d) shows the detected temporal saliency map using the proposed homography-based method. In this example, the camera follows the moving toy train from right to left, while the calender is moving downward. Thus, intuitively, the attention region should correspond to the toy train. The saliency map also suggests that the second attended region corresponds to the moving calender. Brighter color represents higher saliency value.

$$DistT(\mathbf{q}_i, \mathbf{q}_j) = \epsilon(\mathbf{q}_i, \mathbf{H}_m), \tag{4}$$

where $\mathbf{q}_j \in \mathbf{L}_m$. The sizes of the inlier sets play dominant role in the current saliency computation. It is well known that the spatial distribution of the interest points is not uniform due to variance in the texture contents of image parts. Sometimes, relatively larger moving objects/regions may contribute less trajectories, while smaller regions but with richer texture provide more trajectories. One example is shown in Figure 2. In these cases, the current temporal saliency definition is not realistic. Larger regions with less points, which often belong to the backgrounds, will be assigned with higher attention values. While foreground objects, which are supposed to be the true attended regions, will be assigned with lower attention values, if they possess more interest-points. To avoid this problem, we incorporate the spanning area information of the moving regions. The spanning area of a homography $\mathbf{H}_m$ is computed as,

$$\alpha_m = \Big(max(x_i^m) - min(x_i^m)\Big) \times \Big(max(y_i^m) - min(y_i^m)\Big), \tag{5}$$

where $\forall \mathbf{p}_i^m \in \mathbf{L}_m$, and α_i is normalized with respect to the image size, such that $\alpha_i \in [0, 1]$. In the extreme cases, where $max(x_i^m) = min(x_i^m)$ or $max(y_i^m) = min(y_i^m)$, to avoid zero values of α_m, the corresponding term in Eqn.5 is replaced with a non-zero constant number (in the experiment, we use 0.1). The temporal saliency value of a target point $\mathbf{p}$ is finally computed as,

$$SalT(\mathbf{p}) = \sum_{j=1}^{M} \alpha_j \times \epsilon(\mathbf{p}, \mathbf{H}_j), \tag{6}$$

where M is the total number of homographies in the scene. In the degenerated cases, where some point correspondences do not belong to inlier sets of any of the estimated homographies, we apply a simplified form of the homography to each of these point correspondences. Suppose $\{\mathbf{p}_t, \mathbf{p}'_t\}$ is one of the "left-out" correspondences. The transformation is defined as a translation matrix $\mathbf{H}_t = [1\ 0\ d_x^t;\ 0\ 1\ d_y^t;\ 0\ 0\ 1]$, where $d_x^t = x'_t - x_t$ and $d_y^t = y'_t - y_t$, and the inlier set $\mathbf{L}_t = \mathbf{p}_t$.

Up to this point, we have the saliency values of individual points and the spanning regions of the homographies, which correspond to the moving objects in the scene. To achieve object-level attention for $\mathbf{H}_m$, the average of the saliency values of the inliers $\mathbf{L}_m$ is considered as the saliency value of the corresponding spanning region. All the image pixels in the same spanning region have the same saliency value. Since the resulting regions are rectangular, it is likely that an image pixel is covered by multiple spanning regions. In this case, the pixel is assigned with the highest saliency value possible. If the pixel is not covered by any spanning region, its saliency value is set to zero. One example of the proposed temporal saliency map computation is demonstrated in Figure 3, where the camera follows a moving toy train from right to left, and apparently the attention region in the sequence corresponds to the moving toy train.

3. SPATIAL ATTENTION MODEL

When viewers watch a video sequence, they are attracted not only by the interesting events, but also sometimes by the interesting objects in still images. This is referred as the spatial attention. Based on the psychological studies, human perception system is sensitive to the contrast of visual signals, such as color, intensity and texture. Taking this as the underlying assumption, we propose an efficient method for computing the spatial saliency maps using the color statistics of images. The algorithm is designed with a linear computational complexity with respect to the number of image pixels. The saliency map of an image is built upon the color contrast between image pixels. The saliency value of a pixel I_k in an image I is defined as,

$$SalS(I_k) = \sum_{\forall I_i \in I} \parallel I_k - I_i \parallel, \tag{7}$$

where the value of I_i is in the range of $[0, 255]$, and $||\cdot||$ represent the color distance metric. This equation is expanded to have the following form,

$$SalS(I_k) = ||I_k - I_1|| + ||I_k - I_2|| + \cdots + ||I_k - I_N||, \tag{8}$$

where N is the total number of pixels in the image. Given an input image, the color value of each pixel I_i is known. Let $I_k = a_m$, and Eqn.8 is further restructured, such that the terms with the same I_i are rearranged to be together,

$$\begin{aligned} SalS(I_k) &= ||a_m - a_0|| + \cdots + ||a_m - a_1|| + \cdots + \cdots, \\ SalS(a_m) &= \sum_{n=0}^{255} f_n ||a_m - a_n||, \end{aligned} \tag{9}$$

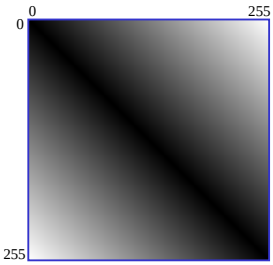


Figure 4: The distance map between the gray-level color values, which can be computed prior to the pixel-level saliency map computation. Brighter elements represent larger distance values.

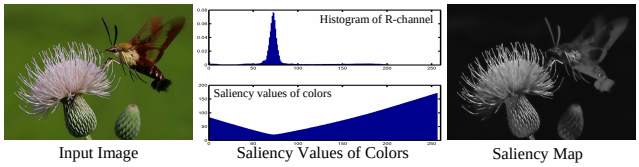


Figure 5: An example of the spatial saliency computation. The left figure shows the input image. The center-top figure shows the histogram of the R-channel of the image, while the center-bottom figure shows the saliency values of the colors. The horizontal axis represents the values of the colors, where $a_n \in [0, 255]$. The saliency values are close to what human expects, since higher frequency indicates repeating information in the image, and therefore, relatively unattractive. The right figure shows the resulting spatial saliency map.

where f_n is the frequency of pixel value a_n in the image. The frequencies are expressed in the form of histograms, which can be computed in $O(N)$ time order. Since $a_n \in [0, 255]$, the color distance metric $\parallel a_m - a_n \parallel$ is also bounded in the range of $[0, 255]$. Since this is a fixed range, a distance map D can be constructed in constant time prior to the saliency map computation. In this map, element $D(x, y) = \parallel a_x - a_y \parallel$ is the color difference between a_x and a_y. One color difference map is shown in Figure 4. Given the histogram $f_{(\cdot)}$ and the color distance map $D(\cdot, \cdot)$, the saliency value for a pixel I_k is computed as,

$$SalS(I_k) = SalS(a_m) = \sum_{n=0}^{255} f_n D(m, n), \qquad (10)$$

which executes in a constant time order. Thus, instead of computing the saliency values of all the image pixels using Eqn.7, only the saliency values of colors $\{a_i, i = 0, \cdots, 255\}$ are necessary for the generation of the final saliency map. One example of the pixel-level spatial saliency computation is shown in Figure 5.

Greatly inspired by the work presented in [15], we propose a hierarchical representation for the spatial attention model based on the pixel-level saliency map computed previously. Two levels of attentions are achieved: attended points and attended regions. Attended points are analogous to the direct response of human perception system to external signals. They are computed as the image pixels with the locally maximum spatial saliency values. On the other hand,

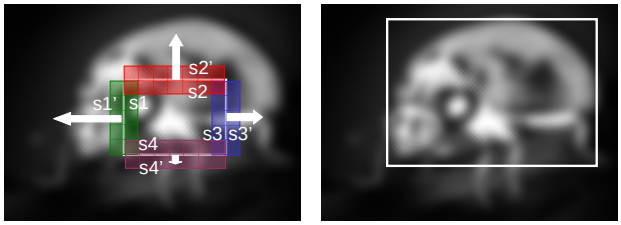


Figure 6: An example of the attended region expansion using the pixel-level saliency map. A seed region is created on the left. Expanding potentials on all four sides of the attended region are computed (shaded regions). The lengths of the arrows represent the strengths of the expansions on the sides. The final attended region is shown on the right.

region-level attention representation provides attended objects in the scene. One simple way to achieve the attended regions is to apply the connected-component algorithm to find the bright regions. However, as shown in Figure 5, pixels with low attention values are embedded in high-value regions. Connected-component algorithm will fail to include these pixels in the attended region. Furthermore, connected-component method tends to generate over-detection of the attended regions. In this paper, we present a region growing technique for detecting the attended regions, which is able to resolve the above mentioned problems. In our formulation, the attended regions are firstly initialized based on the attended points computed previously. Given an attended point $\mathbf{c}$, a rectangular box centered at $\mathbf{c}$ with the unit dimensions is created as the seed region $\mathbf{B_c}$. The seed region is then iteratively expanded by moving its sides outward by analyzing the energy around its sides. The attended region expansion algorithm is described as follows,

1. For each side $i \in \{1, 2, 3, 4\}$ of region $\mathbf{B}$ with length l_i, two energy terms $E(s_i)$ and $E(s_i')$ are computed for both its inner and outer sides s_i and s_i', respectively, as shown in Figure 6. The potential for expanding side i outward is defined as follows,

$$EP(i) = \frac{E(s_i)E(s_i')}{l_i^2}, \qquad (11)$$

   where l_i^2 is for the purpose of normalization.

2. Expand the region by moving side i outward with a unit length if $EP(i) > Th$, where Th is the stopping criteria for the expansion. In the experiment, the unit length is 1 pixel.

3. Repeat steps 1 and 2 until no more side of $\mathbf{B}$ can be further expanded, i.e., all the corresponding expansion potentials are below the defined threshold.

It should be noted that the expansion potential defined in Eqn.11 is designed in such a way, that the attended region is expanded if and only if both the inner and outer sides have high attention values. The expansion stops at the boundary between the high value regions representing the interesting objects and the low value regions for the background. A demonstration of the expanding process is shown in Figure 6.

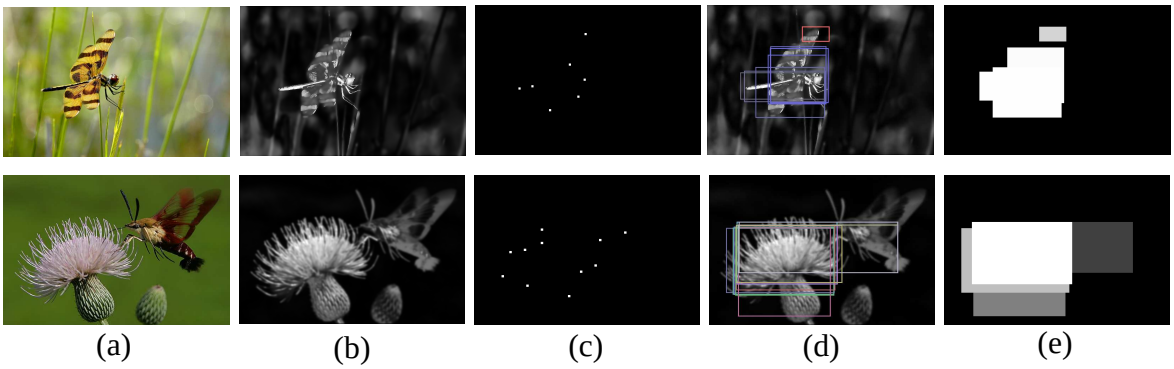


Figure 7: The results of spatial attention detection on two testing images. Column (a) shows the input images; column (b) shows the pixel-level spatial saliency maps; column (c) presents the detected attention points; column (d) shows the expanding boxes from the attention points in (c); finally, column (e) shows the region-level saliency maps of the images.

It is possible that the attended regions initiated using different attended points eventually cover the same image region. In this case, a region merging technique is applied to merge the attended regions that cover the same target image region by analyzing the overlapping ratio between the regions. To be consistent with the temporal attention model, the final spatial saliency map reveals the attended regions in the rectangular shapes. Detailed results of the spatial attention detection on two images are shown in Figure 7.

4. DYNAMIC MODEL FUSION

In the previous sections, we presented the temporal and spatial attention models separately. These two models need to collaborate in a meaningful way to produce the final spatiotemporal video saliency maps. Psychological studies reveal that, human vision system is more sensitive to motion contrast compared to other external signals. Consider a video sequence, in which the camera is following a person walking, while the background is moving in the opposite direction of the camera's movement. In general, humans are more interested in the followed target, the walking person, instead of the his surrounding regions, the background. In this example, motion is the prominent cue for the attention detection compared to other cues, such as color, texture and intensity. On the other hand, if camera is being static or only scanning the scene, in which motion is relatively uniform, then the human perception system is attracted more by the contrasts caused by other visual stimuli, such as color and shape. In summary, we propose the following criteria for the fusion of temporal and spatial attention models,

1. If strong motion contrast is present in the sequence, temporal attention model should be more dominant over the spatial attention model.
2. On the other hand, if the motion contrast is low in the sequence, the fused spatiotemporal attention model should incorporate the spatial attention model more.

Based on these two criteria, simple linear combination with fixed weights between two individual models is not realistic and would produce unsatisfactory spatiotemporal saliency maps. Rather, we propose a dynamic fusion technique, which satisfies the aforementioned criteria. It gives

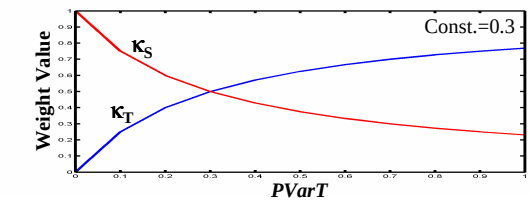


Figure 8: Plots of the dynamic weights, κ_T and κ_S, with respect to the pseudo-variance $PVarT$ of the temporal saliency map, where $Const = 0.3$. As it is clear in the figure, the fusion weight of the temporal attention model increases with the pseudo-variance of the temporal saliency values.

a higher weight to the temporal attention model, if high contrast is present in the temporal saliency map. Similarly, it gives a higher weight to the spatial model, if the motion contrast is relatively low.

Finally, the spatiotemporal saliency map of an image I in the video sequence is constructed as,

$$Sal(I) = \kappa_T \times SalT(I) + \kappa_S \times SalS(I), \qquad (12)$$

where κ_T and κ_S are the dynamic weights for the temporal and spatial attention models, respectively. These dynamic weights are determined in terms of the variance of $SalT(I)$. One special situation needs to be considered carefully. Consider one scene with a moving object whose size is relatively small compared to the background. The variance of the temporal saliency map in this case would be low by the overwhelming background saliency values and does not truly reflect the existence of the moving object. In this case, we compute a variance-like measure, *pseudo-variance*, which is defined as $PVarT = max(SalT(I)) - median(SalT(I))$. The weights κ_T and κ_S are then defined as,

$$\kappa_T = \frac{PVarT}{PVarT + Const}, \quad \kappa_S = \frac{Const}{PVarT + Const}, \qquad (13)$$

where $Const$ is a constant number. From Eqn.13, if the motion contrast is high in the temporal model, then the value of $PVarT$ increases. Consequently, fusion weight of

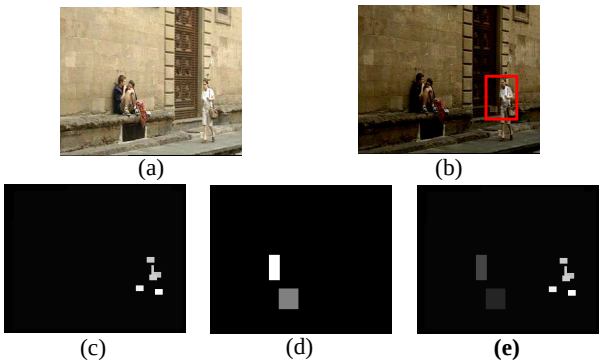


Figure 9: An example of model fusion. This sequence shows two people sitting and another person walking in front of them. (a) is the key-frame of the input sequence. (c) shows the temporal saliency map based on the planar transformations. (d) shows the region-level spatial saliency map using color contrast. (e) is the combined spatiotemporal saliency map. Obviously, the moving object (the walking person) catches more attention than the still regions (sitting persons). Thus, it is assigned higher attention values. The region that corresponds to the interesting action in the scene is shown in (b).

the temporal model, κ_T, is also increased, while the fusion weight of the spatial model, κ_S, is decreased. The plots of κ_T and κ_S with respect to $PVarT$ are shown in Fig.8. One example of the spatiotemporal attention detection is shown in Fig.9, which shows a person is walking in front of the two sitting people. The moving object (walking person) is highlighted by the detected attention region.

5. PERFORMANCE EVALUATION

To demonstrate the effectiveness of the propose spatiotemporal attention model, we have extensively applied the method on two types of video sequences, labelled Testing Set 1 and Testing Set 2. The testing sequences are obtained from feature films and television programs. Testing Set 1 contains nine video sequences, each of which has one object moving in the scene, such as moving cars and flying airplanes. The detailed results of Testing Set 1 are shown in Figure 10. The following information is presented: the representative frames of the testing videos (Figure 10(a)), the temporal saliency maps of the representative frames (Figure 10(b)), the spatial saliency maps of the representative frames (Figure 10(c)), the final spatiotemporal saliency maps (Figure 10(d)) and the detected regions that correspond to the prominent actions in the videos (Figure 10(e)). It should be noted that, for those videos that have very rich texture, the spatial attention model generates meaningless saliency maps. However, with the help of the proposed dynamic model fusion technique, the temporal attention model becomes dominant and is able to detect the regions where interesting actions happen (as shown in Figure 10(e)). This exactly fits human perceptual reactions to motion contrast in these types of situations regardless of visual texture in the scene.

The second testing set, Testing Set 2, contains video sequences without prominent motions. The videos are mainly focusing on the static scene settings or with uniform global motions, i.e., there is no motion contrast in the scene. In this

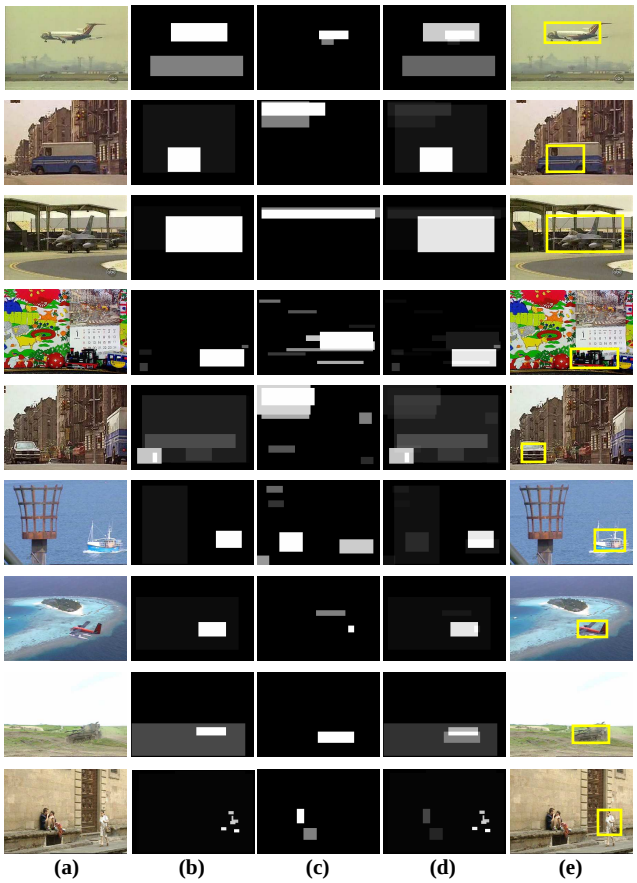


Figure 10: Spatiotemporal attention detection results for the testing videos in Testing Set 1. Column (a) shows the representative frames of the videos; column (b) shows the temporal saliency maps; column (c) shows the spatial saliency maps; column (d) shows the fused spatiotemporal saliency maps; and column (e) shows the regions that correspond to potential interesting actions in clips. It should be noted that when rich texture exists in the scene, temporal attention model is able to detect the attended regions using motion information, while the spatial model fails.

case, the spatial attention model should be dominant over the temporal model. Some results on the testing sequences in Testing Set 2 are shown in Figure 11. The presented results include the following: the representative key-frames of videos (Figure 11(a)), the pixel-level spatial saliency maps (Figure 11(b)), the expanding regions (Figure 11(c)), the attended points in the representative frames (Figure 11(d)) and the attended regions in the representative frames (Figure 11(e)). Temporal saliency maps are not shown since they are uniform and carry less information.

Assessing the effectiveness of a visual attention detection method is a very subjective task. Therefore, manual evaluation by humans is an important and inevitable element in the performance analysis. In our experiments, we have invited five assessors with both computer science and non-computer science backgrounds to evaluate the performance of the proposed spatiotemporal attention detection framework. Borrowing the evaluation ideas from [15], each assessor is asked

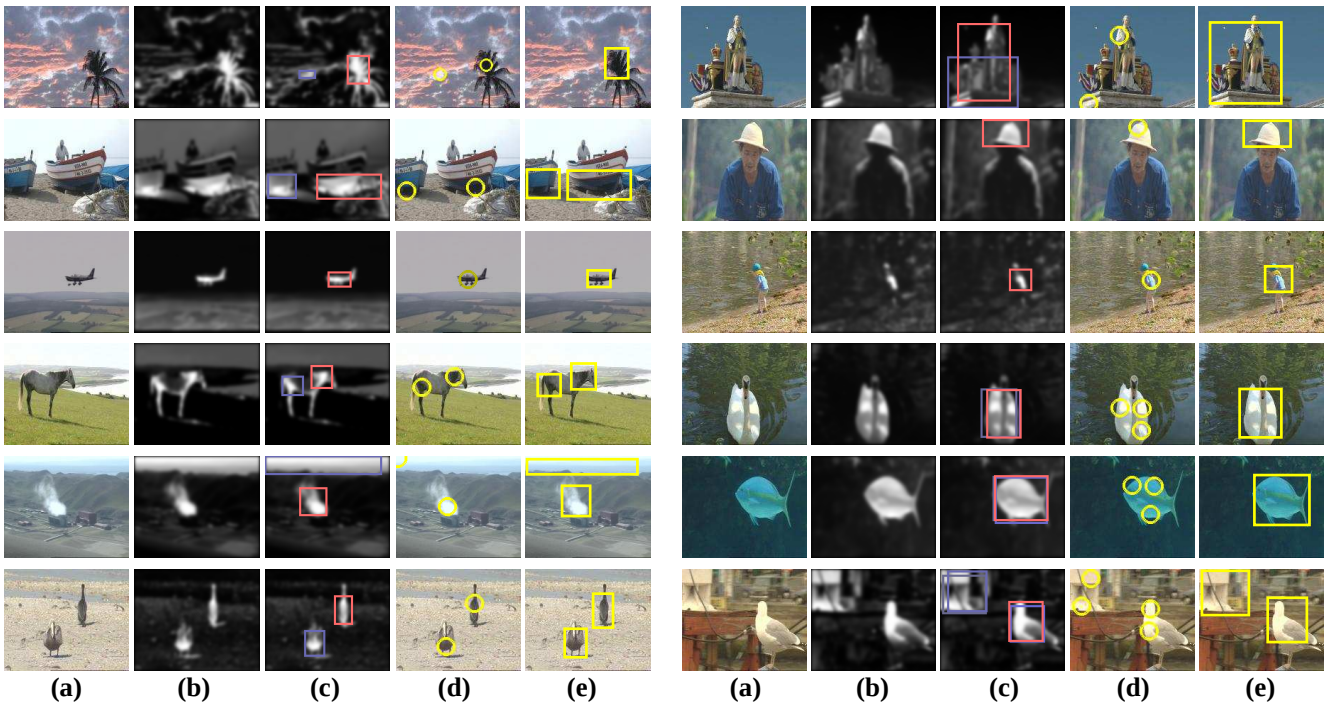


Figure 11: Spatiotemporal attention detection results for the testing videos in Testing Set 2. Only the spatial saliency maps are shown, since there is no motion contrast in the scene. Column (a) shows the representative frames of the videos; column (b) shows the pixel-level spatial saliency maps; column (c) shows the extended bounding boxes using the expansion method proposed in Section 3; column (d) shows the detected attended points in the representative frames; and column (e) shows the detected attended regions in the images. Corresponding evaluation results are shown in Figure 12. Note that column (e) shows different information from column (c). If the extended bounding boxes overlaps (second example in the first row), they are merged to produce a single attended region in the scene. Also, small bounding boxes are removed in the attended region generation.

to give a vote on how satisfactory he or she thinks the detected attended region is for each testing sequence. There are three types of satisfactions, *good*, *acceptable* and *failed*. *Good* represents the situations where the detected attended regions/actions exactly match what the assessor thinks. As pointed out by [15], it is somehow difficult to define the *acceptable* cases. The reason is that different assessors have different views even for the same video sequence. One attended region considered inappropriate by one assessor may be considered perfect to another. In our experimental setup, if the detected attended regions in a video sequence do not cover the most attractive regions, but instead cover less interesting regions, the results are considered *acceptable*. As described by this definition, being *acceptable* is subjective to individual assessors. For instance, in the last example in Figure 10, one assessor considers the walking person is more interesting than the other two sitting people. Then, the current results shown is considered *good* to this assessor. However, another assessor may be attracted by the sitting people the most, then by the walking person. In this case, the current result is considered *acceptable* to the second assessor.

We have performed the evaluation on both testing sets with three categories: (1) Testing Set 1 with moving objects in the scene; (2) Testing Set 2 with detected attended points; and (3) Testing Set 2 with detected attended regions in the scene. Figure 12 shows the assessment of all three categories. In this result table, element in row M and column N represents the proportion of the votes on category M with satisfactory level N. The assessment shown in the table demonstrates that the proposed spatiotemporal attention detection framework is able to discover the interesting objects and actions with more than 90% satisfaction rates. The results of attended point detection have a lower satisfaction rate than the other two region-level attention representations. The reason for this is that, the attended regions possess contextual information among image pixels, and therefore, have richer contents in terms of semantic meanings than image pixels. On the other hand, attended points are isolated from each other, and human perception system responds to them very differently for different persons. Due to the lack of semantic meanings, more disagreements between assessors emerged for the detected attended points, and therefore, has lowered the satisfactory score.

Another interesting observation from the experiments is that, as the texture content in the imagery becomes richer, the attention detection performs with lower satisfactory rate. This is clearly shown in the results (Figure 11). For videos that have prominent objects with relatively plain background settings, the proposed attention detection method performs well and produces very satisfied attended regions. On the other hand, if the background settings are much richer, some false detections are generated. This is actually a good simulation to the human vision system. As pointed by the psy-

System Performance Evaluation			
Data Set	Good	Acceptable	Failed
Testing Set 1 (Moving Objects)	0.82	0.16	0.02
Testing Set 2 (Attended Points)	0.70	0.12	0.18
Testing Set 2 (Attended Regions)	0.80	0.12	0.08

Figure 12: System performance evaluation for three categories, *Testing Set 1 with moving objects*, *Testing Set 2: attended point detection* and *Testing Set 2: attended region detection*.

chological studies in [6], human vision system is sensitive to the difference or contrast between the target region and its neighborhood. In the situations where the background settings are relatively uniform, the contrast between the object and the background is larger. Thus, human vision system is able to pick up the target region very easily. On the other hand, if rich background settings are present in the scene, the contrast between the object and the background is less comparing to the former cases, human vision system is distracted by other regions in the scene and less capable to find the target object.

6. CONCLUSIONS

In this paper, we have presented a spatiotemporal attention detection framework for detecting both attention regions and interesting actions in video sequences. The saliency maps are computed separately for the temporal and spatial information of the videos. In the temporal attention model, interest-point correspondences and geometric transformations between images are used to compute the motion contrast in the scene. The areas of the spanning regions of the motion groups are incorporated in the motion contrast computation. In the spatial attention model, we have presented a fast algorithm for computing the pixel-level saliency map using the color histograms. A hierarchical attention representation is established. Rectangular attended regions are initialized based on the attended points. They are further iteratively expanded by analyzing the expansion potentials along their sides. To achieve the spatiotemporal attention model, a dynamic fusion technique is applied to combine the temporal and spatial models. The dynamic weights of the two individual models are controlled by the pseudo-variance of the temporal saliency values. Extensive testing has been performed on numerous video sequences to demonstrate the effectiveness of the proposed framework, and very satisfactory results have been obtained.

7. REFERENCES

[1] J.C. Baccon, L. Hafemeister and P. Gaussier, "A Context and Task Dependent Visual Attention System to Control A Mobile Robot ", *ICIRS*, 2002.

[2] O. Boiman and M. Irani, "Detecting Irregularities in Images and in Video", *ICCV*, 2005.

[3] L-Q. Chen, X. Xie, W-Y. Ma, H.J. Zhang and H-Q. Zhou, "Image Adaptation Based on Attention Model for Small-Form-Factor Devices", *ICMM*, 2003.

[4] W-H. Cheng, W-T. Chu, J-H. Kuo and J-L. Wu, "Automatic Video Region-of-Interest Determination Based on User Attention Model", *ISCS*, 2005.

[5] J.A. Driscoll, R.A. Peters II and K.R. Cave, "A visual attention network for a humanoid robot", *ICIRS*, 1998.

[6] J. Duncan and G.W. Humphreys, "Visual Search and Stimulus Similarity", *Psychological Review*, 1989.

[7] L.L. Galdino and D.L. Borges, "A Visual Attention Model for Tracking Regions Based on Color Correlograms", *Brazilian Symposium on Computer Graphics and Image Processing*, 2000.

[8] J. Han, K.N. Ngan, M. Li and H.J. Zhang, "Towards unsupervised attention object extraction by integrating visual attention and object growing", *ICIP*, 2004.

[9] Y. Hu, D. Rajan and L-T. Chia, "Adaptive Local Context Suppression of Multiple Cues for Salient Visual Attention Detection", *ICME*, 2005.

[10] L. Itti, C. Koch and E. Niebur, "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis", *PAMI*, 1998.

[11] T. Kohonen, " A Computational Model of Visual Attention", *IJCNN*, 2003.

[12] O. Le Meur, D. Thoreau, P. Le Callet and D. Barba, "A Spatio-Temporal Model of the Selective Human Visual Attention", *ICIP*, 2005.

[13] D.G. Lowe, "Distinctive Image Features From Scale-Invariant Keypoints", *IJCV*, 2004.

[14] Z. Lu, W. Lin, X. Yang, E.P. Ong and S. Yao, "Modeling visual attentions modulatory aftereffects on visual sensitivity and quality evaluation", *T-IP*, 2005.

[15] Y.F. Ma and H.J. Zhang, "Contrast-Based Image Attention Analysis by Using Fuzzy Growing", *ACM Multimedia*, 2003.

[16] R. Milanese, H. Wechsler, S. Gill, J.-M. Bost and T. Pun, "Integration of Bottom-Up and Top-Down Cues for Visual Attention Using Non-Linear Relaxation ", *CVPR*, 1994.

[17] A. Nguyen, V. Chandrun and S. Sridharan, "Visual Attention Based ROI Maps From Gaze Tracking Data", *ICIP*, 2004.

[18] A. Oliva, A. Torralba, M.S. Castelhano and J.M. Henderson, "Top-Down Control of Visual Attention in Object Detection", *ICIP*, 2003.

[19] N. Ouerhani and H. Hugli, "Computing Visual Attention from Scene Depth", *ICPR*, 2000.

[20] N. Ouerhani, J. Bracamonte, H. Hugli, M. Ansorge and F. Pellandini, "Adaptive Color Image Compression Based on Visual Attention", *ICIAP*, 2001.

[21] O. Oyekoya and F. Stentiford, "Exploring human eye behaviour using a model of visual attention", *ICPR*, 2004.

[22] C. Peters and C.O. Sullivan, "Bottom-Up Visual Attention for Virtual Human Animation", *CASA*, 2003.

[23] K. Rapantzikos, N. Tsapatsoulis and Y. Avrithis, "Spatiotemporal Visual Attention Architecture for Video Analysis", *MMSP*, 2004.

[24] F. Stentiford, "A Visual Attention Estimator Applied to Image Subject Enhancement and Colour and Grey Level Compression", *ICPR*, 2004.

[25] A. Treisman and G. Gelade, "A feature-integration theory of attention", *Cognitive Psychology*, 1980.