An Evaluation of Motion in Artificial Selective Attention

Trent J. Williams Bruce A. Draper
Colorado State University
Computer Science Department
Fort Collins, CO, U.S.A, 80523
E-mail: {trent, draper}@cs.colostate.edu

Abstract

The Difference of Gaussian (DoG) saliency maps originally proposed by Koch and Ullman had information channels for intensity, opponent colors, and edge orientations, but from the beginning it was suggested that additional channels could and should be added. This paper addresses selective attention in video sequences, and adds motion channels to the saliency maps. The resulting attention windows display better-than-random correlation to the eye fixations of human subjects.

This paper is not the first to add motion data to saliency maps. It is, however, the first paper we know of to explicitly compare the performance of saliency maps with and without a motion channel. The surprising negative result is that adding motion channels does not improve the performance of saliency-based selective attention, as measured by correspondence with human eye fixations. We draw two conclusions from this experiment: (1) although motion is clearly a critical attentional cue, the saliency map model may not extend as easily to motion data as some people (ourselves included) thought; and (2) the standard evaluation metric for selective attention algorithms – better than random correlation to human eye fixations – is outdated. Our community needs to focus on comparative studies between algorithms instead.

1. Introduction

One of the hallmarks of human vision is selective attention. We are not simply the passive observers of scenes, but instead actively seek out information through a complex, multi-step process of selective attention. The first step is *overt attention* in the form of eye and body movements that fixate our foveas on points in a scene. On every fixation, a two degree field of view window falls within the fovea and is therefore available for high resolution processing. The remaining steps of selective attention tend to be grouped together, and are collectively known as *covert attention* since they cannot be externally observed. There are reasons to believe that covert attention includes early spatial attention [23] followed by feature-based attention [5] and finally late attention [1]. Capacity theory [7] integrates these results by suggesting that vision is a multi-stage pipeline (or conveyor belt [2]), with limited resources at every step. Attention is the process by which data is reduced to fit the available resources at every step of the visual process (see also [14]).

Computer vision systems may use selective attention to select windows from larger images (e.g. [6, 26]). In this context, selective attention systems can be thought of as analogues to human overt attention and/or the first, spatial stage of human covert attention. Most of these systems employ a version of saliency maps, as first introduced by Koch and Ullman [12]. In this approach, the image is divided into low-level information channels, such as intensity, opponent color, and edge orientation channels. These channels are then filtered at multiple scales with Difference-of-Gaussians (DoG) filters that simulate on-center/off-surround processing. The resulting saliency values are summed across scales and channels to produce saliency maps. Attention windows are selected at the peaks of the saliency map, with an inhibition-of-return function to prevent repeatedly selecting the same window in temporal image sequences.

Saliency maps have been highly influential in the computer vision community, but the evaluation of saliency maps as a cognitive model has been lax. Most reported studies that evaluate saliency maps as cognitive models compare attention windows generated by saliency maps to random sequences of attention windows. This is a very low standard of comparison. Moreover, although many saliency-based attention systems have been implemented, too few studies measure the impact of specific information channels and/or algorithm steps (e.g. normalization).

This paper reports on an experiment in adding a motion channel to saliency-based bottom-up selective attention. The traditional saliency model of selective attention in

Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)
1063-6919/05 $20.00 © 2005 IEEE

still images uses seven information channels: one intensity channel, two opponent color channels, and four edge orientation channels. Increases in computing speed make it possible to apply selective attention to video sequences, however, not just still images, making it possible to introduce information channels based on motion, as recently done by Itti [8]. This makes intuitive sense, since there is wide-spread psychological support for motion being an important clue in bottom-up selective attention.

If the principles behind saliency maps are correct, then integrating motion should be a straight-forward process of convolving motion vectors with DoG filters at multiple scales, as with any source of information. The resulting system should do a better job of extracting attention windows that match human eye tracking data than systems without motion channels do. This paper, however, reports an experiment that does not conform to this prediction. Motion channels were added to a selective attention system. Although the resulting system produced better-than-random attention windows, there was no net improvement in the quality of the selected attention windows when compared to the same system without motion, as measured by correspondence to human data.

In computer science, negative experimental results must be interpreted with caution. It is not possible to exhaustively search the space of all motion detection algorithms and/or all computational variations on the saliency model. Perhaps another implementation will show an improvement when motion is added where this system did not. This experiment suggests, however, that adding motion to saliency maps may not be a straightforward process. It also calls into question the methods by which saliency maps are evaluated as cognitive models.

2. Related Work

Saliency-based selective attention was first introduced by Koch and Ullman [12] as a computational model of feature integration theory [23]. It has since been extended and refined by many researchers, including [10, 9, 18]), and a publicly available implementation called the Neuromophic Vision Toolkit (NVT) has been widely distributed. All versions of saliency models work with the same basic underlying principals. The image is first divided into independent "information channels", such as color, intensity, and edge orientation. The information in each channel is represented across scales as an image pyramid. For example, the intensity channel is extracted by calculating the intensity for every pixel in the source image, and then building an image pyramid from the resulting intensity image. Impulses are then detected in every channel by convolving the images in the pyramid with a Difference of Gaussians (DoG) mask. The absolute value of the impulse response at every pixel and scale is then interpreted as a measure of salience.

Attention systems vary in how salience values are integrated across information channels and scales. Koch and Ullman originally suggested multiple methods for normalizing the impulse responses before averaging them across scales for each channel [12]. The resulting channel-specific maps were then renormalized before being summed to create the final saliency map. Itti compared multiple normalization methods, before recommending the specific, highly non-linear scheme that is currently embedded in NVT [8]. Park et al combined salience maps using independent component analysis [18]. The version of selective attention analyzed in this paper was created by Draper and Lionelle [4], who determined empirically that the attention windows selected by NVT were highly sensitive to minor similarity transformations of the image. Their analysis suggested that this sensitivity was largely due to NVT's non-linear normalization step, and created SAFE, a selective attention system without non-linear normalization that is less sensitive to minor image transformations.

SAFE is also different from NVT in that it does not combine information across scales, but instead selects fixation points in a three-dimensional x, y, and $scale$ space. This is consistent with human overt attention, which fixates on a point in 3D space through vergence and focus. It is not clear whether human covert spatial attention selects scales as well as 2D locations, although there is some evidence that it may [16].

Saliency maps were introduced as a cognitive model of feature integration theory. At least two teams of researchers have claimed to find rough neural correlates to saliency maps in primate vision systems using single-cell recording techniques [3, 13]. More importantly for the purposes of this paper, there have been several attempts to compare the attention windows selected by NVT and related systems to human eye tracking data on the same images. Such studies represent an attempt to explicitly verify saliency maps as a model of human overt bottom-up attention.

In one such study, Parkhurst et al. compared attention windows selected by NVT to human eye tracking data on several types of still images [19]. They analyze attention windows as temporal sequences, which significantly complicated their analysis. Nonetheless, they were able to show that sequences of fixations in human eye tracking data correlated more closely to the fixations selected by NVT than to random sequences of points. More recently, Ouerhani et al. tried a simpler form of analysis [17]. They measured the salience values of pixels in still images as computed by NVT, and then compared the average salience of the points fixated on by human subjects to the salience of random fixation points. Again, they showed that NVT was a better then random predictor of human overt attention.

In the work that inspired this paper, Itti added a motion

channel to NVT, and applied the enhanced NVT to the task of finding attention windows in video clips [8]. He then compared human eye tracking data to the saliency values computed by NVT, using a comparison method similar to Ouerhani, et al. His conclusion was that motion-enhanced NVT was a better than random match to human data.

All of these studies compared the predictions of NVT to random attention windows, and found that NVT's predictions were a better match to human eye tracking data. This is a fairly low standard of comparison. Moreover, all of the studies above except Parkhurst et al. took an "all or nothing" approach, comparing the windows predicted by NVT over all information channels to human eye tracking data. Only Parkhurst et al. varied the relative weights of information channels, in order to find the configuration that most closely matched the human data. They found that the relative weights of information channels depended strongly on the image domain (perhaps as a result of top-down influences), but that in general, intensity and color channels were more significant than edge orientation channels (see Figure 7 of [19]).

Privetera and Stark [20] evaluated different functions for extracting impulses or other measures of salience from a single image channel. They filtered intensity images with 9 different functions, each of which could be considered salient, and compared human eye tracking data to the resulting measures. One of the functions tested was the Laplacian of the Gaussian, which is very similar to the DoG filter used in systems discussed above. They analyze the data by several different evaluation metrics, and do not conclude which filter is the best predictor of human eye tracking data. What is interesting in their study, however, is that almost all of the functions tested are better than random predictors of eye fixations, and that several functions tended to outperform Laplacians. This suggests a need for comparisons among attention algorithms, and fewer comparisons to random functions.

Finally, it should be noted that not all computational models of attention are based on Koch and Ullman's salience model. Another starting point is that saliency implies rarity, leading to the "rarity" based attention mechanisms of Schiele and Crowley [21] and Walker, et al. [25] Kadir and Brady present an entropy based selective attention system [11], while Tsotsos et al. present a neural network model for selective tuning [24]. Perhaps most interesting of all, Sun and Fisher present a computational model of feature-based (as opposed to spatial) attention [22].

3. Experiments

In this paper we add a motion channel to a selective attention system called SAFE [4] (see above). We have been using SAFE for a couple of years as a front end to object recognition, and wanted to extend it to video sequences in a manner similar to how Itti extended NVT [8]. At the same time, we wanted to measure the impact of motion data by comparing the performance of SAFE with and without a motion channel. As described below, the results are surprising: attention windows extracted with motion are no better than those extracted without motion in their ability to predict eye fixations.

We interpret this result cautiously. Motion is clearly a powerful bottom-up attentional cue for humans, and there may well be a way to integrate motion into saliency maps. There are many motion detection algorithms and several parameters with regard to saliency maps. We did not and could not explore all possible combinations. We do suggest that one straightforward extension of saliency maps to motion data does not work.

More importantly, this negative result exposes how poorly selective attention models are evaluated as models of human behavior. The standard measure of comparison is whether an attention algorithm is better than random at predicting human eye fixations. When SAFE is extended with motion as described below it easily passes this test, even though it performs no better than when the motion data is omitted. We should start de-emphasizing comparisons to random behavior, and emphasizing experiments which demonstrate progress by comparing new systems to established baselines.

3.1. Adding Motion to SAFE (version 1)

The general idea for adding motion to SAFE was to first extract motion vectors between successive frames of a video sequence, and to use these vectors as a new information channel. To do this, we first needed an algorithm to extract motion vectors. We chose a publicly available implementation of the well-known Lucas-Kanade motion detection algorithm[1] [15]. This algorithm takes two approximately registered images as inputs and returns a motion vector (magnitude and direction) for each pixel.

To apply the motion algorithm to color video sequences, we first extracted the individual frames from the videos and converted them to gray-scale images. The Lucas-Kanade algorithm was applied to temporally adjacent pairs of images to produce motion vectors. The magnitudes of the motion vectors were stored in the form of an image and was used as the base of an image pyramid in SAFE. As with all other channels, the images in the motion pyramid were convolved with a DoG filter, and the resulting salience values were added to the salience values of the other channels. In SAFE (unlike NVT), there is no non-linear normalization of image channels, and saliency values are not summed over

[1] available at http://www.cs.ucf.edu/ edu/ vision/source.html.

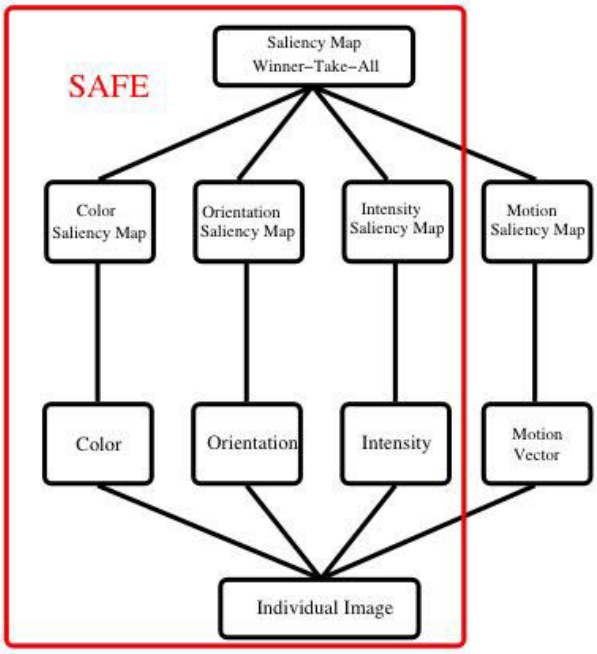

Figure 1. This is a visual depiction of the combination of existing SAFE still-image channels and the motion channel. The existing still-image SAFE channels are included in the box, and the motion channel is what has been added.

scales. Figure 1 shows a graphical representation of the existing SAFE system and the addition of the motion channel that we added. We will refer to the combination of SAFE and motion detection as SAFE-M.

To go into more detail, the Lucas-Kanade algorithm takes 3 parameters: the pyramid size, the size of the smoothing window, and the number of iterations. As a first pass, we used a pyramid level of 1, (typically 1-3), a window size of 2 (typically 1-4), and 2 iterations (typically 1-5).

4 Data

The test data for this experiment was provided by Dr. Laurent Itti at the University of Southern California, and is the same data used in [8]. The data consists of video clips of varying complexity and content, ranging from a circle moving across the screen with a constant background to kids running and playing in a park. All the videos have corresponding eye tracking data. The eye tracking data is the pixel fixation point for every frame in the video. There are fixation points for at least four subjects for every video clip. There are a total of eight subjects, but only five participants viewed all the clips. The data was recorded at 240Hz and the error was minimized by calibrating the eye tracking

total avg. eye movement per frame	5.83 +- 1.92
total distance	4.75 +- 3.73
avg. distance moved.	0.76 +- 0.53
avg. max distance per frame	3.64 +- 3.22

Table 1. These measurements (in pixels) show the statistics of all the subjects viewing all video frames. Average eye movement shows the number of times the eyes were recored to have moved during a single frame. Total distance is the Euclidean distance that the eyes moved from the start of the frame to the end of the frame. Average distance is the average distance all the subjects moved their eyes from one eye tracking sample to the next. Max Distance per frame is the average of the farthest two recorded points for a given frame.

equipment every 5 video clips during data collection. The video frames were displayed at 33.185ms/ frame, which means that there were approximately 30 frames displayed per second. Since the sampling rate of the eye-tracking system was faster than the frame display rate, there are approximately 8 eye tracking samples per frame. In addition to the x and y position that the subjects were looking at, the data also includes the "action" of the eye at the time of recording. The "action" is an indication of whether the eyes were fixated, pursuing an object, saccading, blinking, or blinking while saccading for every eye tracking sample. In extracting the raw eye data, we used only the eye data that was recorded while there was either fixation or smooth pursuit. The other actions do not reflect the actual bottom-up attention that we attempt to model so we did not use them. Further details on this data can be found in [8].

To justify using all 8 eye tracking samples for every frame, statistics were calculated on all of the videos. Table1 show the statistics of the eye tracking data across all the videos. For a given frame, we calculated the maximum distance that the eyes moved and averaged that value across all the frames. Across a total of 8190 frames, the eyes moved an average maximum distance of 3.641 +- 3.221 pixels. This means that, on average, the eyes moved only 3.641 pixels during the time period of one frame.

The frame size was constant across all videos (640 x 480 pixels). We tested 11 videos, ranging in length from 177 frames to 1651 frames (5 sec - 55 sec). Unfortunately, the videos are blurry and have a lot of noise. The artifacts in these clips most certainly had an impact on the performance results. Subjects, however, saw the same blurry images that were provided to the automatic attention systems.

4.1. Evaluation Methodology

The goal of the experiment was to evaluate the impact of incorporating motion into a bottom-up attentional model.

We therefore compare the performance of SAFE with motion data to the original version of SAFE with only still image channels (intensity, opponents colors, and edge orientations). We also compare the performance of SAFE with motion to a version in which the still image channels have been removed, leaving only the motion channel.

Two measures were used to evaluate versions of SAFE. For a given version of the algorithm, one measure compares average salience values across the image to salience values within the fixation windows of the human subjects. If the salience values are meaningful, they should be higher inside fixation windows. The other measure compares the top N attention windows per frame as selected by SAFE to the fixation points of human subjects. For this data set, one degree of visual angle corresponds to 22 pixels on the viewing screen. We used 2 degrees as the approximate size of the fovea, implying that a window 44 pixels high and wide was used to represent the foveal viewing field on the screen. When comparing human eye fixations to selected attention windows, we assume that the fixation point and window match is the centers of both are within 44 pixels.

4.2. Results and Discussion

Table 2 shows that human subjects tend to fixate on points that SAFE indicates are salient. The average salience value for a pixel in the combined (motion + still) system was 12.12 ± 9.06, while the regions that the subjects attended to averaged 37.13. This implies that pixels within fixation windows have salience values that are on average 2.76 standard deviations above the norm. This is strongly better-than-random behavior, the traditional yardstick for evaluating attention systems. When only still image features are used, however, the average pixel salience is 10.49 ± 7.96, while pixels that are attended to have an average salience of 33.09 ± 10.49. This implies that pixels within fixation windows have salience values that are on average 2.84 standard deviations above the norm. In other words, according to this measure, although both versions of the system performed well above random, the version without the motion information was a better predictor of human fixations than the motion enhanced system.

When evaluated alone, the motion channel also produced salience values with above random performance, although the results are not nearly as good as either the combined or still feature versions. When only motion is used, the average salience of attended pixels is 1.39 standard deviations above the norm.

The second measure determines whether the top N attention windows extracted by SAFE correspond to the points that humans attended to. If SAFE's windows correspond to human fixation points, SAFE is doing a good job of picking out the salient points in the video. We varied the number of windows extracted per image in SAFE. We tested the top 1, 5, 10, 20, 30, 40 and 50 windows per frame, and ranked salience values. Figure 2 shows the percentage of human fixation windows that are within the top N attention windows selected by SAFE. (There are an average of 7.36 human fixation points per frame, the result of both multiple subjects and the fact that subjects can saccade during a frame, producing multiple fixation points.)

Parameters	Human	SAFE (ver. 1)
combined:	37.13 +- 20.54	12.124 += 9.05864
original:	33.09 +- 19.35	10.487 += 7.96268
motion:	10.36 +- 5.79	5.092 += 3.804155

Table 2. The table compares the average salient values of the fixation points to the average of all the salient values.

By this measure, the still image (original) channels consistently outperform both the combined and motion channels for all 7 parameters. In other words, the still image system is more likely to generate attention windows that overlap human fixation points. Fig 2 shows that the addition of motion, in general, actually decreases the correspondence of suggested salience and human fixation points.

Both evaluation measures therefore agree on a conclusion that surprised us: adding motion to SAFE did not improve its performance as measured by comparison to human eye fixations. In fact, it may have made it worse. This seemed to contrast with Itti's result, which showed that adding motion to NVT produced much better than random attention windows.

Several possibilities occurred to us. We had tested a single parameter setting of the Lucas-Kanade motion algorithm. We therefore repeated the experiments on two video sequences using three different sets of parameters. The results are shown in Table 3 and Figure 3. It turns out that there is no qualitative difference in the major finding: the system still performs as well or better without the motion channel.

We also worried whether we had defined the motion channel correctly. Edge orientation information is usually incorporated into salience maps by using multiple subchannels, one for each edge orientation. Motion vectors are two dimensional data sources, and we had combined the information into a single motion magnitude channel. Perhaps this was incorrect. We therefore repeated the experiments on two videos again, this time using two motion channels, one for motion in the x direction, the other for motion in the y direction. As Table 4 and Figure 4 show, there is no qualitative difference in performance once again. Even with two directional motion channels, the motion information does not improve performance.

Video 1	Human	SAFE (ver. 2)	Human	SAFE (ver. 3)	Human	SAFE (ver. 4)
combined:	27.83 +- 6.33	10.22 +- 6.67	28.03 +- 6.44	10.39 +- 6.68	28.13 +- 7.08	10.55 +- 1.09
original:	24.58 +- 5.38	8.75 +- 5.75	24.58 +- 5.38	8.75 +- 5.76	24.58 +- 5.38	8.75 +- 5.76
motion:	8.34 +- 3.72	4.33 +- 3.23	8.51 +- 3.77	4.49 +- 3.32	8.64 +- 3.95	4.69 +- 2.39
Video 2	Human	SAFE (ver. 2)	Human	SAFE (ver. 3)	Human	SAFE (ver. 4)
combined:	70.58 +- 44.06	4.87 +- 13.26	73.16 +- 44.61	4.92 +- 13.51	82.35 +- 46.64	5.19 +- 1.15
original:	67.19 +- 44.26	4.64 +- 11.91	67.19 +- 44.26	4.64 +- 11.91	67.19 +- 44.26	4.64 +- 11.91
motion:	14.86 +- 4.95	4.87 +- 4.10	15.68 +- 7.24	7.04 +- 5.22	25.21 +- 10.51	10.70 +- 5.07

Table 3. We compare the average salience value of a human fixation point to all the salience points in the images. This experiment explores the difference in performance in varying the parameters. (ver. 2: pyramid level 1, window size 2, iterations 5. , ver.3: pyramid level 1, window size: 4, iterations 2, ver. 4: pyramid level 3, window size 2, iterations 5)

Video 1	Human	SAFE (ver. 2)	Human	SAFE (ver. 3)	Human	SAFE (ver. 4)
combined:	35.82 +- 9.58	13.71 +- 9.65	35.51 +- 9.48	13.76 +- 9.63	34.77 +- 9.30	13.88 +- 9.74
original:	24.58 +- 5.39	8.75 +- 5.76	24.58 +- 5.39	8.75 +- 5.76	24.58 +- 5.38	8.75 +- 5.75
motion:	19.09 +- 7.53	8.94 +- 8.21	18.79 +- 7.54	8.95 +- 8.24	18.17 +- 7.41	9.12 +- 8.51
Video 2	Human	SAFE (ver. 2)	Human	SAFE (ver. 3)	Human	SAFE (ver. 4)
combined:	60.43 +- 37.72	8.65 +- 10.13	62.630 +- 37.79	8.69 +- 10.25	87.17 +- 44.36	9.00 +- 11.60
original:	67.19 +- 44.26	4.64 +- 11.91	67.19 +- 44.26	4.64 +- 11.91	67.19 +- 44.26	4.64 +- 11.91
motion:	28.42 +- 9.78	8.79 +- 7.81	29.82 +-10.61	8.82 +- 7.86	43.59 +- 20.35	9.12 +- 8.57

Table 4. We ran the same experiments using 2 motion channels and compared the average salience value of a human fixation point to all the salience points in the images. (ver. 2: pyramid level 1, window size 2, iterations 5. , ver.3: pyramid level 1, window size 4, iterations 2, ver. 4: pyramid level 3, window size 2, iterations 5)

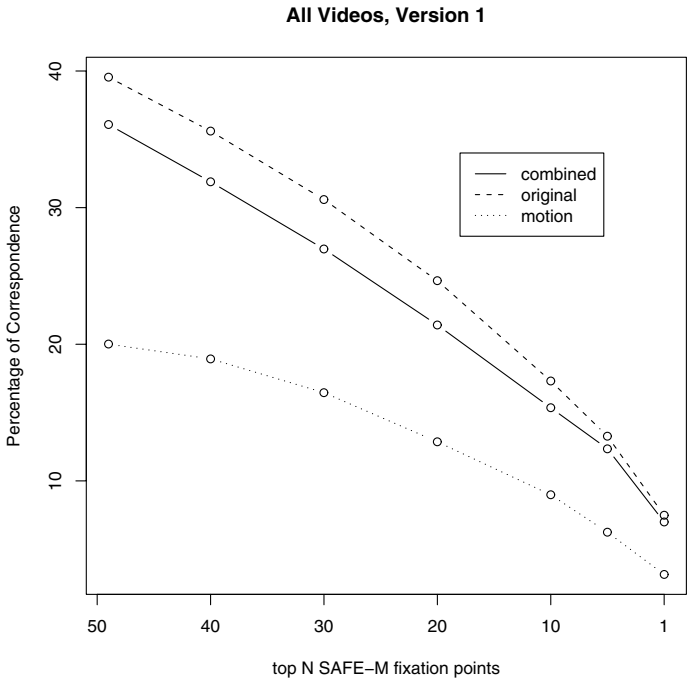

Figure 2. This graph shows the percentages of the corresponding top 50, 40, 30,20,10,5,1 ranked salient values by the SAFE-M system and whether or not the subjects were attending to one of those top N regions in the video.

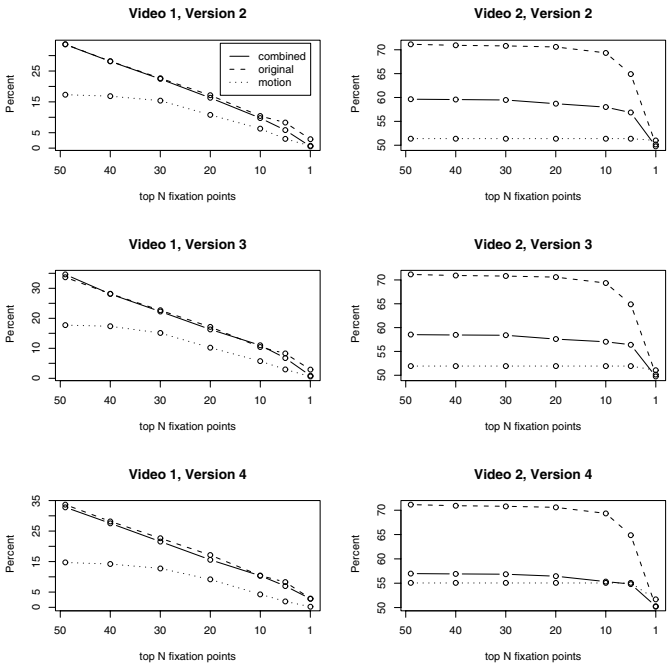

Figure 3. The graph shows the correspondence using 1 motion channel. We used 3 different parameter settings on 2 different videos. (ver. 2: pyramid level 1, window size 2, iterations 5, ver.3: pyramid level 1, window size 4, iterations 2, ver. 4: pyramid level 3, window size 2, iterations 5)

There remain many differences between this experiment and the one performed by Itti. The most obvious difference is the difference in salience systems (NVT vs SAFE). There are also differences in how the motion vectors are computed. Interestingly, however, Itti never compared NVT with motion to NVT without motion. His paper simply showed that NVT with motion produces much better than random attention windows.

5. Conclusion

Negative results in computer science must be viewed with caution. This paper shows that one attempt to incorporate motion into saliency maps failed, in the sense that motion data did not improvement the system's predictions of human eye fixations. There is no way to know, however, whether a different motion extraction system or different parameters within the saliency algorithm might have produced a different result.

There are two conclusions we can draw from this work, however. The first is that the assumption that motion can be easily incorporated into saliency maps should be questioned. It may well be possible (in fact, we believe that it is), but clearly the details of exactly how motion is integrated matter. Not all methods of integrating motion data will work.

More significantly, the measures typically used to evaluate selective attention systems are outdated. The standard measure used is better than average correlation to human eye fixations. By this measure, the integration of motion into SAFE was a success. Nonetheless, it is clear that the motion data did not improve the performance of SAFE, suggesting that the "better than random" measure of performance is inadequate. We need to start measuring performance relative to other selective attention algorithms, not random attention windows.

References

[1] *Inattentional Blindness*. MIT Press, Cambridge, MA., 2000.
[2] *Blackwell Handbook of Perception*, chapter Visual Attention, pages 272–310. Blackwell, 2001.
[3] C. Colby and M. Goldberg. Space and attention in parietal cortex. *Annual Review of Neuroscience*, 22:319–349, 1999.
[4] B. Draper and A. Lionelle. Evaluation of selective attention under similarity transformations. *Computer Vision and*

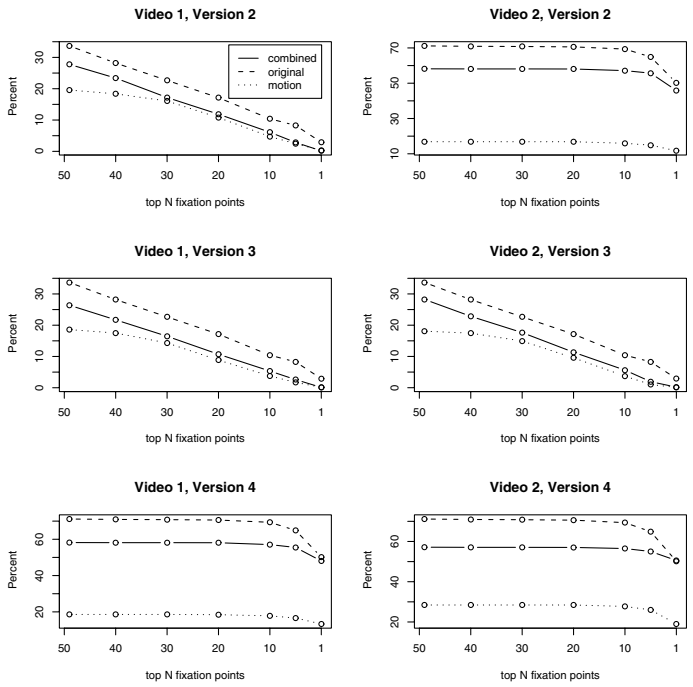

Figure 4. The graph shows the performance of correspondence using 2 motion channels: x and y. We used 3 different parameter settings on 2 different videos. (ver. 2: pyramid level 1, window size 2, iterations 5, ver.3: pyramid level 1, window size 4, iterations 2, ver. 4: pyramid level 3, window size 2, iterations 5)

Image Understanding, to appear, April 2005. An earlier version of this paper appeared at the Workshop on Attention and Performance in Computer Vision, Graz, Austria, 2003.

[5] J. Duncan and G. W. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458, 1989.

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 264–271, Madison, WI, June 2003. IEEE CS Press.

[7] T. C. Handy. Capacity theory as a model of cortical behavior. *Journal of Cognitive Neuroscience*, 12(6):1066–1069, 2000.

[8] L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions of Image Processing*, 13(10):1304–1317, October 2004.

[9] L. Itti and C. Koch. A comparison of feature combination strategies for saliency-based visual attention systems. SPIE human vision and electronic imaging IV (HVEI'99), 1999.

[10] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trasactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[11] T. Kadir and M. Brady. Scale, saliency and image description. *International Journal of Computer Vision*, 45(2):83–105, 2001.

[12] C. Koch and S. Ullman. Shifts is selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, 4:219–227, 1985.

[13] M. Kusunoki, J. Gottlieb, and M. Goldberg. The lateral intraparietal area as a salience map: the representation of abrupt onset, stimulus motion, and task revelance. *Vision Research*, pages 1459–1468, 2000.

[14] N. Lavie. Perceptual load as a necessary condition for selective attention. *Journal of Experimental Psychology: Human Perception and Performance*, 21(3):451–468, 1995.

[15] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision (darpa). In *Proceedings of the 1981 DARPA Image Understanding Workshop*, pages 121–130, April 1981.

[16] A. Oliva and P. G. Schyns. Coarse blobs or fine edges? evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology*, 34:72–107, 1997.

[17] N. Ouerhani, R. von Wartburg, H. Hugli, and R. Muri. Empirical validation of the saliency-based model of visual attention. *Electronic Letters on Computer Vision and Image Analysis*, 3(1):13–24, 2004.

[18] S.-J. Park, J.-K. Shin, and M. Lee. Biologically inspired saliency map model for bottom-up visual attention. In *International Workshop on Biologically Motivated Computer Vision*, pages 418–426, Tubingen, 2002. Springer-Verlag.

[19] D. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, 2002.

[20] C. M. Privetera and L. W. Stark. Algorithms for defining visual regions-of-interest: Comparison with eye fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(9):970–982, 2000.

[21] B. Schiele and J. L. Crowley. Object recognition using multidimensional receptive field histograms. In *European Conference on Computer Vision*, pages 610–619. Springer-Verlag, 1996.

[22] Y. Sun and R. Fisher. Object-based visual attention for computer vision. *Artificial Intelligence*, 146(1):77–123, 2003.

[23] A. M. Treisman and G. Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12:97–136, 1980.

[24] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo. Modeling visual attention vis selective tuning. *Artificial Intelligence*, 78(1-2):507–545, 1995.

[25] K. Walker, T. Cootes, and C. Taylor. Locating salient fcial features using image invariants. In *3rd International Conference on Face and Gesture Recognition*, pages 242–247, Nara, Japan, 1998. IEEE CS Press.

[26] D. Walther, L. Itti, M. Reisenhuber, T. Poggio, and C. Koch. Attentional selection for object recognition – a gentle way. In *International Workshop on Biologically Motivated Computer Vision*, pages 472–479, Tubingen, 2002. Springer-Verlag.