https://doi.org/10.2352/ISSN.2470-1173.2019.12.HVEI-205

Object-based and multi-frame motion information predict human eye movement patterns during video viewing

Zheng Ma, Jiaxin Wu, Sheng-hua Zhong, Stephen J. Heinen;
The Smith-Kettlewell Eye Research Institute, San Francisco, CA, USA
College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China

Abstract

Compared to low-level saliency, higher-level information better predicts human eye movement in static images. In the current study, we tested how both types of information predict eye movements while observers view videos. We generated multiple eye movement prediction maps based on low-level saliency features, as well as higher-level information that requires cognition, and therefore cannot be interpreted with only bottom-up processes. We investigated eye movement patterns to both static and dynamic features that contained either low- or higher-level information. We found that higher-level object-based and multi-frame motion information better predict human eye movement patterns than static saliency and two-frame motion information, and higher-level static and dynamic features provide equally good predictions. The results suggest that object-based processes and temporal integration of multiple video frames are essential to guide human eye movements during video viewing.

Introduction

Eye movement information is a reliable indicator of observers' attention allocation and regions of interest [1]. This is likely because visual acuity is highest in a small foveal region[2], and eye movements reorient the fovea to different scene elements that require high resolution viewing [3, 4]. Therefore, studying eye movement patterns can reveal observers' inner representation of the visual world.

Studies using static images show that although eye movement locations are somewhat predicted by low-level visual saliency, they are better predicted by higher-level semantic 'meaning' information in different spatial regions [5].

However, it is not known how low- level visual saliency and higher-level cognitive derived information drive human eye movements in dynamic stimuli such as videos. On one hand, there are many abrupt luminance changes and multiple sources of motion information across the entire visual field during video viewing. Therefore, it is plausible that these low level dynamic features predict eye movements. In fact, it has been shown that flicker and motion information provide better prediction than static features such as luminance and color [6]. On the other hand, videos provide richer semantic information than static images, and people might track the motion histories of different objects and follow the storyline of the entire video. Therefore, it is possible that higher-level cognitive-derived information better predicts gaze patterns. For example, both human and monkey gaze tends to cluster around biologically relevant social stimuli during video watching [7].

In the current study, we directly compare how low-level saliency and higher-level information guide eye movements during video watching. We collected human eye movement data while they watched videos depicting various life events. We then generated seven eye movement prediction maps based on different types of information, a Static Saliency Map, an Object Map, Two- and Multi-Frame Flicker Maps and Two- and Multi-Frame Optical Flow Maps, as well as a centering bias map. The Static Saliency Map reflect bottom-up saliency based on local differences in color, intensity and orientation. The Two-Frame Flicker Map and Optical Flow Map reflect bottom-up transient information that could be captured by low level motion detectors. On the other hand, the Object Map reflects information that can only be obtained after an object is recognized. Similarly, the two Multi-Frame dynamic features reflect information that allows memory and integration over longer time periods. We found that the Multi-Frame Flicker and Optical Flow Maps, as well as the Object Map, better predict human eye movements than the lower-level maps. We also found that the Object Map performed equally well as the two multi-frame dynamic feature maps, and provide significantly better prediction than the two two-frame dynamic feature maps. These results suggest that top-down, higher-level object-based information and temporal integration of the previous frames play an important role in guiding human eye movements during video watching.

Method

Human Eye Movement Data Collection

Six observers participated in the experiment. All had normal or corrected to normal vision. The protocol for the study was approved by the Institutional Review Board at the Smith-Kettlewell Eye Research Institute, and also adhered to the Declaration of Helsinki. Informed consent was obtained for experimentation with human observers.

We collected eye movement data while observers watched videos in an established video dataset, the SumMe dataset [8]. The SumMe dataset contains 25 videos depicting various life events such as kids playing, scuba diving and river crossing. The videos last from one to six minutes and are minimally edited. During the experiment, all of the original videos were resized to have the same width (1920 pixels; 43.6 degrees of visual angle). The audio was muted to ensure only visual stimuli guided the eye movements. Videos were presented using Psychtoolbox-3 [9, 10] for MATLAB.

The experiment was divided into six blocks, each of which contained 4-5 videos and lasted about 10 minutes. A chin rest was used to minimize head movements. Observers watched the entire videos without instructions or additional tasks. An Eyelink-1000 eye tracker recorded the location of the observer's right eye at 1000 Hz. At the beginning of each block, eye movements were calibrated using a standard 9-point calibration method. Since the temporal resolution of the eye movement data is higher than the frame rate of the videos (15 to 30 Hz), before the data were analyzed we averaged gaze positions of the same frame (33 to 67 gaze samples per frame) to obtain a single eye movement location of each video frame.

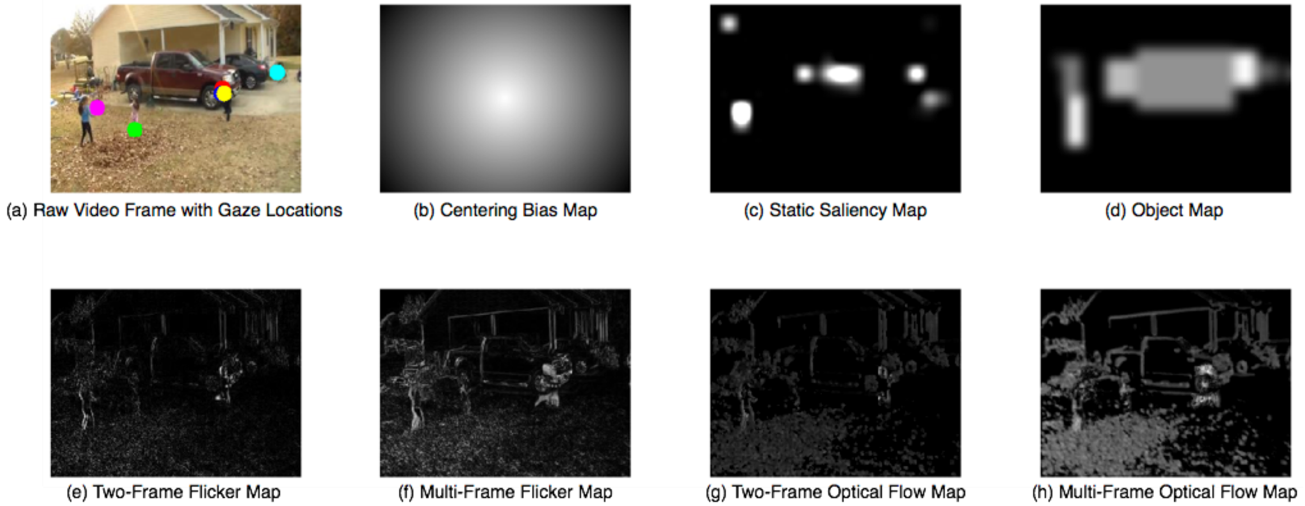

(a) Raw Video Frame with Gaze Locations
(b) Centering Bias Map
(c) Static Saliency Map
(d) Object Map
(e) Two-Frame Flicker Map
(f) Multi-Frame Flicker Map
(g) Two-Frame Optical Flow Map
(h) Multi-Frame Optical Flow Map

Figure 1. A sample video frame from the 'Kids Play in Leaves' video of the SumMe dataset (a), together with its different types of prediction maps (b to h). The colored dots in (a) show gaze locations of the six observers. Note that the gaze locations from three observers were very close to each other, as indicated by the almost overlapping red, blue and yellow dots. In prediction maps, higher intensity pixels indicate higher probability an eye movement was made to that location..

Computation of Eye Movement Prediction Maps

For each video frame, we generated seven maps to predict human gaze location. Each map was based on a different type of feature information. In each map, the higher the value assigned to a given pixel, the higher the probability that an eye movement was made to that location.

Centering Bias Map

People in most studies on video watching show a centering bias [11]. Therefore, as an initial validation of our methods, we asked whether our observers watching videos from the SumMe dataset showed this characteristic bias. To this end, we first created a Centering Bias Map (Figure 1b). For that map, the value of each pixel is proportional to the inverse of its Euclidean distance from the screen center. Therefore, in this map the center pixel of the screen has the largest value and the four corners of the video frame have the smallest one.

Static Saliency Map

The Static Saliency Map (Figure 1c) depicts the uniqueness of different locations in terms of low-level static features, including color, intensity, and orientation. It can be considered as a static map calculated through pure bottom-up processes and does not include temporal or features derived through cognitive processes. We ran the Saliency Toolbox [12], which implements an algorithm that computes salience at each location over the entire image [13]. For each pixel, higher values indicate more salient locations.

Object Map

Each Object Map (Figure 1d) is also derived from static information within a single video frame, but the map reflects higher-level cognitive processes beyond low-level features, since the objects must be recognized. The Object Map depicts locations that contain objects with semantic meaning in each frame. The SumMe dataset contains a total of 109,813 frames across the 25 videos. To obtain the Object Map in an efficient way, we used the Tensorflow Object Detection Application Programming Interface [14] to automatically detect objects that are presented in each video frame.

We ran the Faster R-CNN ResNet-101 model [15, 16] pre-trained on the COCO dataset [17], which detects the presence of and determines the locations of 80 different categories of objects, such as a person, a plane, or a table. Since all of the SumMe videos were recorded during real life events, the COCO categories cover most of the objects presented in the videos. We used the model-generated confidence score of the presence of an object at each pixel to create the scores for the Object Map. The higher the value, the more probable there is a meaningful object present at that location. We further applied a Gaussian filter of 1 degree of visual angle to smooth the rectangular boundaries of the detected objects.

Two- and Multi-Frame Flicker Map

Besides static low- and higher-level features, we also consider dynamic features. The first type of dynamic feature we considered is flicker. Flicker measures abrupt changes in luminance at the same pixel across contiguous video frames. To capture transient flicker information that is obtained through bottom-up processes, we computed a Two-Frame Flicker Map (Figure 1e), which contains the absolute luminance difference between the current frame and immediately preceding frame. We also computed a Multi-Frame Flicker Map (Figure 1f), which contains the maximum absolute luminance change across the previous five frames. Since the Multi-Frame Flicker Map characterizes luminance changes across a longer time interval than the Two-Frame Flicker Map, it may reflect temporal integration required for higher-level cognitive processes such as memory.

Two- and Multi-Frame Optical Flow Map

The second type of dynamic feature we considered was optical flow. In the computer vision literature, optical flow measures how motion information is distributed in an image [18]. To capture transient motion information that could be simply detected by lower-level motion detectors, for each frame, we computed the Two-Frame Optical Flow Map (Figure 1g) by applying the Horn and Schunck method [18] to calculate the optical flow between the current frame and the previous frame. To study the role of higher-level cognitive processes, we investigated how well motion information from multiple previous frames predicts eye movement patterns. To this end, we generated a Multi-Frame Optical Flow Map (Figure 1h) that integrates optical flow information across 10 previous frames. For example, to generate the Multi-Frame Optical Flow Map of the 10th frame of a video, two-frame optical flow was first computed for all of the first 10 frames, and then summed together to get the map of the 10th frame. This map tracks the longer range of motion information that could be obtained if the observer temporally integrated motion information across multiple frames.

Computation Details

For the sake of computational efficiency, we first converted all frames in the SumMe dataset to images of 320 by 240 pixels. We then generated each map using the methods described above. Finally, for each video frame, each map was normalized to the range of 0 to 1. A sample video frame with its different types of prediction maps are shown in Figure 1.

Evaluation of Eye Movement Prediction Maps

We used the receiver operating characteristic (ROC) method and area under curve (AUC) values to evaluate how well each map predicts human eye movements [19]. For each map, at each given threshold, if the value at a pixel is greater than the threshold, then the pixel is treated as a predicted gaze location. We then used the proportion of real gaze locations across all human observers included in the predicted area as the "true positive rate", and use the size of the predicted gaze area over the area of the entire image as the "false positive rate." We applied the ROC tool provided by the GBVS toolbox [20] to obtain the true and false positive rates at various different thresholds levels. We calculated the AUC value from the ROC curve to serve as a measure of how well the map predicts human gaze location. Higher AUC values suggest better predictability.

We first calculated the AUC values of each map for each video frame, then obtained the average AUC value for each movie. We were therefore able to treat each movie as a 'subject', and perform statistical tests among different maps to compare their ability to predict human eye movements.

Results

Consistent with previous studies [11], we found that the Centering Bias Map has the highest AUC value (0.812), and predicts human gaze patterns significantly better than all the other maps, suggesting that people have a strong bias to look at the center of the screen (all *p*s < 0.001).

The AUC values of the remaining six types of prediction maps are shown in Figure 2. A repeated one-way ANOVA showed a significant difference among the AUC values of the proposed prediction maps ($F(5, 120) = 14.1$, $p < 0.001$), suggesting that different maps differently predict human eye movements during video watching. We then used paired t-tests to directly compare different maps with each other and applied Bonferroni correction to correct for familywise error rates of multiple post-hoc comparisons.

The critical comparisons are between the low-level feature maps and higher-level maps. These comparisons were made for the static features and the two types of dynamic features respectively. For the static features, we found that the AUC value of the Static Saliency Map is significantly lower than the AUC value of the Object Map ($t(24) = 5.67$, $p < 0.001$). For the dynamic flicker features, the AUC value of the Two-Frame Flicker Map is significantly lower than the AUC value of the Multi-Frame Flicker Map ($t(24) = 7.52$, $p < 0.001$). Similarly, for the dynamic optical flow features, the AUC value of the Two-Frame Optical Flow Map is significantly lower than that of the Multi-Frame Optical Flow Map ($t(24) = 6.11$, $p < 0.001$). Together, these results suggest that for both static and dynamic features, information that can invoke higher-level cognitive processes can better predict eye movements than low-level features.

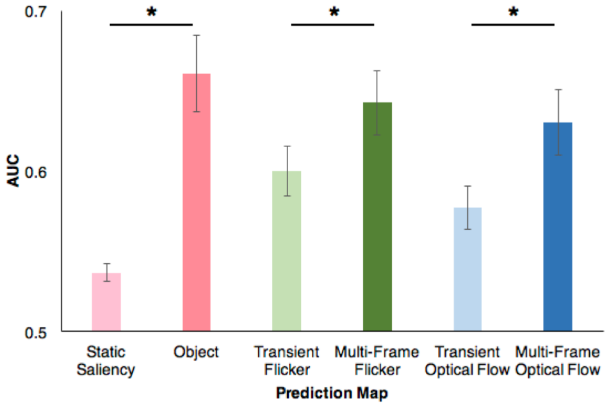


Figure 2. AUC values of different prediction maps. Error bars show standard error of the mean across the 25 videos in the SumMe dataset. Significant critical pairwise comparisons (p < 0.001) are indicated by *.

An additional question could be answered by the current data is how different the static and dynamic features are. Therefore, we further compared the predictability of static and dynamic features. The Static Saliency Map has a significant lower AUC value than the average of the two lower-level Two-Frame dynamic features ($t(24) = 3.39$, $p = 0.002$), and it is also lower than that of the average of the two higher-level Multi-Frame dynamic features ($t(24) = 4.91$, $p < 0.001$). These results suggest that static bottom-up salient features in videos do not predict human eye movement patterns as well as either bottom-up or higher-level dynamic features.

The story is different for the higher-level Object Map. The Object Map provides a better prediction than the average of the two lower-level Two-Frame dynamic features ($t(24) = 3.08$, $p = 0.005$). However, there is no significant difference between the Object Map and the pool of the two higher-level Multi-Frame dynamic features ($t(24) = 1.02$, $p = 0.32$). These results suggest that the three higher-level feature maps predicted human eye movements equally well.

Together, the results suggest that information obtained through higher-level cognitive processes can predict human eye movements during video viewing better than other information. However, bottom-up transient dynamic information still predicts eye movements better than those using only static features.

Discussion

Previous studies found that human observers tend to look at meaningful regions of static images [5]. For videos, low-level dynamic features such as flicker and optical flow can better predict human eye movements [6]. On the other hand, studies have also shown that people also tend to look at biologically relevant stimuli in videos [7]. There is still a lack of direct comparison in terms of how low-level and higher-level features can guide eye movements during video viewing.

In the current study, we recorded eye movements from observers while they watched videos with no additional instructions. We generated multiple eye movement prediction maps to determine whether low- or higher-level features better predict human eye movement patterns.

Importantly, we found that compared to feature information that can be obtained through bottom-up processes, information that requires higher-level cognitive processes better predicts human eye movement locations in video frames. This is true for both static and dynamic features. The locations of objects identified by an object detection algorithm provides better predictions than those identified by static saliency algorithms. The results suggest that object recognition is important for guiding observers' eye movements during video viewing. For dynamic features, flicker and optical flow across multiple frames may require higher level cognitive processes to store and integrate information. We found that these multi-frame dynamic features provide a better prediction than the two-frame dynamic features. Therefore, our results suggest that eye movements during video viewing are guided to a greater degree by higher-level temporal integration, rather than purely low-level abrupt changes in luminance and motion.

We also found that while the static saliency information predicts eye movements the poorest, static object-based information predicts them as well as the two types of long-range dynamic features that we tested. Together, the results indicate that during video watching, previous history of luminance change and motion information, as well as object semantics may be the best predictors of where people look while watching videos. Furthermore, the better prediction by motion features found in previous work (e.g. [6]) may be due to higher-level factors that longer duration motion conveys, such as object-based motion tracking, rather than simply low-level abrupt luminance change or motion signal.

The results of the current study potentially guide video compression and video retargeting for different devices. If we can better predict human gaze locations during video viewing, we can then select compression methods to ensure that visual quality is the highest at image areas where humans devote the most attention [21, 22].

References

[1] J. Henderson, "Human gaze control during real-world scene perception", Trends in Cognitive Sciences, vol. 7, no. 11, pp. 498-504, 2003.

[2] R. L. DeValois & K. K. DeValois, Spatial Vision, New York: Oxford University, 1988.

[3] N. H. Mackworth and A. J. Morandi, "The gaze selects informative details within pictures," Perception & Psychophysics, vol. 2, no. 11, pp. 547–552, 1967.

[4] J. Najemnik and W. S. Geisler, "Optimal eye movement strategies in visual search," Nature, vol. 434, no. 7031, pp. 387–391, 2005.

[5] J. M. Henderson and T. R. Hayes, "Meaning-based guidance of attention in scenes as revealed by meaning maps," Nature Human Behaviour, vol. 1, no. 10, pp. 743–747, 2017.

[6] P. K. Mital, T. J. Smith, R. L. Hill, and J. M. Henderson, "Clustering of Gaze During Dynamic Scene Viewing is Predicted by Motion," Cognitive Computation, vol. 3, no. 1, pp. 5–24, 2011.

[7] S. V. Shepherd, S. A. Steckenfinger, U. Hasson, and A. A. Ghazanfar, "Human-Monkey Gaze Correlations Reveal Convergent and Divergent Patterns of Movie Viewing," Current Biology, vol. 20, no. 7, pp. 649–656, 2010.

[8] M. Gygli, H. Grabner, H. Riemenschneider, and L. V. Gool, "Creating Summaries from User Videos," In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8695. Springer, Cham

[9] D. H. Brainard, "The Psychophysics Toolbox," Spatial Vision, vol. 10, pp. 433-436, 1997.

[10] D. G. Pelli, "The Video Toolbox software for visual psychophysics: Transforming numbers into movies," Spatil Vision, vol. 10, 437-442, 1997.

[11] V. Tosi, L. Mecacci, and E. Pasquali, "Scanning eye movements made when viewing film: Preliminary observations," International Journal of Neuroscience, vol. 92, no. 1-2, pp. 47–52, 1997.

[12] D. Walther and C. Koch, "Modeling attention to salient proto-objects," Neural Networks, vol. 19, no. 9, pp. 1395–1407, 2006.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 11, pp. 1254–1259, 1998.

[14] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/Accuracy Trade-Offs for Modern Convolutional Object Detectors," in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, Jan. 2017.

[16] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," in Advances in Neural Information Processing Systems, pp. 379-387, 2016

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in Computer Vision – ECCV 2014 Lecture Notes in Computer Science, pp. 740–755, 2014.

[18] B. K. Horn and B. G. Schunck, "Determining optical flow," Artificial Intelligence, vol. 17, no. 1-3, pp. 185–203, 1981.

[19] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search.," Psychological Review, vol. 113, no. 4, pp. 766–786, 2006.

[20] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency", in Proceedings of Neural Information Processing Systems (NIPS), 2006.

[21] U. Engelke, D.P. Darcy, G.H. Mulliken, S. Bosse, M.G. Martini, S. Arndt, J.N. Antons, K.Y. Chan, N. Ramzan, and K. Brunnström, Psychophysiology-Based QoE Assessment: A Survey. IEEE Journal of Selected Topics in Signal Processing, 2017. 11(1): p. 6-21.

[22] P. L. Callet, and E. Niebur, Visual Attention and Applications in Multimedia Technologies. Proceedings of the IEEE, 2013. 101(9): p. 2058-2067.

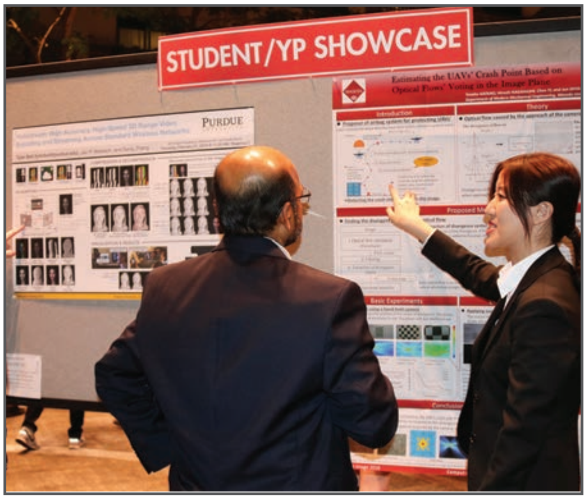
Author Biography

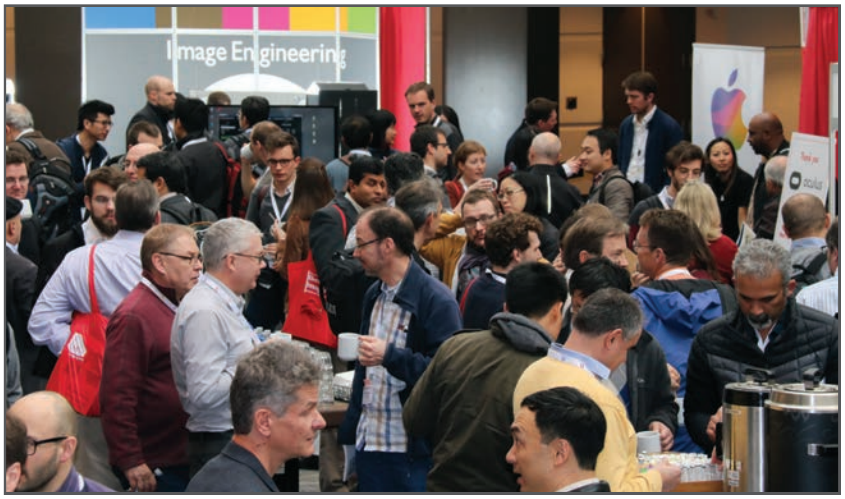
Zheng Ma received her BS in psychology from the Peking University (2011) and her MA and PhD in psychology from the Johns Hopkins University (2013, 2016). Since then she has been a postdoctoral fellow at the Smith-Kettlewell Eye Research Institute. Her work has focused on human eye movements in dynamic visual scenes and the interaction between perception and action.

Jiaxin Wu received her BS (2015) and MS (2018)in computer science and software engineering from Shenzhen University. Her work has focused on video content analysis and deep learning.

Sheng-hua Zhong received her BS in optical information science and technology from Nanjing University of Posts and Telecommunication (2005), her MS in signal and information processing from Shenzhen University (2007), and her PhD in computer science from the Hong Kong Polytechnic University. She was a postdoctoral fellow at the Johns Hopkins University (2013-2014). She is now an assistant professor at Shenzhen University. Her interests include multimedia content analysis, brain science, and machine learning.

Stephen J. Heinen received his dual BA in psychology and mathematics from the Wright State University (1981), and his MA and PhD in experimental psychology from the Northeastern University (1987, 1988). Since then she has been a scientist and then a senior scientist at the Smith-Kettlewell Eye Research Institute. His work has focused on the neural mechanisms of human visual perception and oculomotor system.





AR is a fundamental and transformative interface and platform.

20+ years from my initial working demo to today's AR displays

Optimistic that it won't take another 20 to reach ubiquitous consumer AR!