OBJECT TRACKING BASED ON TIME-VARYING SALIENCY

Sheng Xu [a,] *, Hong Huo [a], Fang Tao [a]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, No.800 Dongchuan Road,
Shanghai, China - affine@sjtu.edu.cn

KEY WORDS: Object, Tracking, Hierarchical, Matching, Machine vision

ABSTRACT:

Visual attention has been widely used in image pre-processing, since it can rapidly detect the region of interest in the given scene. This paper presents a novel technique to track the moving object, which is based on the motion saliency model. The salient region is computed by the combination of multi-feature maps and motion saliency map, which vastly reduce the amount of information in further image processing. Next, a single matching method, normalized color histogram, is used to measure the similarity for tracking processing. Experimental results, found in AVSS 07, are reported, which validate our model useful and effective.

1. INTRODUCTION

Object tracking in video streams has been one of the most popular topics in computer vision, since it serves as a means to prepare data for surveillance, perceptual user interfaces, object-based video compression, and driver assistance. Tracking over time typically involves matching objects in consecutive frames using points, lines or blobs, based on their motion, shape, and other visual information. That is to say, tracking may be considered to be equivalent to establishing coherent relations of image features between frames with respect to position, velocity, shape, texture, color, and etc(Wang 2003). However, not all information in these frames is necessary, but only some details around objects should be paid attention to. Now, visual attention is the ability to rapidly detect the interesting parts of a given scene, on which higher vision tasks can focus. Thus, only a few parts of the image information are selected for further object tracking.

Simulating human vision system (HVS), visual attention represents a basic tool for computer vision. Until now, the model of computational visual attention has been widely investigated during the last two decades. Most of known computational models(Itti 1998; Kadir 2001; Yee 2002; Itti 2003; Shic 2007; Stentiford 2007) rely on the feature integration theory presented by Treisman(Treisman 1980). The saliency-based model of Koch and Ullman(Koch 1985), which is one of the best of the most prominent theory models of visual attention, was first presented, and gave rise to numerous development and computational implementation. Especially, Itti(Itti 1998) presents one of the most prominent computational models of attention based on Koch model. And many improved Itti models are discussed in recent years. Another important computation model(Stentiford 2007), presented by Stentiford, is implemented by measuring the similarity between pixels and finding the rarity. The model of attention makes the assumption that an image region is salient if it possesses few features in common with other regions in the image, since it is noted that saliency implies rarity. However, as argued by Gilles(Gilles 1998), the converse is not necessarily true. Motivated by the promising results obtained by Gilles, Kadir presents a visual saliency, which investigates the local complexity approach. All above mentioned computation models aimed, however, at extracting the interest region from static scene. Some of the rare visual attention models which take in consideration the motion saliency are presented by Yee and Walther(Yee 2002), Itti et al.(Itti 2003; Itti 2005), Shic and Sassellati(Shic 2007).

This paper reports a computational model of dynamic visual attention which combines static multi-feature maps and motion feature map to detect salient locations in the video sequences. Therefore, the model obtains a motion saliency map (or region of interest) related to static and dynamic feature. The locations, detected by the motion saliency model, can be described as the seed point in further roughly segmentation. Then, tracking problem is posed as a sub-image matching problem. And, the normalised color histogram algorithm is applied to measure the similarity from frame to frame. In this processing, only the normalised color histogram in the segmented salient regions is computed to reduce the amount of information.

The layout of the remainder of this presentation is as follows. The next section will detail the computational saliency model. Related computational models appear in itti(Itti 1998) and shic's(Shic 2007) paper. Following this, a discussion of object matching is provided since visual tracking is based on finding the similar object in the coherent frames. Section 4 will show the example of processing of the object tracking. A concluding discussion rounds out the paper.

2. MODEL OF TIME-VARYING VISUAL ATEETNION

The computational model of dynamic visual attention consists of a static saliency map which discriminates salient scene locations based on the static early vision features and motion feature that pops out motion phenomena. The two salient maps are then combined into a final time-varying saliency map. The relational diagram for the dynamic saliency model is shown in Figure. 1.

* Corresponding author. Tao Fang. E-mail: tfang@sjtu.edu.cn; phone:+86-021-34204758

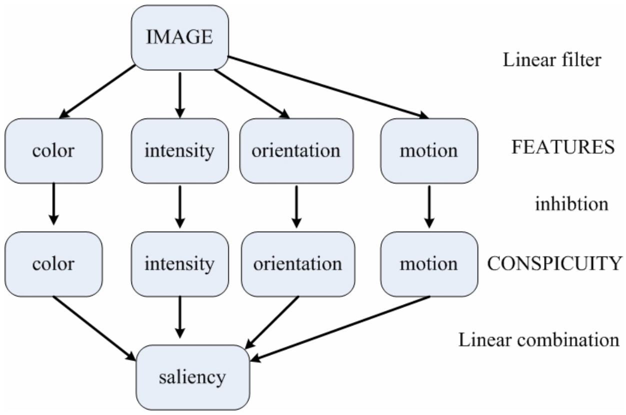


Figure 1. The relation diagram of the extended dynamic saliency map

2.1 The static saliency map

We use a custom-written version of the Itti mode(Itti 1998) to operate over the static saliency map. It can be computed in followed main stages.

Feature maps

In the Itti model, an input static color image is computed by three human different early vision features, such as intensity, color, and orientation. This leads to a multi-feature representation of the scene.

- Intensity feature

$$I = (r+g+b)/3$$

- Four color channels are created, such as red, green, blue, and yellow:

$$R = r-(g+b)/2$$
$$G = g-(r+b)/2$$
$$B = b-(r+g)/2$$
$$Y = (r+g)/2-|r-g|/2-b$$

Then each two chromatic features are computed by the opponent filters.

$RG = |R-G|$, and $BY = |B-Y|$

- Local orientation features are obtained from I using oriented Gabor pyramids $O(\theta)$, where $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$ is the preferred orientation.

Then, seven different features F_i are used in this model.

Conspicuity maps

In the second stage, each feature map F_i is computed to create its conspicuity map C_j, which highlights the important parts in each specific feature by center-surround differences. So, multi-scale *difference-of-Gaussians* (DoG) filters, which can be implemented using Gaussians pyramids, are used for a multi-scale representation in a scene. I is then computed as the Gaussian pyramid $I(\sigma)$, where σ is the pyramid scale, in the following filter-subtract-decimate manner:

$$I^0(\sigma+1) = G * I^0(\sigma) \quad (1)$$

$$I(\sigma+1) = \text{SUBSAMPLE}[I^0(\sigma+1)] \quad (2)$$

with $I^0(0) = I$, $\sigma = [0..8]$, G the Gaussian filter, and SUBSAMPLE a function which subsamples the input image by a factor of 2. Then the center-surround differences $I(c,s)$ in intensity feature yield the feature maps, with $c \in \{2,3,4\}$, and $s = c+\delta$, $\delta \in \{3,4\}$. (More details can be shown in itti model(Itti 1998))

At the same time, two color maps $RG(c,s)$ and $BY(c,s)$ are obtained. Furthermore, local orientations at multi-scale $O(\sigma,\theta)$ are computed by taking the real component of spatial Gabor filtering over levels of the Laplacian pyramid, and orientation feature maps $O(c,s,\theta)$ are gained.

In total, 42 feature maps are computed: six for intensity, 12 for color, and 24 for orientation.

So, for feature intensity, color, and orientation, their conspicuity maps are computed as follows:

$$\overline{I} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(I(c,s)) \quad (3)$$

$$\overline{C} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \left[N(RG(c,s)) + N(BY(c,s))\right] \quad (4)$$

$$\overline{O} = \sum_{\theta \in \{0^\circ,45^\circ,90^\circ,135^\circ\}} N\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} N(O(c,s,\theta))\right) \quad (5)$$

where operator N means the normalization. In these above-mentioned expressions, each item is assumed to have the same contribution to the conspicuity maps.

The static saliency map

Finally, three feature conspicuity maps are combined into the static saliency map S_s . The integration can be computed as follows:

$$S_s = \frac{1}{3}\left(N(\overline{I}) + N(\overline{C}) + N(\overline{O})\right) \quad (6)$$

It is clear that the static saliency map is purely data-driven and does not require any prior knowledge about the scene. Namely that, the competitive method is totally bottom-up, and only three early vision features are used to simulate the human vision system to determine the region of interest in the scene.

2.2 Dynamic saliency map

The dynamic saliency map S_d is another part of the saliency map, which discriminates moving objects in the scene. The original itti model does not involve the motion scene. This is discussed in later work(Yee 2002; Itti 2003; Itti 2005; Shic 2007). However, Yee and itti model seem to be a mismatch between the theory and implementation (especially in itti 2003). Otherwise, itti's surprise model concerns the statistical differences in the video streams, which cannot obtain any explanation in electrophysiology. In contract, the Extended Itti model, presented by shic, simulates the original itti model in static scene, and uses some different formulation in motion saliency, which can capture some of the main "pop-out" phenomena that we would expect.

In the extended itti model, a compromise approach is applied to compute motion saliency, a variation of time-varying edge detection (shown in Jain et al. (1995)). The dynamic “edginess” of a point, $E_{s,t}$, is combined by the product of the spatial and temporal derivatives:

$$E_{s,t}(x,y,t) = D_s I(x,y,t) \cdot D_t I(x,y,t) \tag{7}$$

where I is the intensity of an image at the scene. Following Jain’s idea, the itti model can be extended in time for motion saliency. For example, the intensity modality $I(\sigma)$ becomes $I(t,\sigma)$, $RG(c,s)$ becomes $RG(t,c,s)$, and so on. And, under the assumption that motion is largely color-blind, motion saliency can be computed by the intensity between different frames only. So, for N frames $I(t,\sigma)$, $t \in [1..N]$, motion feature maps can be implemented by the following steps:

1) Compute the N-th order first temporal derivative, $M_t(t,\sigma)$.

2) Compute the spatial derivative, $M_s(t,\sigma,\theta)$:

$$M_s(t,\sigma) = Im\left\{O_c(t,\sigma,\theta)\right\} \tag{8}$$

3) Compute the motion feature map $M(t,\sigma,\theta)$:

$$M(t,\sigma,\theta) = M_s(t,\sigma,\theta) \cdot M_t(t,\sigma,\theta) \tag{9}$$

Then, the motion conspicuity maps can be computed as follows:

1) Compute the direction of motion for each orientation to obtain positive and negative directional features. The positive directional feature $M_+(t,\sigma,\theta)$ is $\sqrt{M(t,\sigma,\theta)}$ at the location where it is positive, and is 0 otherwise: the negative directional feature is computed similarly.

2) Compute the directional contribution to motion conspicuity, $M_d(t,\sigma,\theta)$:

$$M_d(t,\sigma,\theta) = N\left(M_+(t,\sigma,\theta)\right) \oplus N\left(M_-(t,\sigma,\theta)\right) \tag{10}$$

3) Compute across-scale contribution for each orientation:

$$M_o(t,\theta) = \bigoplus_{\sigma=0}^{8} N\left(M_d(t,\sigma,\theta)\right) \tag{11}$$

4) Finally, the motion conspicuity map is:

$$S_d = \bar{M}(t) = \sum_{\theta \in \left\{0^\circ,45^\circ,90^\circ,135^\circ\right\}} N\left(M_o(t,\theta)\right) \tag{12}$$

2.3 The final saliency map

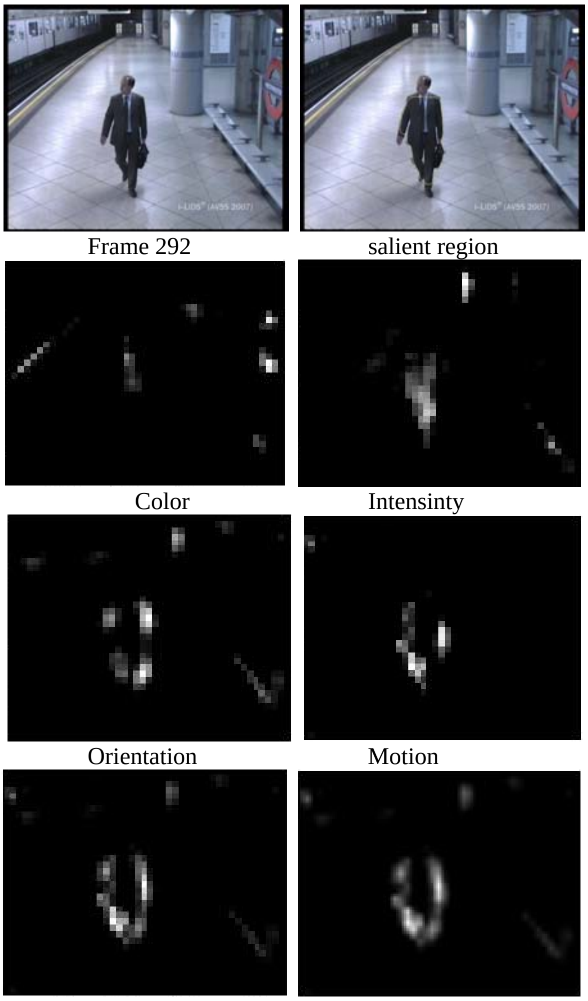
Finally, the saliency map can be integrated by the static saliency map S_s and the time-varying one S_d , according to the following equation:

$$S = w_s S_s + w_d S_d \tag{13}$$

Here, the weight w_s and w_d are set 1.

And then, the most salient locations of the frames are selected by the WTA network. The detected locations in each frame are used for further image sequence processing, such as matching and tracking, etc.

Figure 2 shows an example of the four conspicuity maps and the final motion saliency map. It also pops out the salient region by WTA.



Frame 292 salient region

Color Intensinty

Orientation Motion

Motion Saliency Map WTA

Figure 2. Detecting the saliency map by combining the four conspicuity maps: color, intensity, orientation, and motion. And WTA is then applied to generate the salient region in the frame.

3. OBJECT TRACKING

For tracking purpose, correspondences are required between frames to form object trajectories. Typical tracking methods(Wang 2003; Wu 2004) are divided into four major categories: region-based, active-contour-based, feature-based and model-based tracking. Here, a region-based tracking strategy is adopted in our tracking system. Thus, the basic idea about the visual tracking is to select the salient locations for further tracking tasks. Therefore, our model detects the salient locations and uses them for further recognition and tracking. However, Kadir(Kadir 2001) points out that it is unlikely that features exist entirely within the isotropic local region and hence several neighboring positions are likely to be equally salient. In such a case, the most salient point in the saliency map from frame to frame could be a result of noise rather than underlying image features. So, a segmented salient region is generated for image matching and tracking, rather than the salient point. In this process, we can choose the salient points as the seed point in segmentation. Then, we propose a new technique to match and track the objects from frame to frame.

In our algorithm, object tracking can be computed in two steps: (1) determining the robust features of the tracked objects; (2) matching salient regions in the video streams. First, in the processing of computing the salient regions, multi-feature maps and combined conspicuity maps are computed, and then only one most salient region ("winner") in one of the feature maps will be pop-out. Thus, this feature can be described as the discriminating feature in the scene, which is the feature that can distinguish a salient region most effectively. Here, we assume that the same object in the consecutive frames within the similar background may apply the same discriminating feature to pop out the salient region(Ouerhani 2003). This characterization is then used to locate the salient regions over time.

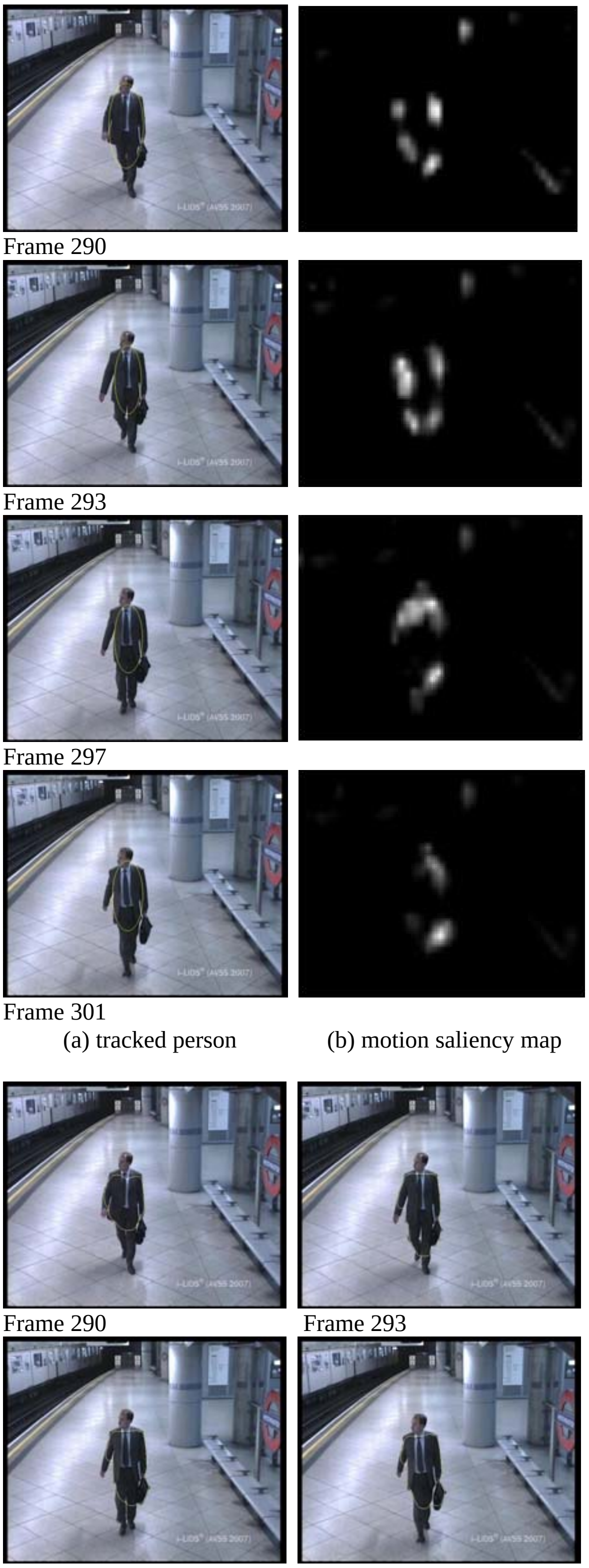
Second, since the salient regions in the discriminating feature map from frame to frame are obtained, the object tracking task can be posed as a regions matching problem. In our tracking algorithm, a new detected salient region after roughly segmentation in the next frame is inserted into the existing object tracking sequence, depending on its similarity with the last inserted region. In our paper, the color histogram is used for the matching problem, since it could be made invariant to illumination changes, as well as geometric transformations. Formally, let x_{t_i} the actual detected most salient point in frame t_i, which is shown in the conspicuity map j_i^*, and the salient region $R\left(x_{t_i}, j_i^*\right)$ can be generated by using the salient point (as seedpoint) to roughly segment the moving object. Then, the color histogram $Ch\left(x_{t_i}, j_i^*\right)$ in the salient region is obtained. Therefore, the similarity can be measured when the following inequality is satisfied:

$$\left|x_{t_{i+1}} - x_{t_i}\right| < \varepsilon \,\&\, \left|Ch\left(x_{t_{i+1}}, j_{i+1}^*\right) - Ch\left(x_{t_i}, j_i^*\right)\right| < \varepsilon \tag{14}$$

The inequality means that only when these two objects have the similar color histogram and their center distance is close, then the salient region in the next frame will be inserted into the tracking sequence. Thus, the salient regions matching between different frames are based on color histogram similarity and the seedpoint spatial proximity. When a detected salient region can not correspond to the existing one, then it means that the tracked object disappears and a new tracking sequence should be initialized.

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Results



Frame 290

Frame 293

Frame 297

Frame 301

(a) tracked person (b) motion saliency map

Frame 290 Frame 293

Frame 297 Frame 301

(c) segmented salient regions

Figure 3. Object tracking based on motion saliency model. There are four sample frames to show the tracking processing. Column (a) shows the tracked moving person; (b) presents the motion saliency map of the current frame; (c) shows the segmented salient regions for further matching and tracking.

To test our proposed tracking method, we select a video recorded in the underground railway, where people go through the channels. This dataset comes from AVSS 2007, which is a sub-set of the i-Lids dataset, and it can be found at http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html. Each video in the dataset consists of 720*576 color images and is recorded at 25 frames per second.

Figure 3 shows the tracking results using the sequence in which a person moves from one side of the underground railway to another. The processing is started from frame 290 shown on the top row, and our model is applied to the example to generate the motion saliency map, segmented salient regions and the final tracked result in Figure 3. Here, the first column represents the tracked object; the second column represents the motion saliency map; and the third column represents the segmented salient regions. It is clear that the motion saliency map is capable to locate the part of moving person, which proves that the motion saliency model works well. Next, a roughly segmentation is used to determine the matching regions in Figure 3(c). And then, the final result is shown in Figure 3(a), in which person is tracked in the consecutive frames. The results show that it is easy to validate our tracking algorithm applicable and effective.

4.2 Discussion

Our tracking model addresses the problem of how tracking processing works based on the computational motion saliency model. Key properties of this algorithm are that its usage of the motion saliency model, and color histogram matching based on the roughly segmentation. The biologically-inspired motion saliency model is presented here in order to simulate vision attention, which is affected by the stimulus colors, intensity, orientation and motion. This process is easily implemented on a computer, and its behaviour on most cases is in good quantitative agreement with the human psychophysical literature. Compared to other image pre-processing methods, motion saliency model has a strong performance in complex natural scenes(Itti 1998). For example, it can rapidly detect the salient region, and effectively guide bottom-up attention. From a computational viewpoint, the major advantage of this approach lies in the massively parallel implementation. Especially, the attended salient regions pop out the region of interest in the scene, and then further image processing is limited in these regions. So, the technique vastly reduces the amount of information that should be used at subsequent stages of processing.

Then, level set method(Chan 2000; Chan 2001) is applied to roughly segment the salient regions. Pre-segmentation is capable to detect the tracked object, and remove inhomogeneous regions. This process can eliminate the irrelated information around the tracked object. But, it is noted that exact segmentation is non-trivial and time-consuming, especially for the level set model. Thus, the trade-off between some inaccurate information and the time-consuming cost is an issue that we hope to address in future work.

The use of the normalised color histogram for sub-image matching is a single way to describe the segmented region and track moving object. It is clear that this method is invariant to rotation, scaling, and other geometric transformations, as well as the intensity changes. However, it is noted that the algorithm performs poor, when compared to other matching method. But, in our case, the algorithm is usable for the reason that the method applies the salient region segmentation prior to matching, which is based on the assumption that motion saliency model is capable to pop out the similar regions in most cases. There are many different matching methods, the effect and efficiency of which we hope to investigate in the future.

5. CONCLUSION

In this paper, we have presented a novel technique for object tracking. Our method is based on the extended saliency model for computing the time-varying saliency map, which can locate some salient regions for distinguishing and tracking the objects. This dynamic saliency model is useful to catch the moving objects. Especially, the salient regions can guide an attentive objects matching processing, which is the base in object tracking over time. Moreover, the saliency model is able to avoid time-consuming computing, since only a few regions of interest should be computed for further object matching and tracking. And also, the paper applies the color histogram for image matching to track these objects in the video sequence, by posing the tracking problem as a sub-image matching problem across pairs of video frames. It is clear that the tracking technique is invariant to illumination changes and geometric transformation. The example, presented to illustrate the process of object tracking, shows the usefulness of dynamic attention-based object tracking. In the future work, more effort will be devoted to the problem on object tracking without pre-segmentation.

ACKNOWLEDGEMENTS

The authors would like to thank Frederick Shic for his ardent help on implementation of the extended itti model.

REFERENCES

Chan, T., Sanderberg, B., Vese, L., 2000. Active contours without edges for vector-valued images. *Journal of Visual Communication and Image Representation*, 11(2), pp. 130-141.

Chan, T., Vese, L., 2001. Active contours without edges. *IEEE Transactions on Image Processing*, 10(2), pp. 266-277.

Gilles, S., 1998. *Robust description and matching of images*, Phd thesis, University of Oxford.

Itti, L., C. Koch, et al, 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11), pp. 1254-1259.

Itti, L., Dhavale, N., Pighin, F., 2003. Realistic Avatar Eye and Head Animation Using a Neurobiological Model of Visual Attention. *In: Proc. SPIE 48th Annual International Symposium on Optical Science and Technology, Bellingham, WA:SPIE Press*.

Itti, L., Baldi, P., 2005. A Principled Approach to Detecting Surprising Events in Video. *In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR).*

Jain, R., Kasturi, R., and Schunck, B.G., 1995. *Machine Vision*. McGraw-Hill Science/Engineering/Math.

Kadir, T., Brady M., 2001. Saliency, scale and image description. *International Journal of Computer Vision*, 45(2), pp. 83-105.

Koch, C., Ullman, S., 1985. Shifts in selective visual attention: towards the underlying neural circultry. *Human Neurobiology*, 4, pp. 219-227.

Ouerhani, N., Hugli H., 2003. A model of dynamic visual attention for object tracking in natural image sequences. *Computational Methods in Neural Modeling, Pt 1*, 2686, pp. 702-709.

Shic, F., Scassellati, B., 2007. A behavioral analysis of computational models of visual attention. *International Journal of Computer Vision*, 73(2), pp. 159-177.

Stentiford, F., 2007. Attention-based similarity. *Pattern Recognition*, 40(3), pp. 771-783.

Treisman, A., Gelade, G., 1980. A feature-integration theory of attention. *Cognitive Psychology*, 12(1), pp. 97-136.

Wang, L. A., Hu, W. M., Tan, T. N., 2003. Recent developments in human motion analysis. *Pattern Recognition,* 36(3), pp. 585-601.

Wu, W. M., Tan, T.N., Wang, L., Maybank, S., 2004. A survey on visual surveillance of object motion and behaviors. *IEEE Trans. SMC Part C*, 34(3), pp. 334-352.

Yee, C., Walther, D., 2002. *Motion detection for bottom-up visual attention*. tech. rep., SURF/CNS, California Institute of Technology.