Confidence Based updation of Motion Conspicuity in Dynamic Scenes

Vivek Kumar Singh, Subhransu Maji and Amitabha Mukerjee,
Dept of Computer Science and Engineering,
Indian Institute of Technology Kanpur, India.
viveksin@usc.edu, {maji, amit}@cse.iitk.ac.in

Abstract

Computational models of visual attention result in considerable data compression by eliminating processing on regions likely to be devoid of meaningful content. While saliency maps in static images is indexed on image region (pixels), psychovisual data indicates that in dynamic scenes human attention is object driven and localized motion is a significant determiner of object conspicuity. We have introduced a confidence map, which indicates the uncertainty in the position of the moving objects incorporating the exponential loss of information as we move away from the fovea. We improve the model further using a computational model of visual attention based on perceptual grouping of objects with motion and computation of a motion saliency map based on localized motion conspicuity of the objects. Behaviors exhibited in the system include attentive focus on moving wholes, shifting focus in multiple object motion, focus on objects moving contrary to the majority motion. We also present experimental data contrasting the model with human gaze tracking in a simple visual task.

1. Introduction

Visual attention refers to the ability of vision systems to rapidly select the most salient data in a scene, so as to drastically reduce the amount of visual information required in high level tasks such as motion tracking, object recognition, etc. Computationally, visual attention models are becoming widespread for static scenes like image compression [11] and segmentation [12] and are also beginning to be used for dynamic scenes [11].

The Itti-Koch model[8], [7] provides a computational model for bottom-up attention in static scenes. Their attention model for static scenes combine intensity, colour and orientation features to produce a saliency map, which which simulates a computational function resident either in LGN or in the V1 in mammalian brains. Finally, a Winner-Take-All (WTA) network (located near the thalamic recticular nucleus) identifies the actual scene location for fixation. A comparison of the saccades from an implementation of Itti's model and psychological data is shown in Figure 1.

2. Dynamic Vision

Visual attention in dynamic scenes differs from static models of gaze fixation in several important ways:

- Object fixation. While saliency maps in static images are indexed on image region (pixels), psychovisual data indicates that in dynamic scenes human attention is object driven, and localized motion is a significant determiner of object conspicuity.
- Motion-based Perceptual Grouping. Objects with similar motion are grouped together in a motion gestalt.
- Confidence Decay. Parts of the scene gain saliency based on input features such as motion, but other parts that have not been visited for some time also face a decay in the confidence of their confidence estimate.
- Relative Motion. Motion salience is predicated on object motion, but even among moving objects, objects moving contrary to the flow gain salience.
- Winner Take All and Inhibition are present as in static attentive mechanisms.

Traditional methods to incorporate motion information have been to include another saliency map for motion called the motion saliency map using the optical flow information, or similar ideas. In [11] optical flow

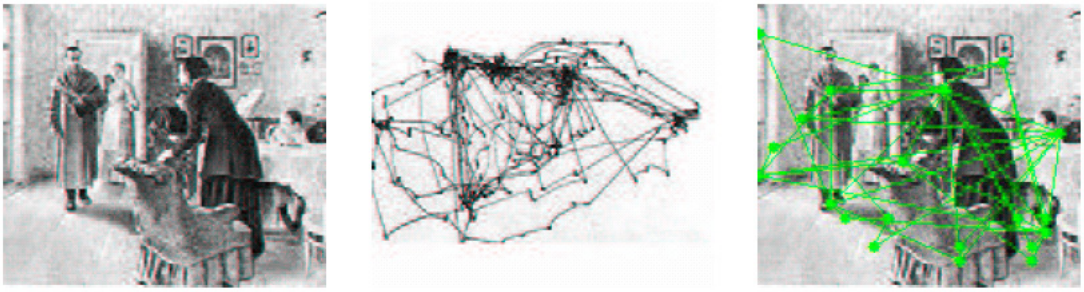

Figure 1. The painting "The Unexpected Visitor", by I.E. Repin (left) the gaze switches by a subject as observed in the pioneering work of Yarbus [14], and gaze switches computational model(right).

is used to locate portions in the images where there is motion and higher saliency is attached to those locations. Though simple to understand these methods are not suited to handle scenes where there is a lot of motion and so a mechanism is needed to prioritize the moving objects.

This work proposes a computational model of visual attention that uses flow segmentation to capture the perceptual grouping of objects. A motion saliency features is computed based on the localized motion conspicuity of the objects (groups). This motion saliency model coupled with a static saliency model constitutes the basic saliency map for simulating the visual attention of humans in dynamic scenes. Finally, a Confidence map is used to update the confidence for the current foveated region. We validate the behavior in the system with experimental data contrasting the model with human gaze tracking in a simple visual task.

Visual attention is mediated by the current task (top-down) as well by features of the scene itself (bottom-up). While the former aspects vary widely, the latter results in relatively stable computational process. While these numbers can vary widely, some experiments [2]suggest that when viewing natural images, task-dependent factors accounted for 39% of the eye movement related information, whereas task-independent factors accounted for 61%.

In our bottom-up attention architecture for dynamic scenes, we build on the existing static attention models by incorporating features for visual motion. Four principles guide the Itti-Koch model: Visual attention is based on multi-featured inputs; saliency of a region is affected by the surrounding context; the saliency of locations is represented by a saliency map, and the Winner Take All and Inhibition of return are suitable mechanisms to allow attention shifts[Figure 3].

2.1. Computational Model

To model the foveal image processing (exponential fall of information from the fovea in human eye) in human eye, we apply multiscale gaussian filtering to the image. The foveal map is modeled as a a linear combination of radially weighted smoothed (gaussian convolved) images. Mathematically,

$$I_f(x,y) = \sum_i f_i(r) \times I_i(x,y) \tag{1}$$

where, $I_f(.,.)$ is the foveated image, f_i is an exponential function of $r(=\sqrt{x^2+y^2})$ and I_i is image obtained by convolving with the ith gaussian. Note that here the gaussian obtained are not layers of gaussian pyramid but we basically generate a pyramid of gaussian filters of different sizes and ith smoothed image is obtained by convolving the filter with $(i-1)$th smoothed image.

For foveating a color image, we first transform the input image in RGB format to a suitable format YCrCb which separates out the color component from intensity Y. The color foveated image is obtained by foveating the intensity image Y and superimposing the original color information (CrCb).

Figure 2 pictorially describes the algorithm for foveating an image.

2.2. Feature Maps for Static Images

First, a number of features $(1 \dots j \dots n)$ are extracted from the scene by computing the so called feature maps F_j . Such a map represents the image of the scene, based on a well-defined feature, which leads to a multi-featured representation of the scene. In his implementation, Itti considered seven different features which are computed from an RGB color image and which belong to three main cues, namely intensity, color, and orientation.

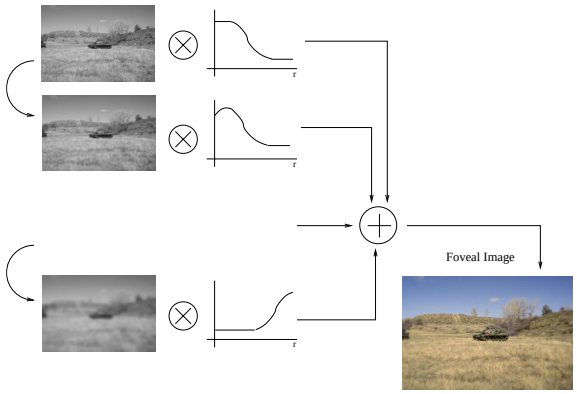

Figure 2. Overview of Foveation Algorithm

- Intensity Feature: $F_1 = I = 0.3 * R + 0.59 * G + 0.11 * B$
- Two chromatic features based on the two color opponency filtersR^+G^- and B^+Y^- where the yellow signal is $Y = \frac{R+G}{2}$. Such chromatic opponency exists in human visual cortex [6]. $F_2 = (R - G)/I$, $F_3 = (B-Y)/I$. The normalization of the features with I decouples hue from intensity.
- Four local orientation features (gabor filters) $F_{4...7}$ according to the angles $\theta \in 0, 45, 90, 135$.

2.3. Conspicuity Maps

In second step, each feature map is transformed in its conspicuity map which highlights the parts of the scene that strongly differ, according to a specific feature, from their surroundings. In biologically plausible models [8], this is usually achieved by using a center-surround-mechanism. Practically, this mechanism can be implemented with a difference-of-Gaussians-filter, DoG, which can be applied on feature maps to extract local activities for each feature type. This method is based on a multi-resolution representation of images. For a feature j, a gaussian pyramid P_j is created by progressively lowpass filtering and subsampling by factor 2 the feature map F_j , using a gaussian filter G:

$$P_j(0) = F_j \tag{2}$$

$$P_j(i) =\downarrow (P_j(i-1) * G) \tag{3}$$

where (*) refers to the spatial convolution operator and $\downarrow$ refers to the downsampling operation. Center-Surround is then implemented as the difference between fine (c for center) and coarse scales (s for surround). Indeed, for a feature j $(1 \dots j \dots n)$,

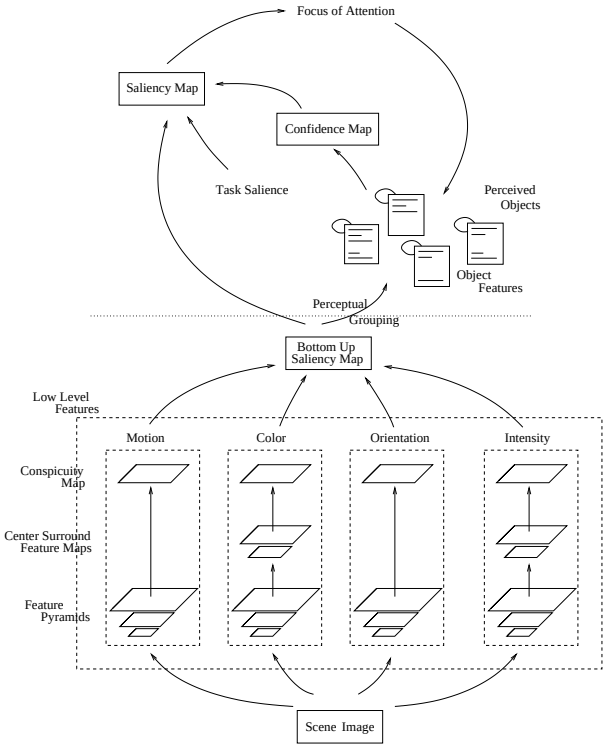

Figure 3. Computational Model of Visual Attention

a set of intermediate multiscale conspicuity maps $M_{j,k}(1 \dots k \dots K)$ are computed according to the equation below, giving rise to (n*K) maps for n considered features.

$$M_{j,k} = |P_j(c_k) - P_j(s_k)| \tag{4}$$

where - is a cross-scale difference operator that first interpolates the coarser scale to the finer one and then carries out a pixel-by-pixel subtraction.

2.4. Motion

To model the effect of neural responses of the motion selective neurons based on the local motion, we have used the optical flow models. Due to the aperture problems and lack of enough texture, flow vectors obtained using the Lucas Kanade approach [9] were very poor.

One approach to obtain good flow vectors is to use a set of image features that would enable object tracking. Since the stimulus (video) that we have used in our experiments mainly contains triangles and quadrilaterals hence the corners of these objects can be used as a feature for tracking. This is also coherent with the belief that to update our mental world model we also rely on the boundary responses.

To extract the corners in the scene and for tracking the selected set of corners in the scene, we have used the pyramidal Lucas Kanade sparse feature tracking algorithm [3]. The implementation of these routines are included in OpenCV [5]. Flow vectors for non-boundary points of the object are obtained by interpolating the flow vectors at the corners.

Figure 4 shows an image (obtained by overlapping 2 consecutive frames) and the corresponding motion feature map between these two frames.

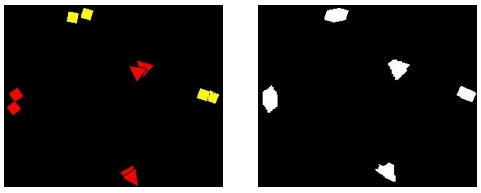

Figure 4. (a) 2 consecutive frames overlapped (b) motion saliency map

3. Perceptual Grouping

Perceptual grouping refers to the grouping of spatial sensory information on the basis of features such as form, shape etc. The concept of perceptual grouping is inspired by psychophysical experiments. A neurological basis is still non-existent. However, long range excitatory connections in V1 and neural interactions are believed to be involved in perceptual grouping. This stage allows to integrate the space-based and object-based theories of attention and hence can be used to explain the observations supporting them. We further believe that this process of grouping is strongly affected by cognitive factors and hence forms a very crucial step in scene understanding.

An obvious question that perceptual grouping raises is what are the features that are involved in grouping. Gestalt laws of perceptual organization tries to answer this question, by a clearly enumerated set of principles (see [13]). Since grouping of scene features into objects is itself a very complicated task, hence in this work we have focused on the ideas which we believe are relevant to our video (see Experiment section for video details). Since our video contains objects with different colors, so clearly the principle of similarity (regions with similar features are grouped) and that of proximity (regions that are close are grouped) based on features plays a significant role. Another important gestalt principle would be grouping in temporal domain based on motion similarity. But the randomness in object motion allows us to safely assume that such grouping may not have a significant effect.

We implement grouping of image parts by their similarity of color features using and spatial proximity using a pyramidal segmentation algorithm, which looks at neighbors of each pixel and group them based on their similarity at each level of pyramid. Segments with size below a certain threshold are dropped.

4. Bottom up saliency map

The low level feature processing decomposes the image into feature maps. Now the question that rises is how do we control a single attentional focus using multiple feature maps. To solve this problem, bottom-up approaches of attention simply combine these feature maps into a single map, which they call *saliency map*. A saliency map is a scalar 2D map which indicates the visual salience topographically. In this work, we have simply taken *saliency map* to be the average of the color, intensity, orientation and motion feature maps. Figure 5 below shows some sample images and their corresponding saliency maps.

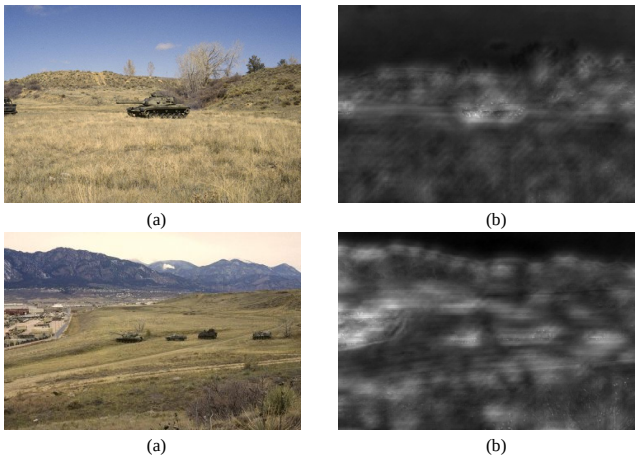

Figure 5. (a)Images, (b)Saliency maps

5. Cognitive factors

Cognitive factors play a very important role in visual attention by constraining attention to an "interesting" subset of image locations. These factors primarily include object recognition, scene understanding and task salience. Each of these is a very difficult problem in itself. Scene understanding requires modeling the long

term memory and psychological factors while for computing task salience, object relationships in context to the task needs to be inferred.

In this work however, our choice of stimulus and the task at hand makes it easier for us model these effects. In particular, long term memory no longer needs to be modeled and the search space for recognizing objects in each frame is much smaller.

Another key task at this stage is object tracking which corresponds to the pursuit human eye movements.

5.1. Object tracking & Information map

Object tracking is achieved using the motion feature maps obtained in low level feature processing stage. However rapidly changing speed and appearance/disappearance of objects in the video have made object tracking trickier than simply updating the object positions using the low level motion features maps. Between successive frames, we need to perform a search in the neighborhood of the object to associate the flow vector with a moving object. Search window size gives a trade off between accuracy of event detections and processing speed. Based on our observations on the object size (10-12 pixels) and speed extremes (maximum of 4-5 pixels per frame), we choose a search window of 20×20 size. To simulate the effect of foveal processing of motion features, the track features of only objects within a certain threshold of foveal radius were updated.

In our experiments, we observed that that when subjects switch attention from one object to another, they fixate on it for a few frames. We believe that during this time they update their internal representation of the object and its features. Fixating on moving objects essentially correspond to pursuit of eye movements for tracking. In this work, we modeled this behavior by encoding the tracking error obtained from foveal feature update and finite optical flow window into an motion error salience map. This *confidence map* highlights the high tracking error regions of objects, which in turn indicates how confident the system is about the object's position. (a low confidence value means a high salience)

The tracking error (in object's position) between successive frames is obtained by calculating the portion of overlap between the predicted position of the object obtained using the motion feature map and the current position obtained by perceptual grouping. We associate a confidence term (or information term) with each object that stores the overlap ratio. So a high overlap ratio means that we have a higher confidence on the position of the object and its features. Hence lower confidence value implies higher salience for that object.

For matching objects in successive frames, we say that if overlap ratio is above a certain threshold then they are instances of the same object and update the track features. This allows the system to continuously track all the objects in the scene. If however the object is moving too fast or its features have not been updated since last few frames then this would result in noisy flow and reduction in the overlap region. This would in turn result in a lower confidence value and hence higher salience. Note that since the flow vectors are not very accurate, the overlap ratio over time may have some error peaks/oscillations. So we enforce that changes in overlap ratio result only in a decrease in confidence value. In general, the confidence value of the object increase only when it falls within the foveal processing region.

5.2. Task driven salience

As discussed before the task at hand plays a significant role in driving our attention. One way to model the effect of task salience is to use a task relevance map as suggested by Navalpakkam et al in [10].

In this work, the task that we have chosen does not involve any semantic concepts/relations between objects and hence task relevance map is rather simplistic and need not be explicitly modeled. Since the task involved keeping a count of red triangles and squares, dominant features relating to task were shape and color, in particular red objects were more task relevant. Thus for this task we we used a weighted R component of image as the task salience map.

5.3. Selection of the point of Attention

Once the saliency Map has been computed the Winner Take All (WTA) and Inhibition of Return are suitable mechanisms to imitate the eye movements and the focus of attention. The WTA network implements a parallel computation and iteratively converges to the point of maximum saliency. Stochasticity can be added so that quicker convergence can be achieved.

However, WTA always selects the point with maximum salience and hence will end up fixating the point of attention at the most salient point forever. The movement of the attention point (to capture other areas of interest i.e comparitive salience) is achieved by inhibiting the salience of the object currently in focus, as done in the work of Backer et al [1]. At each iteration the saliency of the object being attended to is decayed,

thus eventually the objects not being attended to will increase in saliency and take the focus of attention. To make sure that saliency decay rate does not overshadow the effect of tracking error salience, decay rate was set a very small value.

6. Comparison with human subjects

To validate our model, we present here a comparison of gaze position predicted by our system on a synthetic video against human gaze position recorded for the same video. We generated a video containing simple geometric objects undering dynamic events (varying object speeds, varying shapes/colors) and recorded the reaction of human gaze on these events. We then compared the gaze positions predicted by our system with the real gaze data obtained from the experiment. Gaze data was collected at University of Rochester, New York using eyelink2 gaze tracker. The visual scanpaths over dynamic scenes in not included here since scanpaths in general were rings (pursuit movements) with a few diagonals (saccadic jumps) that did not reveal much information about the eye movements.

6.1. Experiment Details

The input video (stimulus for the experiment) contains objects with simple geometric shapes (triangle and squares) circling around in the screen with changing frequencies. Objects keep appearing and disappearing throughout the video and change their colors and their shapes (triangle or square). Figure 4-(a) shows an overlapping sample snapshot of the video. The first column in Table 1 shows the occurence of various events in the video.

In the experiment, subjects were asked to concentrate specifically on the object shapes and colors as they may change over time and that they would have to keep a count of the objects with different shapes and colors. Furthermore they were told that their performance will be evaluated at the end of the video, by asking questions, a majority of which are on the red objects. The prior knowledge of the domain of questions were aimed to analyze the effect of task on feature selection.

Table 1 shows the ground truth and comparison between response time (in frames) to different video events recorded on the humans and those predicted by our system. First column lists the events while other columns indicate the frames at which our model/subject responded to the corresponding event (Video frame rate = 15 fps). NIL in table 1 means these events had no effect on the gaze. * indicates that the gaze was already close to that location and no saccadic eye movement was observed.

Events	GT	S1	S2	S3	Model
Ro appears	1	1	1	1	1
Ro appears	28	30	43	46	28
Ro appears	81	82	87	86	82
Yo appears	117	124	124	NIL	118
Yo appears	178	190	190	182	179
Color change - Yellow to Red	240	241	242	247	246
Ro appears	316	317	317	320	317
Ro appears	352	354	354	354	353
Ro appears	485	487	487	490	490
Bo appears	574	580	580	576	575
Bo appears	612	NIL	NIL	NIL	613
Color change - Blue to Red	871	872*	872*	878	871
Ro appears	908	913	913	911	909
Yo appears	981	989	989	985	982
Color change - Yellow to Red	1134	NIL	1143	1137	1138

Table 1. Table showing events in the video and the frame at which objects involved in the event were attended, Ro, Yo and Bo are Red, Yellow and Blue Object respectively, GT is the ground Truth, S1, S2 and S3 are various subjects.

7. Modifications and Extensions

In this section we show that including local motion conspicuity cue leads to better gaze prediction results. In Figure 7, the ball at the center moving in the opposite direction of the others, is the most salient object in the image. The motion of objects in the local neighborhood clearly influence how conspicuous an object is.

We have implemented a computational model of this based on clustering of the image regions into regions of similar motion. This can be done by clustering flow vectors. However in many synthetic scenes flow vectors are sparse because of the lack of sufficient texture so alternatively we have used the Motion Segmentation Algorithm [4] to get the connected components(motion segments) as seen in Figure 6- (row 2). The saliency of each object is determined by the local motion conspicuity of the object. A gaussian mixture model is fitted into the data of the motion directions of the objects within a certain radius and the conspicuity of the

object is inversely proportional to the prior of the gaussian to which it belongs, quantifying how rare the motion of the object is within the local group. One of the possible measures satisfying the above properties is the following:

$$C(O) = (1 - P(g))^n; g = argmax_a \frac{d(O, \mu_a)}{\sigma_a} \qquad (5)$$

Here n is a constant and d(**a**,**p**) gives the distance between the **a** and **p**. Higher values of n biases the saliency towards the less probable directions. In our experiments [Figure 6], we have used a n value of 3. The model has been tested on these examples and the results are compared with the previous model[Figure 7]. Finally the model is tested on some realistic scenes [Figure 6]- (row 1). Though it is known that human gaze is highly task driven, the gaze predicted by our model is visually very acceptable demonstrating the efficacy of our model.

8. Future Work

In this work, we extended the conventional Itti-Koch model to handle dynamic scenes. We have successfully implemented our model and showed its efficiency and accuracy by comparing it with the gaze data obtained from the experiments on human subjects. We then take the model a step further by providing a notion of comparing the saliency of the moving regions and perceptual grouping. As shown in the example videos this becomes important when there are many moving objects in the scene.

One very important tool that we felt is needed while working on the project was a a performance measure to compare the performance of any model as against humans as well as other computational models. We believe that such a measure can possibly be formulated using the Levenstein distance between the gaze vectors.

Using a model of visual attention we can impose an ordering on the objects in the scene (according to their saliency), thus we can use this approach for ranking these objects in order to speed up the content based image retrieval systems. Another possible extension would be to integrate a perceptual database with this model and use the model for object recognition.

9. Acknowledgments

We thank Dr Dana Ballard, Professor at University of Rochester, New York for his invaluable comments on synthetic videos and help in conducting the gaze experiments.

References

[1] G. Backer and B. Mertsching. Two selection stages provide efficient object-based attentional controlfor dynamic vision. In *International Workshop on Attention and Performance in Computer Vision*, April 2004.

[2] R. J. Baddeley and B. W. Tatler. An information theoretic analysis of eye movements to natural scenes. In *ECVP*, 2005.

[3] J.-Y. Bouguet. Pyramidal implementation of the lucas kanade feature tracker, 2000.

[4] G. R. Bradski and J. W. Davis. Motion segmentation and pose recognition with motion history gradients. *Mach. Vision Appl.*, 13(3):174–184, 2002.

[5] I. Corporation. Open source computer vision library 0.9.6.

[6] S. Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637):68–71, 1997.

[7] L. Itti. *Models of Bottom-Up and Top-Down Visual Attention*. PhD thesis, California Institute of Technology, Pasadena, California, Jan 2000.

[8] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, Mar 2001.

[9] B. Lucas and kanade T. An iterative image registration technique with an application to stereo vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.

[10] V. Navalpakkam and L. Itti. A goal oriented attention guidance model. In *Lecture Notes in Computer Science*, volume 2525, pages 453–461, Nov 2002.

[11] N. Ouerhani. *Visual Attention: Form Bio-Inspired Modelling to Real-Time Implementation*. PhD thesis, University of Neuchatel, 2004.

[12] N. Ouerhani and H. H. *MAPS: Multiscale Attention-Based PreSegmentation of Color Images*, volume 2695. Springer Berlin / Heidelberg, 2003.

[13] E. C. Schwab and H. C. Nusbaum. *Pattern Recognition by Humans and Machines: Visual Perception*, volume 2. Harcourt Brace Jovanovich, 1986.

[14] A. L. Yarbus. Eye movements and vision. *Neuropsychologia*, 6:389–390, December 1968.

Figure 6. The results of running the motion saliency part on the image sequence of the walking person is shown(Note: the attended point is computed taking into account both motion and static saliency maps). The first row is the input and the red dot is the final attended location computed by the algorithm. The second rows shows the regions of motion obtained by motion segmentation. The third row shows the weights of the regions with higher weight representing higher saliency. Notice the system is able to locate the region of motion accurately and assign higher salience to those regions.

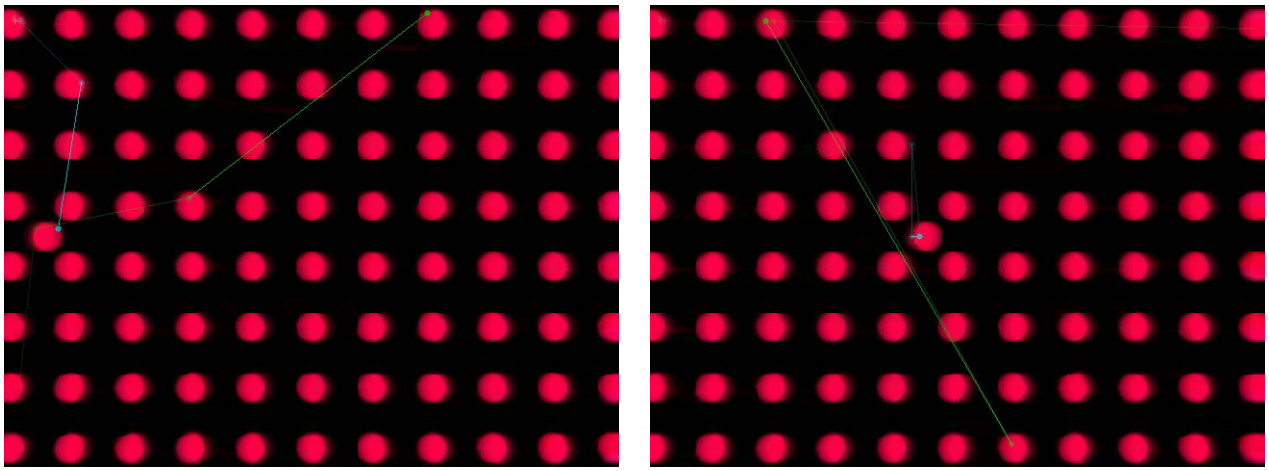

Figure 7. Left image represents a first three and the right one three intermediate frames of the video overlayed on a single image. Here the center ball(initially at the left) at rest initially and starts to move in a direction opposite to the other balls.The gaze point predicted by the new motion saliency model and the previous model is shown as the cyan and green dot respectively (The line indicates the saccadic eye movements). The new model's gaze point fixate at the center ball most of the time, while the previous model's gaze jumps from one ball to another.