Motion Saliency Outweighs Other Low-level Features While Watching Videos

Dwarikanath Mahapatra, Stefan Winkler and Shih-Cheng Yen

Department of Electrical and Computer Engineering
National University of Singapore
4 Engineering Drive 3, Singapore 117576

ABSTRACT

The importance of motion in attracting attention is well known. While watching videos, where motion is prevalent, how do we quantify the regions that are motion salient? In this paper, we investigate the role of motion in attention and compare it with the influence of other low-level features like image orientation and intensity. We propose a framework for motion saliency. In particular, we integrate motion vector information with spatial and temporal coherency to generate a motion attention map. The results show that our model achieves good performance in identifying regions that are moving and salient. We also find motion to have greater influence on saliency than other low-level features when watching videos.

Keywords: Visual attention, region of interest, salient features

1. INTRODUCTION

Understanding why people look at certain areas in an image has been an intriguing problem in vision research for a long time. To answer this question, there have been numerous psychophysical experiments where observers were shown pictures, and the pattern of their eye movements was recorded using eye-trackers. These experiments have given rise to various theories regarding the pattern of eye movements. A widely accepted theory states that the attention mechanism selects the most salient region in a scene, and the eyes are directed to that location. Essentially, the brain computes a saliency map which highlights the salient locations of the image based on local image characteristics such as color, intensity, orientation, etc. Itti and Koch[1] justify such a method to determine saliency maps. However, there are others which discard the concept of saliency maps and state that semantic factors influence the scan path of the eye.[2]

The path followed by the eye while attending to a scene is the result of different underlying factors. Visual saliency may be a dominating factor and so can be semantic content. The visual world is highly structured, and in this environment the human visual system (HVS) has developed a strategy to attend to the various stimuli present. We often move our eyes purposefully to actively gather information from our surroundings. Generally our tasks dictate the pattern of our eye movements. The attended scenes may be as simple as the horizon or may be more complex like a crowded street. One thing is certain: we do not attend to each and every location of the scene. One may spend hours without looking at the bush in the corner, because it did not "stand out". How then do we "decide" where to look?

The properties of the stimulus play a strong role in observed fixation locations. Yarbus[3] demonstrated the active task-dependent nature of fixation locations. Saccades to a simple shape or an object often land near its centroid.[4,5] When viewing natural images, observers tend to fixate at regions with higher local contrast such as borders or edges.[6] Salient points attracting attention is a popular hypothesis,[1,7] which is appealing in passive viewing conditions such as when we are watching television. But it has limited applicability in real life because it is purely stimulus driven. Henderson *et al.*[2] have come up with the hypothesis that semantic content rather than visual saliency influences eye movements during visual search in real-world scenes. There are even information-theoretic models that attempt to explain the pattern of eye movements.[8]

Corresponding author: S.-C. Yen (eleys@nus.edu.sg).
S. Winkler is now with Symmetricom, San Jose, CA 95131.

Human Vision and Electronic Imaging XIII, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6806, 68060P, © 2008 SPIE-IS&T · 0277-786X/08/$18

Considerable effort has gone into the analysis of eye data where subjects look at static images. In our work we use video clips instead of static images. The aim is to determine the role of motion vis-a-vis other low level features in attracting attention. The importance of low-level image characteristics like local contrast is well known,[6] but the role of motion has not been investigated nearly as well. Abrams *et al.*[9] establish the fact that motion is a strong cue in attracting attention. The role of motion raises some pertinent questions: How can we define a framework to include moving objects? Which is more important for attention – low-level image features or motion? How do these two compare with each other and possibly interact?

In prior works related to eye-gaze tracking the subject was engaged in free viewing[10,11] or involved in active visual search.[12] Video clips are used less often as stimuli. Carmi *et al.*[13] have used their model to compute saliency maps for frames in a video and compared the salient regions with fixated regions. The same authors have also explored the role played by memory during prolonged viewing of video clips.[14]

A simple way of including moving objects is the use of motion vectors. As it is done in video compression, the motion vector for a block of fixed size is often a good representation of where that particular block moved in the subsequent image frame. Then arises the question of keeping blocks of pixels belonging to the same object close together over time. This is where the concept of temporal and spatial coherency[15] comes in.

In this paper we investigate the role of motion vis-a-vis other low-level features in attracting attention. For this purpose we have used video clips of Hollywood movies and natural scenes. The saliency map model[1] is used to generate a saliency map for each frame of the clips and to determine the saliency values at fixated locations. We also calculate a "motion saliency" map based on motion vectors and coherency. The corresponding motion saliency values at fixated locations are determined and compared with the previous saliency values. The aim of our work is not to propose a new model for saliency calculation, but to conduct a comparative study of the separate roles of motion and other low-level image features in attracting attention. For that we generate saliency maps for low-level features as well as for motion.

The paper is organized as follows: Section 2 gives details about the type of data used for our experiments. In Secction 3 our approach in distinguishing between fixations and saccades from the eye data is detailed. Section 4 gives a brief overview of how a static saliency map is obtained and our approach to generate a motion saliency map. In Section 5 we present our results and analysis. Section 6 concludes the paper.

2. VISUAL STIMULI

Monochromatic movie sequences were presented monocularly to a monkey at a distance of 57 cm, using a 20" color monitor (Viewsonic P810) at a resolution of 800×600 pixels. The luminance distribution of the monitor (0-80 cd/m^2) was gamma-corrected prior to the experiments. During the course of the experiments, three different graphics cards were used, with refresh rates of 85, 120 and 150 Hz. Movie frames were digitized from DVD copies of Hollywood movies like "The Big Lebowski" and "Everest". The movie files were de-interlaced, and cropped to remove any visible borders. There were no attempts to remove the compression artifacts from the video sequences as they were fairly minor.

The individual frames were then converted to grayscale and presented at rates of 3, 4 or 6 refreshes per frame (corresponding to the 85, 120 and 150 Hz graphics cards), making the frame rate of the movies 28.3 Hz, 30 Hz and 25 Hz, respectively. No temporal up-sampling was attempted, i.e. the identical movie frame was presented for 3, 4 or 6 refreshes. Individual movie frames had dimensions of 640×480 pixels and were positioned at the center of the screen, occupying $32^\circ \times 24^\circ$ of visual angle (0.05°/pixel). For each recording site, a short movie sequence ranging in duration from 15-30 sec was selected, and the clip was presented 60-200 times to the dominant eye.

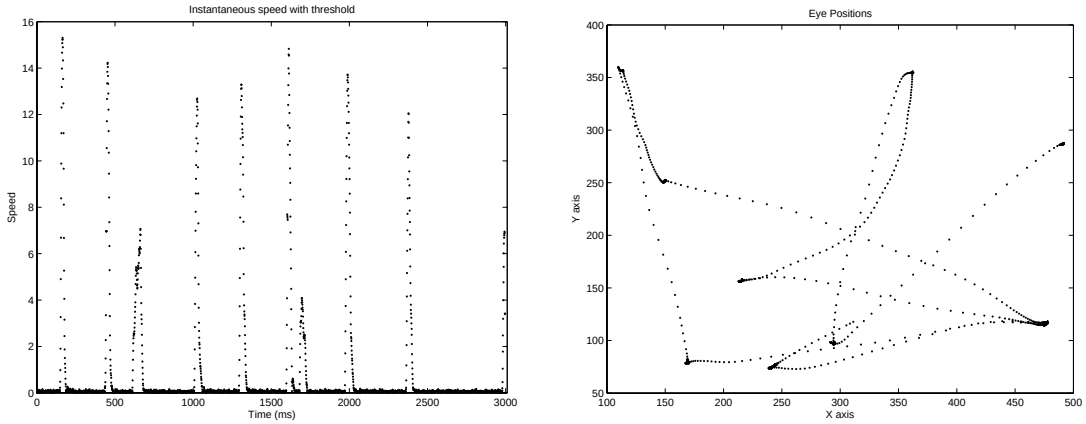
The eye movement data was collected from a macaque monkey while it freely viewed the movies. With the animal's head restrained, eye movements were tracked using surgically implanted scleral eye coils and recorded at a sampling rate of 1 kHz. The animal was trained to keep its gaze within the bounds of the movie during a trial, after which, the animal would be given a juice reward. If the gaze of the animal fell outside of the movie boundaries, the trial was aborted and the data discarded. Each trial lasted approximately 3 seconds.

3. SEPARATING FIXATIONS FROM SACCADES

Eye movements are typically categorized as fixations, when the eye pauses in a certain position, and saccades, when it moves to another position. The resulting series of fixations and saccades is called a scan path. Most information from the eye is made available during a fixation. Hence, the regions of fixations along a scan path show what information on the stimulus were processed during an eye tracking session.

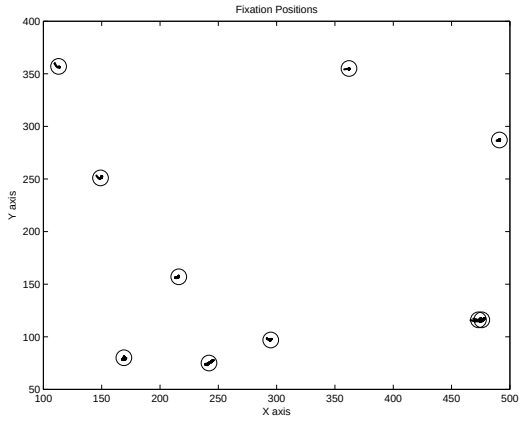
Methods to separate eye movement data into fixations and saccades can be broadly classified into two main categories, namely spatial and temporal methods. Temporal methods use features such as duration of stay and local adaptivity of the eye, whereas spatial methods use spatial features of eye data. Interested readers can refer to detailed taxonomy by Dario *et al.*[16] We have used a technique known as velocity-threshold identification (I-VT) to do the separation. This technique exploits the fact that during saccades, eyes move at a greater speed than the threshold velocity.

In I-VT, classifying eye positions into fixations and saccades is done using the instantaneous velocity of eyes at corresponding time instances. The eye-tracker samples the eye position at regular time intervals. Hence, the distance between consecutive samples can be used as a measure of instantaneous speed of eye movement. Depending upon whether the speed is greater or less than the threshold, eye positions are labeled as saccades or fixations. It is important to get a correct value of the threshold. Generally, researchers have worked with fixed thresholds such as ($< 100°$/sec) for fixations and ($> 300°$/sec) for saccades. In our work, we have used adaptive thresholds which is equal to an average of all the instantaneous speed values. The threshold value adjusts with the presence of drift in the fixation regions.



(a) Eye speed

(b) Eye positions



(c) Fixations

Figure 1. Obtaining fixation points from eye tracking data.

During saccades the eye moves quickly between two fixation regions. In the speed plot (Figure 1), we observe distinct peaks during a saccade. Although small peaks are observed during fixation regions, corresponding highest values fall below the threshold. These small peaks correspond to eye drift during fixations. An adaptive threshold was found to be effective by correctly differentiating between saccade peaks and drift peaks. In a few trials, fixations of small sizes, called blips, are obtained. Blips containing less than 30 samples are too small to be significant and are ignored in the analysis.

4. SALIENCY

4.1 Contrast Saliency

Saliency is a concept which states that there are regions in a scene that are more "attractive" than their neighbors and hence draw attention. Attention can be due to bottom-up cues or top-down influences. Bottom-up cues include local image characteristics from the stimuli that the HVS processes as and when they occur, e.g. local image contrast, presence of edges, orientation angles, color. Top-down influences refer to prior knowledge, like searching for a particular shape from among numerous objects. We use bottom-up cues and motion information to generate two saliency maps.

A saliency map of an image is a representation of the salient regions of the image. These are the regions that are most likely to draw attention. The conspicuity of a location in the visual scene determines the level of activity of the corresponding units in different feature maps. The saliency map combines the information of all the feature maps into one global measure of conspicuity. Saliency at a given location is determined primarily by how different this location is from its surroundings in terms of color, intensity, orientation, motion, depth etc. From the many locations that are salient the most salient region stands out.

The human visual system is sensitive to low frequency components with the cortical cells tuned to low level details like color, intensity and orientation edges. The Itti-Koch model[1] takes color, local contrast and orientation edges into account. For our work, since the visual stimuli are in gray scale, we disable the color part of the model and compute the saliency map based on orientation and intensity only. For the rest of this report, we call this map the contrast saliency map. Another significant deviation from the Itti-Koch model is that we do not implement a WTA (winner take all) algorithm to determine the most salient region in the image. The rationale behind this is that since we are focusing on the fixation regions and comparing their saliency values with a control value, we do not need to calculate the visually most salient part of the image. Moreover, we analyze video clips where the scenes keep changing and the WTA was deemed irrelevant.

The procedure for calculating the saliency map of a frame can be summarized in the following steps:

1. Linear filtering of the image at 8 spatial scales is done to extract features corresponding to luminance and edge properties.

2. For each individual feature, the spatial contrast map is calculated by taking the difference of the feature maps across several scales. A multiscale representation of the image ensures that we have information on the region of attention and its surroundings (center-surround). The contrast is calculated by taking the absolute difference of these scales:
$$I_s(k,l) = |I(k) \ominus I(l)|, \tag{1}$$
where $I(k)$ is the intensity map at scale k and $I(k,l)$ is the contrast map. The center is given by the levels $k \in \{1,2,3\}$ and the surround is given by levels $l = k + m, \quad m \in \{3,4\}$ in the Gaussian pyramid. An example is shown in Figure 2.

3. The feature maps are subjected to iterative convolutions with a Difference of Gaussian filter which suppresses isolated noisy regions.

4. Conspicuity maps of all the features are combined to get the final saliency map.

The saliency at a fixated point is the value of the saliency map at that location averaged over the number of frames for which the fixation persisted. To classify whether a fixated point is truly salient we compare its saliency value with a reference called the "control" value. The control is calculated in the following manner. For each of the frames corresponding to one fixation we take a patch of 30×30 pixels around each sampled position and sum up all the pixel values in the saliency map. This value is then divided by the number of pixels in the patch (900) and the number of eye positions in the frame. Thus we end up getting a saliency value per pixel per eye position for one frame. We repeat the calculation for all the frames corresponding to a given fixation and take the average. This value is representative of the average saliency of the frame. If the saliency value of the fixated location is greater than the average saliency of the frame we consider it salient. A higher control value might be due to the fact that motion attracted attention or because of top down factors.

4.2 Motion Saliency

We define *motion saliency* as attention due to motion. Since the stimuli are video clips, motion is undoubtedly a strong factor in capturing the viewers attention. Abrams *et al.*[9] have shown that onset of motion captures attention. Models have been developed that emulate the response of the middle temporal (MT) area of the primate cortex which is selective of the velocity in visual stimuli.[17] Attention models have been used to not only sense and analyze eye movements, but also guide them by using a special kind of gaze-contingent information display.[18]

We use spatial and temporal coherency[15] for computing a motion attention map, highlighting regions that are moving and are visually salient. The steps involved in the process can be summarized as follows:

1. Motion Vectors

   Motion vectors are an integral part of many video compression algorithms, where they are used for motion compensation. The idea behind block matching is to divide the current frame into a matrix of blocks that are then compared with the corresponding block and its neighbors in the previous frame to determine a motion vector that estimates the movement of a block from one frame to another.

   We calculate the motion vectors using the Adaptive Rood Pattern Search (ARPS).[19] The ARPS algorithm makes use of the fact that the general motion in a frame is usually coherent, i.e. if the blocks around the current block moved in a particular direction then there is a high probability that the current block will also have a similar motion vector. This algorithm uses the motion vector of the macro block to its immediate left to predict its own motion vector.

2. Motion Intensity

   The motion intensity I_t, a measure of induced motion energy or activity, is computed as the normalized magnitude of motion vectors:

   $$I_t(x,y) = \frac{\sqrt{dx^2{}_{x,y} + dy^2{}_{x,y}}}{MaxMag}, \tag{2}$$

   where $dx_{x,y}, dy_{x,y}$ denote the components of motion vectors at location x, y, and $MaxMag$ is the maximum magnitude in the motion vector field.

3. Spatial Coherency

   Spatial coherency indicates which are the blocks of pixels that belong to the same object. This is achieved by calculating the entropy over a block of pixels. The higher the entropy the smaller the probability of that particular block belonging to the same object. The rationale behind this is simple. Entropy gives a measure of randomness in the image. The greater the entropy, the higher the randomness. In this case lower entropy implies greater relationship between pixels and thus a higher probability of the group of pixels belonging to the same object.

   Thus the spatial coherency at pixel x, y, considering a window of size 8×8, is given by

   $$C_s(x,y) = -\sum_{i=1}^{N} p_s(i) \log p_s(i), \tag{3}$$

where $p_s(i)$ is the probability of occurrence of pixel intensity i and $N = 8 \times 8$.

4. Temporal Coherency

   The motivation behind the temporal coherency map is the same as that of the spatial coherency map. Instead of a block of pixels we have a group of pixels over different image frames. Higher entropy implies greater motion and hence a higher measure of saliency. We analyze a maximum of 7 frames prior to the frame corresponding to the fixated location and compute the entropy for all the pixel points:

$$C_t(x, y) = -\sum_{i=1}^{M} p_t(i) \log p_t(i), \tag{4}$$

   where $p_t(i)$ is the probability of occurrence of pixel intensity i at the corresponding location for different frames. For the temporal coherency map $M = 7$, as the effect of motion in one frame on the scan path of the eye lasts for 5-7 frames.

5. Combining the Maps

   The three maps are combined into motion saliency as follows:

$$B = I_t \times C_t \times (1 - I_t \times C_s), \tag{5}$$

   which is used for all further analysis. In the temporal coherency map higher entropy implies greater motion over that particular area. Since our aim is to determine *motion salient* regions, this is a direct indicator of interesting regions. For this reason the intensity map is multiplied with the temporal coherency map. However, in the spatial coherency map greater entropy indicates disparate objects. Our aim is to group objects together. This is the justification for the third term, which in essence assigns higher value to pixels belonging to the same object. Thus the output of the motion saliency map are regions in the image that belong to one object and are in motion.

An example of the various maps for a frame from one of our test clips is shown in Figure 2. The value of motion saliency is the value of motion attention map at the fixated location. The control value is calculated in a method similar to the one used to calculate control for contrast saliency above.

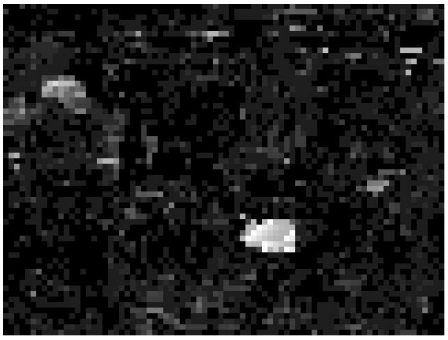
5. RESULTS AND ANALYSIS

The focus of our work is to investigate the role of motion in attention when watching videos. A very elementary approach is to analyze the role of motion alone. We tried to segment areas of the frame according to their motion magnitude, calculated from the motion vectors obtained. There were two approaches to this: First we assigned each block of pixels their overall magnitude and generated a contrast map, as described in Eq. (1). We repeated the same procedure using the orientation of the vectors. The results of both methods were similar but not very encouraging. In another variation of the same method we normalized the motion magnitudes to the range $[0, 1]$ and assigned different intensity values to a range of motion values. Thus we converted the motion map to an intensity map and segmented the frame according to the intensity values, or equivalently motion magnitude. The results – although better than in the previous case – were still unsatisfactory, because they were not coherent with the observed pattern of fixations. Thus we conclude that the magnitude of motion alone is not a very influential factor in grabbing attention.

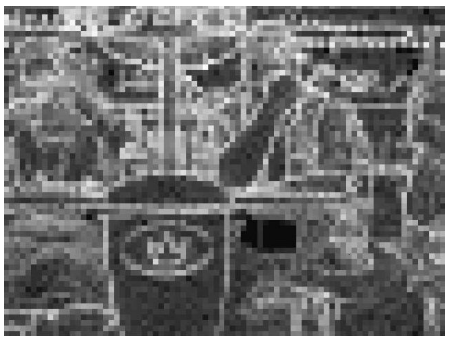
To group together moving objects we explored the concept of coherency. The results obtained using temporal and spatial coherency maps were significantly better, thus establishing the fact that motion is relevant when it is coherent over space and time. Random motion, even of high magnitude with respect to others, is not very salient.

We analyzed a total of 1231 trials over 10 movies with a total of 104,635 frames. Out of a total of 8553 fixation points, we characterize those that are contrast salient and motion salient. Some of the fixations are due to contrast saliency as well as motion saliency. The number of fixations due to each individual saliency

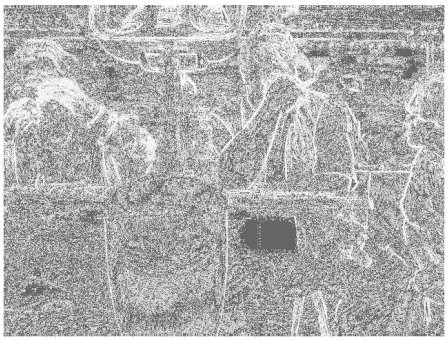
(a) Original Frame



(b) Intensity/Contrast



(c) Spatial Coherency



(d) Temporal Coherency

(e) Motion Saliency

Figure 2. Different components of the motion saliency map for the frame shown in (a).

measure was determined to get an idea about their influences in attracting attention. Table 1 shows the results while using only motion vectors to characterize motion saliency. 81% of all fixations were salient and out of the salient fixations, there were 34% fixations that were both motion salient as well as contrast salient. There were less motion salient fixations 56% than contrast salient fixations 70%. 21.5% of fixations were exclusivelt motion salient and the number of exclusively contrast salient fixations were 24.5%.

From Table 2 we observe that 85% of all fixations were salient. Out of the salient fixations, there were 44% fixations that were both motion salient as well as contrast salient. There were more motion salient fixations 74% than contrast salient fixations 70%, leading us to conclude that motion is a dominant cue over contrast (intensity and edge contrast) in attracting attention in videos. This is supported by the number of fixations that were exclusively motion 25.5% and exclusively contrast salient 22%.The number of contrast salient fixations is the same as in Table 1 as we have used the same model to calculate the contrast map. Using coherency to calculate a motion attention map shows a greater number of fixation points as being motion salient, which agrees with our observations.

Table 1. Fixation statistics when using motion vectors alone

Movie clip	# trials	contrast salient	motion salient	contrast & motion salient	only contrast salient	only motion salient	total salient fixations	total fixations	% salient
Biglebowski1	136	617	406	239	378	167	784	1050	75
Biglebowski2	100	356	201	113	243	88	444	645	69
Everest1	159	792	909	525	267	384	1276	1412	83
Everest2	101	429	300	220	209	80	509	659	77
Galapagos1	100	442	506	309	133	197	639	741	86
Galapagos2	130	442	571	297	145	174	616	714	86
Galapagos3	101	318	299	212	106	87	405	479	85
Galapagos4	112	414	399	258	156	141	555	674	82
Galapagos5	112	487	531	341	146	190	677	803	84
Galapagos6	180	779	791	460	319	331	1109	1376	81
	1231	**5076**	**4813**	**2974**	**2102**	**1839**	**6915**	**8553**	**81**

Table 2. Fixation statistics when using coherency.

Movie clip	# trials	contrast salient	motion salient	contrast & motion salient	only contrast salient	only motion salient	total salient fixations	total fixations	% salient
Biglebowski1	136	617	456	255	362	201	818	1050	78
Biglebowski2	100	356	258	146	210	112	468	645	73
Everest1	159	792	999	581	211	418	1210	1412	86
Everest2	101	429	344	236	193	108	537	659	80
Galapagos1	100	442	558	322	120	236	678	741	92
Galapagos2	130	442	520	301	141	219	661	714	93
Galapagos3	101	318	336	222	96	114	432	479	90
Galapagos4	112	414	448	273	141	175	589	674	87
Galapagos5	112	487	589	362	125	227	714	803	89
Galapagos6	180	779	864	497	282	367	1146	1376	83
	1231	**5076**	**5372**	**3195**	**1881**	**2177**	**7253**	**8553**	**85**

The data in Table 2 show some interesting trends. Firstly, the clips 'Biglebowski1' and 'Biglebowski2' had an unusually low percentage of salient points. The clips had a lot of clutter and objects that could have caught the monkeys attention even if they were not salient. For example there were instances when the monkey fixated at trees in the background. Balls and bottles were other common objects that caught the animal's attention. In a few clips the monkey tracked moving objects, like a rolling ball or a man floating in one of the scenes. These instances of tracking were a result of saliency due to moving objects and the observations agreed with our motion model wherein the tracked objects were identified as motion salient.

Another interesting observation was that if the scene did not change over a few clips then the fixations followed a pattern that were not coherent with the saliency model. This can be attributed to the influence of high level factors, which would explain prolonged fixation on trees and even fruits. We even observed instances where the animal kept staring at a particular location, ignoring any motion and low level features. This could be related to past experiences or even a result of monotony of the scene contents. A preference to look at faces was also observed which is a well documented fact.

The results show a significant influence of motion on attention. The rapid motion of a bird across the scene, movement of hands or objects drew immediate attention. Even ripples on the surface of water drew attention. Although the eye fixation data were that of a monkey, studies show that there is good correlation between overt fixations of human and monkey subjects.[20]

6. CONCLUSIONS

We have developed a motion attention model that shows which objects are moving and salient. We have used block matching to compute motion vectors for estimating the motion of blocks of pixels, which we found an efficient way to include motion in the framework. By incorporating spatial and temporal coherency with motion information, we were able to segment objects that were moving and salient. Our experiments show that motion is indeed an important cue vis-a-vis other low-level features in attracting attention when watching videos.

In future work, we want to focus on two issues with respect to motion and attention. The first are biological motion models that imitate the response of the middle temporal (MT) area of the primate cortex, which is selective of the velocity in visual stimuli.[17] Second, the observations from this data set underline the importance of texture. It has been estimated that texture contrast is approximately 10 times more important than intensity contrast in attracting human overt attention.[21] We shall look into the applicability of such models in our study.

ACKNOWLEDGMENTS

We thank Dr. Charles M. Gray at Montana State University (Bozeman) for sharing the eye tracking data with us. We thank Ajay Kumar Mishra and Suranga Chandima Nanayakkara for their help in implementing an earlier version of the algorithm.

This work was supported by grants from the Singapore Ministry of Education AcRF Tier 1 Fund (R-263-000-355-112, R-263-000-355-133), the NUS Cross-Faculty Research Grant, and the Singapore Science and Engineering Research Council.

REFERENCES

1. L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision Research* **40**, pp. 1489–1506, 2000.
2. J. M. Henderson, J. R. Brockmole, M. S. Castelhano, and M. Mack, "Visual saliency does not account for eye movements during visual search in real-world scenes," in *Eye Movements: A Window on Mind and Brain*, R. van Gompel, M. Fischer, W. Murray, and R. Hill, eds., pp. 537–562, Elsevier, 2007.
3. A. Yarbus, *Eye Movements and Vision*, Plenum Press, New York, NY, USA, 1967.
4. D. Melcher and E. Kowler, "Shapes, surfaces and saccades," *Vision Research* **39**, pp. 2929–2946, 1999.
5. D. Vishwanath and E. Kowler, "Localization of shapes: Eye movements and perception compared," *Vision Research* **43**, pp. 1637–1653, 2003.
6. P. Reinagel and A. Zador, "Natural scene statistics at the center of gaze," *Network: Computation in Neural Systems* **10**, pp. 341–350, 1999.
7. L. Itti and P. Baldi, "A principled approach to detecting surprising events in video," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 631–637, 2005.
8. L. Renninger, P. Verghese, and J. Coughlan, "Eye movements can be understood within an information theoretic framework," in *Computational and Systems Neuroscience*, 2005.
9. R. A. Abrams and S. E. Christ, "Motion onset captures attention," *Psychological Science* **14**(5), 2003.
10. C. Kayser, K. J. Nielsen, and N. K. Logothetis, "Fixations in natural scenes: Interaction of image structure and image content," *Vision Research* **46**(16), pp. 2535–2545, 2006.
11. V. Dragoi and M. Sur, "Image structure at the center of gaze during free viewing," *Journal of Cognitive Neuroscience* **18**, pp. 737–748, 2006.
12. B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision Research* **45**(5), pp. 643–659, 2005.
13. L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Processing* **13**(10), pp. 1304–1318, 2004.
14. R. Carmi and L. Itti, "The role of memory in guiding attention during natural vision," *Journal of Vision* **6**(9), 2006.
15. Y. Ma, L. Lu, H. Zhang, and M. Li, "A user attention model for video summarization," in *Proc. ACM Multimedia*, pp. 533–542, 2002.

16. D. D. Salvucci and J. H. Goldberg, "Identifying fixations and saccades in eye-tracking protocols.," in *Proc. ACM Symposium on Eye Tracking Research & Applications*, pp. 71–78, 2000.
17. E. P. Simoncelli and D. J. Heeger, "A model of neuronal responses in visual area MT," *Vision Research* **38**, pp. 743–761, March 1998.
18. E. Barth, M. Dorr, M. Böhme, K. R. Gegenfurtner, and T. Martinetz, "Guiding the mind's eye: Improving communication and vision by external control of the scanpath," in *Proc. SPIE Human Vision and Electronic Imaging*, **6057**, (San Jose, CA, USA), 2006.
19. Y. Nie and K.-H. Ma, "Adaptive rood pattern search for fast block-matching motion estimation," *IEEE Trans. Image Processing* **11**, pp. 1442–1448, December 2002.
20. W. Einhäuser, W. Kruse, K.-P. Hoffmann, and P. König, "Differences of monkey and human overt attention under natural conditions," *Vision Research* **46**, pp. 1194–1209, 2006.
21. D. Parkhurst and E. Niebur, "Texture contrast attracts overt visual attention in natural scenes," *European Journal of Neuroscience* **19**, pp. 783–789, 2004.