Active Fixation Control to Predict Saccade Sequences
Supplementary Material

Calden Wloka Iuliia Kotseruba John K. Tsotsos
Department of Electrical Engineering and Computer Science
York University, Toronto, Canada
calden, yulia_k, tsotsos@cse.yorku.ca

1. Retinal Transform

Our implementation follows the same steps as outlined in [3] with few minor changes. In particular, we refitted the generalized Gamma distribution to better simulate rod vision and reimplemented the interpolation function in CUDA.

As with most foveation algorithms, our approach starts by building a Gaussian pyramid and then for each pixel the appropriate level of the pyramid is sampled depending on how far the pixel is from the current gaze point. The level of the pyramid to sample from is computed as follows

$$L_{x,y} = \frac{\frac{\pi}{180}(\text{atan}((D_{\text{rad}} + \text{dotpitch})\frac{1}{D_{\text{view}}}) - \text{atan}((D_{\text{rad}} - \text{dotpitch})\frac{1}{D_{\text{view}}}))}{(\epsilon_2(\alpha * (EC + \epsilon_2))). * log(\frac{1}{CT_0})} \tag{1}$$

where D_{rad} is the radial distance between the point (x, y) and the current gaze point, EC is the eccentricity from the fovea center for point (x, y) in degrees, dotpitch is the size of the pixel of the monitor in meters and D_{view} refers to the viewing distance. α, ϵ_2 and CT_0 are constants from [2]. In this equation the numerator represents the maximum spatial frequency that can be represented at the given distance from the current gaze point and the denominator is the maximum spatial resolution that can be resolved by the eye.

However, [2] only considered cone vision. For more biologically accurate results we augment it with rod vision following [3] using the generalized Gamma distribution as proposed in [4]. Since in [4] cell counts were used to fit a distribution function, we adjust the parameters to convert it to the units that we use, namely the levels of pyramid. Therefore, we set the parameters of the generalized Gamma distribution as follows: $\alpha = 2.46, \beta = 121.8, \gamma = 0.77, \sigma = 861.27$ and $\mu = -1$. To find corresponding levels of the pyramid for each pixel we compute the generalized gamma distribution for EC and plug it into the equation (1) as the denominator.

Finally we compute the foveated image using a bi-cubic interpolation routine for 3D volumes to sample the required level of the pyramid for each pixel. Our code is a CUDA reimplementation of the ba_interp3 function [1].

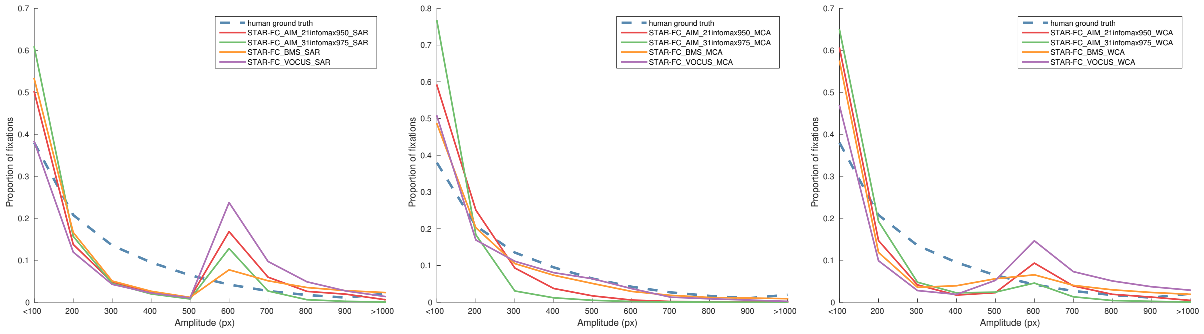
Note that we compute cone distribution for each color channel separately, but rod distribution only for the intensity channel (the image is first converted to YCrCb color space), since rods do not contribute to color vision. Rod and cone functions contribute 40% and 60%, respectively, to the final transformed image.

The viewing conditions in all our experiments match the experimental conditions reported for the CAT2000 dataset [1], namely all stimuli span 45 degrees and the viewing distance is set to 1.06 m.

2. Saccade amplitudes

Figure 1 shows plots of the fixation amplitudes that demonstrate the effect of using different saliency algorithms in the periphery and blending strategies. Both SAR and WCA blending strategies (Figure 1a and Figure 1c) lead to a pronounced spike in the distribution, which approximately corresponds to the diameter of the central field. MCA, on the other hand, produces a more even distribution of the amplitudes (Figure 1b). Figure 2 shows fixation amplitudes for all tested bottom-up saliency algorithms. Note that all of them greatly underestimate the number of short saccades (< 100 px) and generally have a much flatter fall off than the human ground truth distribution.

[1] https://www.mathworks.com/matlabcentral/fileexchange/21702-3d-volume-interpolation-with-ba-interp3--fast-interp3-replacement



(a) SAR

(b) MCA

(c) WCA

Figure 1: Plots of fixation amplitudes demonstrating the effects of different strategies for combining peripheral and central fields of STAR-FC (SAR, MCA and WCA), and different bottom-up saliency algorithms in the peripheral field (AIM, VOCUS and BMS).

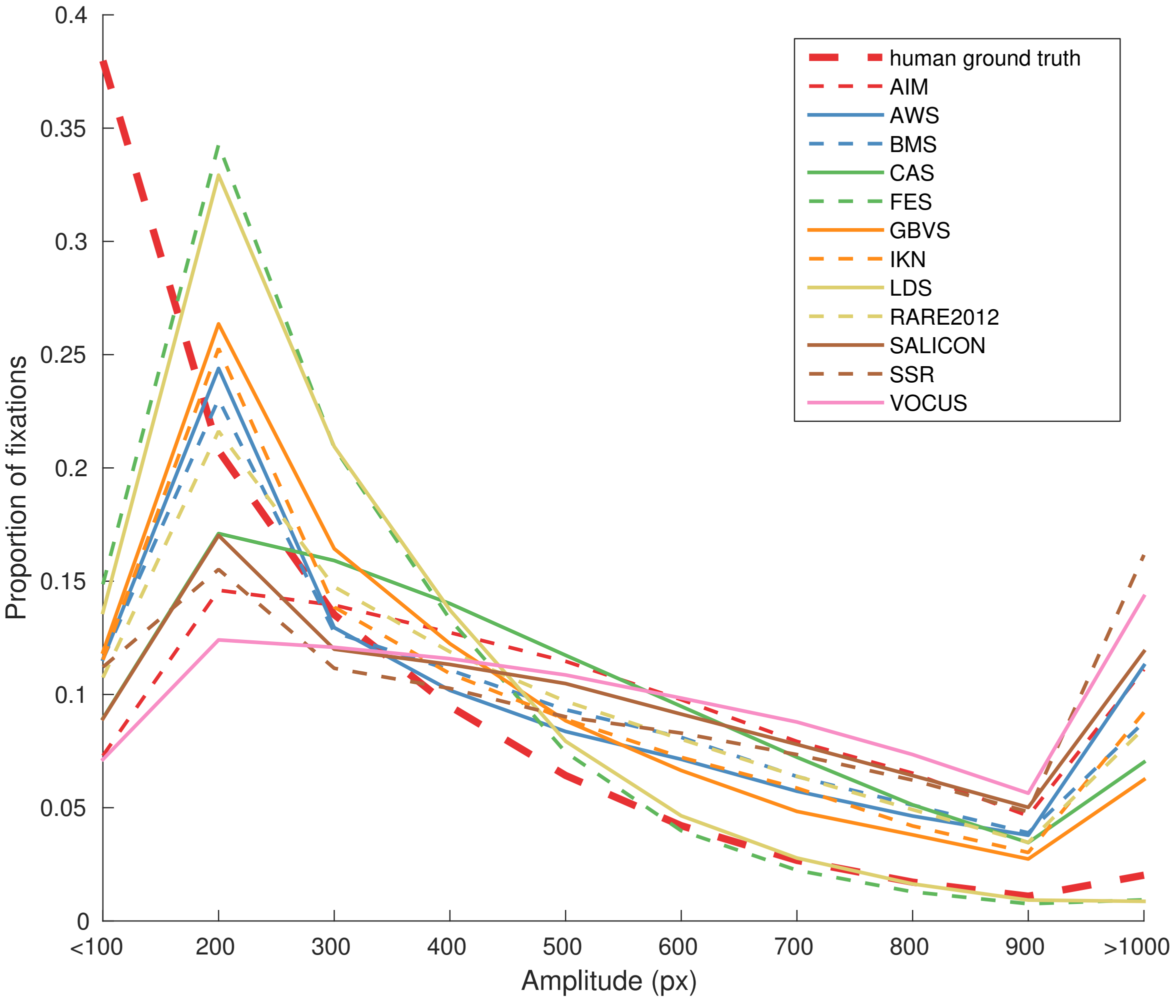
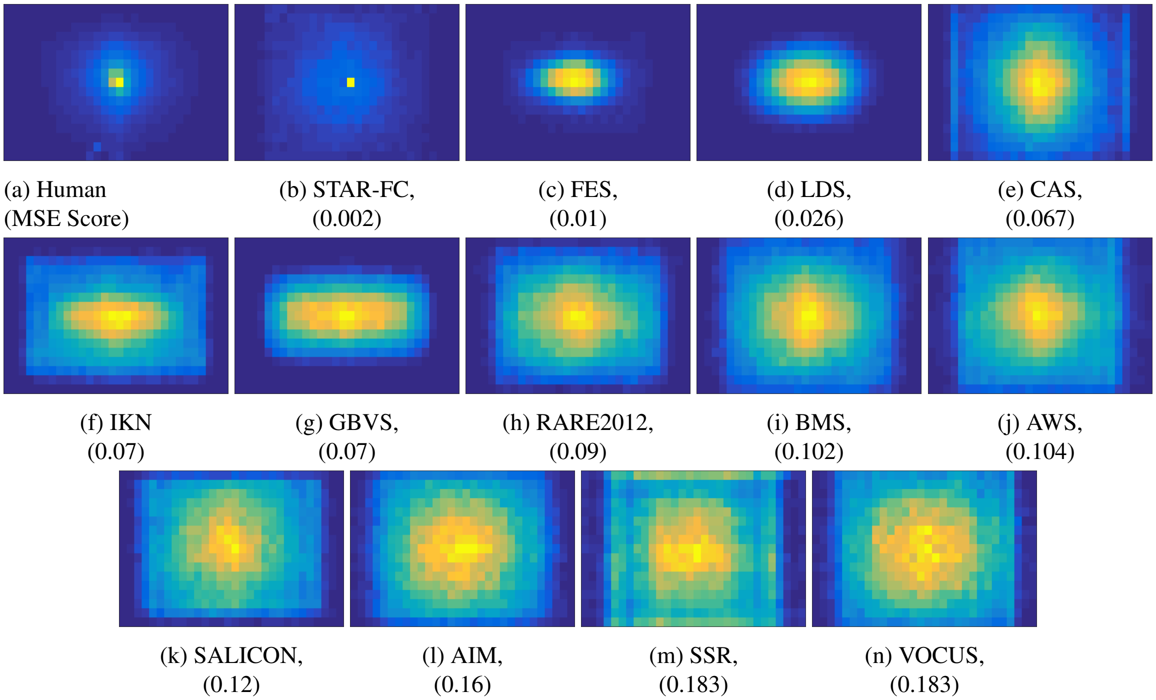


Figure 2: Fixation amplitudes for 12 state-of-the-art bottom-up saliency algorithms.

3. 2D Histograms of Fixations

Figure 3 shows 2D histograms of fixations for all saliency algorithms with MSE scores.



(a) Human (MSE Score)

(b) STAR-FC, (0.002)

(c) FES, (0.01)

(d) LDS, (0.026)

(e) CAS, (0.067)

(f) IKN (0.07)

(g) GBVS, (0.07)

(h) RARE2012, (0.09)

(i) BMS, (0.102)

(j) AWS, (0.104)

(k) SALICON, (0.12)

(l) AIM, (0.16)

(m) SSR, (0.183)

(n) VOCUS, (0.183)

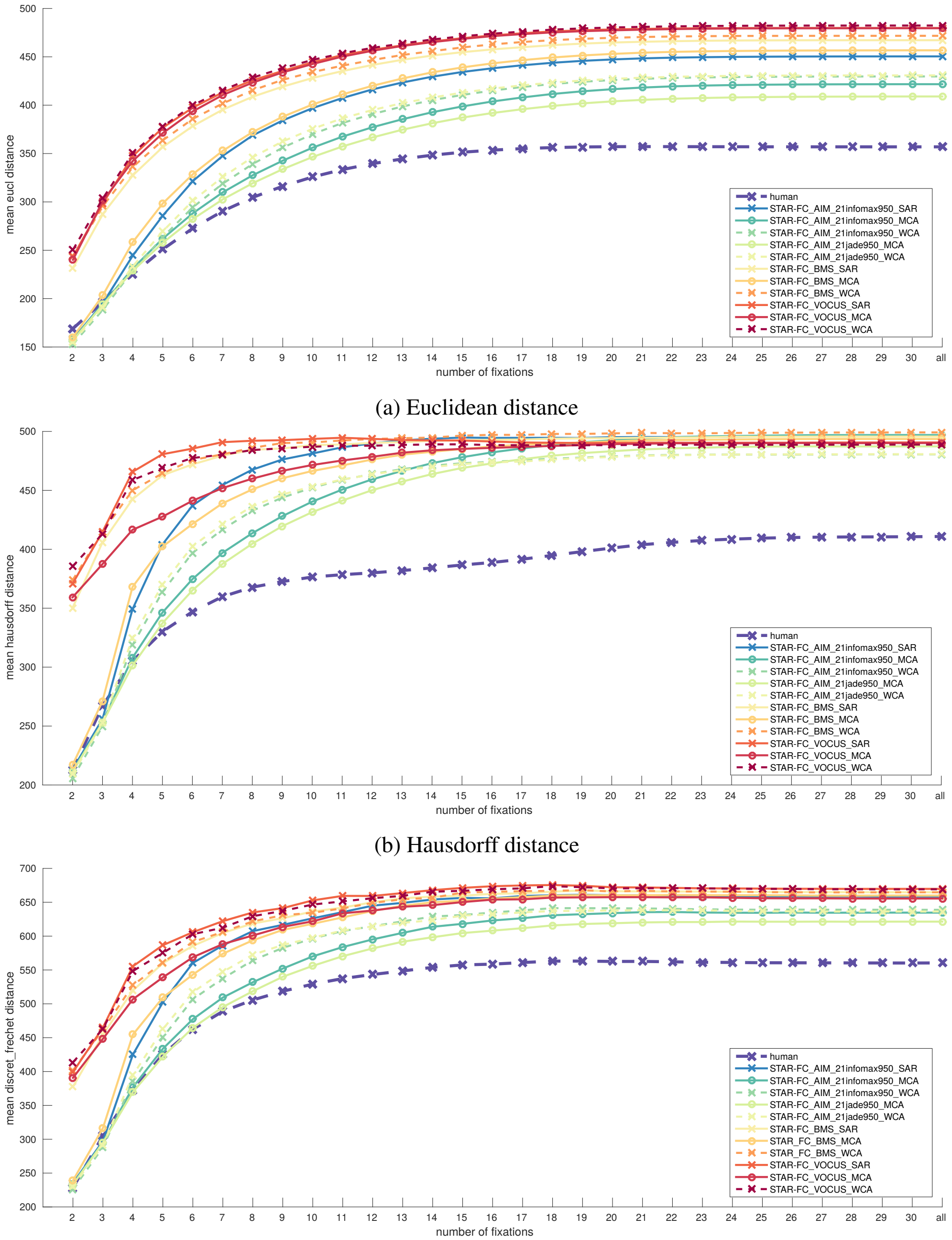
Figure 3: 2D histograms of fixation locations over the CAT2000 dataset for all tested bottom-up saliency algorithms. Mean-squared-error (MSE) scores between model and human distributions are shown in parentheses under each model name. The algorithms are sorted by MSE in ascending order, starting with STAR-FC (AIM with 21infomax950 basis and MCA blending strategy), which is an order of magnitude closer to the human distribution than the best bottom-up algorithm (FES).

4. Trajectory Scores

Figure 4 and Figure 5 show trajectory scores for full sequence length for various STAR-FC variants and all tested saliencey algorithms. Note that as the sequences get longer they begin to diverge and the trajectory scores saturate.

Figure 6 shows the AUC score for the first 5 fixations for all saliency algorithms and our best performing STAR-FC model (using AIM with 21infomax950 basis and MCA blending strategy) split by category. For human fixations we report the AUC for the average pairwise distance. As we noted in the paper, it correlates well with inter-observer consistency for different categories of images (e.g. high IO consistency for Sketch translates to a smaller average pairwise distances across all metrics, whereas the opposite is true for the Jumbled category).

Note that saliency algorithms tend to follow the same trends as human inter-observer scores. In general, categories with high IO consistency such as Affective or Sketch are not as challenging as categories with lower human to human fixation consistency. One major exception is the Low Resolution category, which has a high degree of IO consistency but is nevertheless extremely challenging for all saliency algorithms. This calls for more investigation of the effects that blurring has on the quality of saliency prediction.

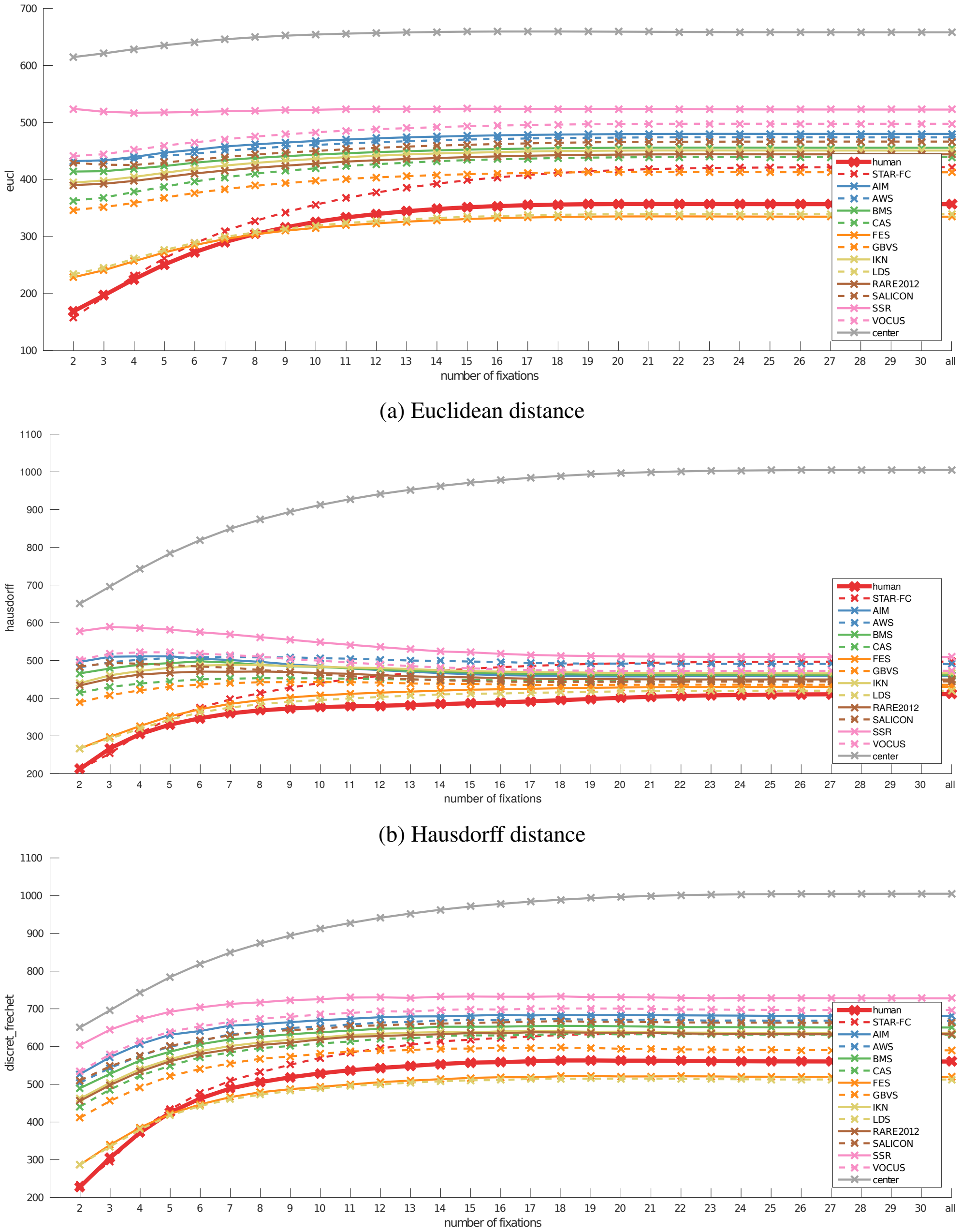


(a) Euclidean distance

(b) Hausdorff distance

(c) Frechet distance

Figure 4: A comparison of fixation prediction scores over the full length of the fixation sequences for variants of STAR-FC.

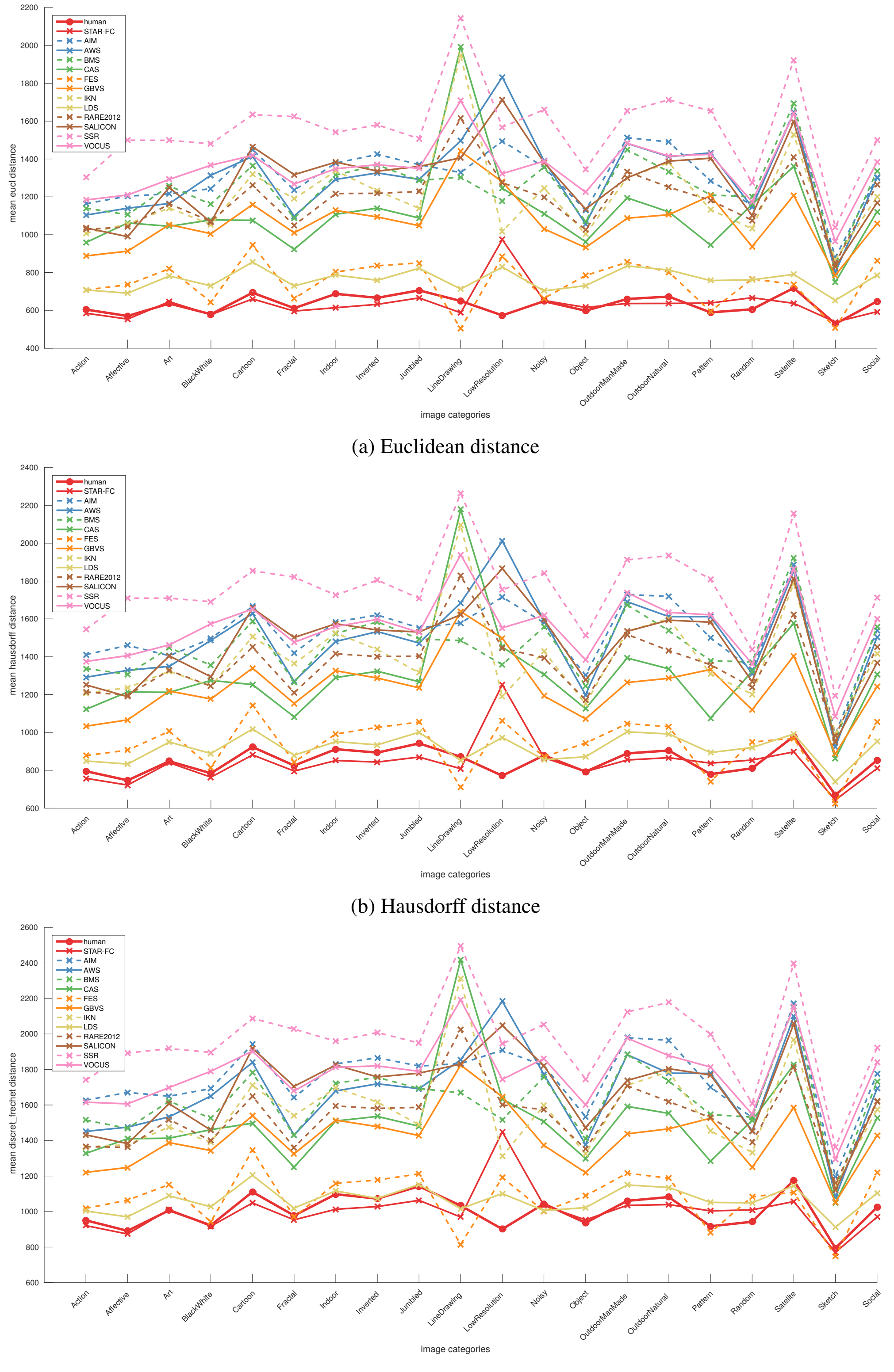


(a) Euclidean distance

(b) Hausdorff distance

(c) Frechet distance

Figure 5: A comparison of fixation prediction scores over the full length of the fixation sequences for all tested saliency algorithms and the best performing STAR-FC model (using AIM with 21infomax950 basis in the peripheral field and MCA blending strategy).



(a) Euclidean distance

(b) Hausdorff distance

(c) Frechet distance

Figure 6: Fixation prediction scores for all tested saliency algorithms and the best performing STAR-FC model (using AIM with 21infomax950 basis in the peripheral field and MCA blending strategy). For each category we measured the mean distance from the human fixation and plotted the area-under-the-curve (AUC) score for the first 5 fixations.

5. Examples of Predicted Fixation Sequences

Below we show some examples of predicted fixations. For clarity we only compare STAR-FC and one saliency algorithm at a time and show only the closest human sequences to each of the predicted sequences. Furthermore, we show results only for the first 3 and 5 fixations. We selected FES as the top performing contrast-based algorithm and SALICON as the top performing CNN-based algorithm for comparison with our best performing STAR-FC model (21infomax950 bases and MCA blending strategy).

In Figure 7 examples from categories with high IO consistency (Affective and Low Resolution) are shown. Figure 8 shows examples from the Satelite category which has low IO consistency.

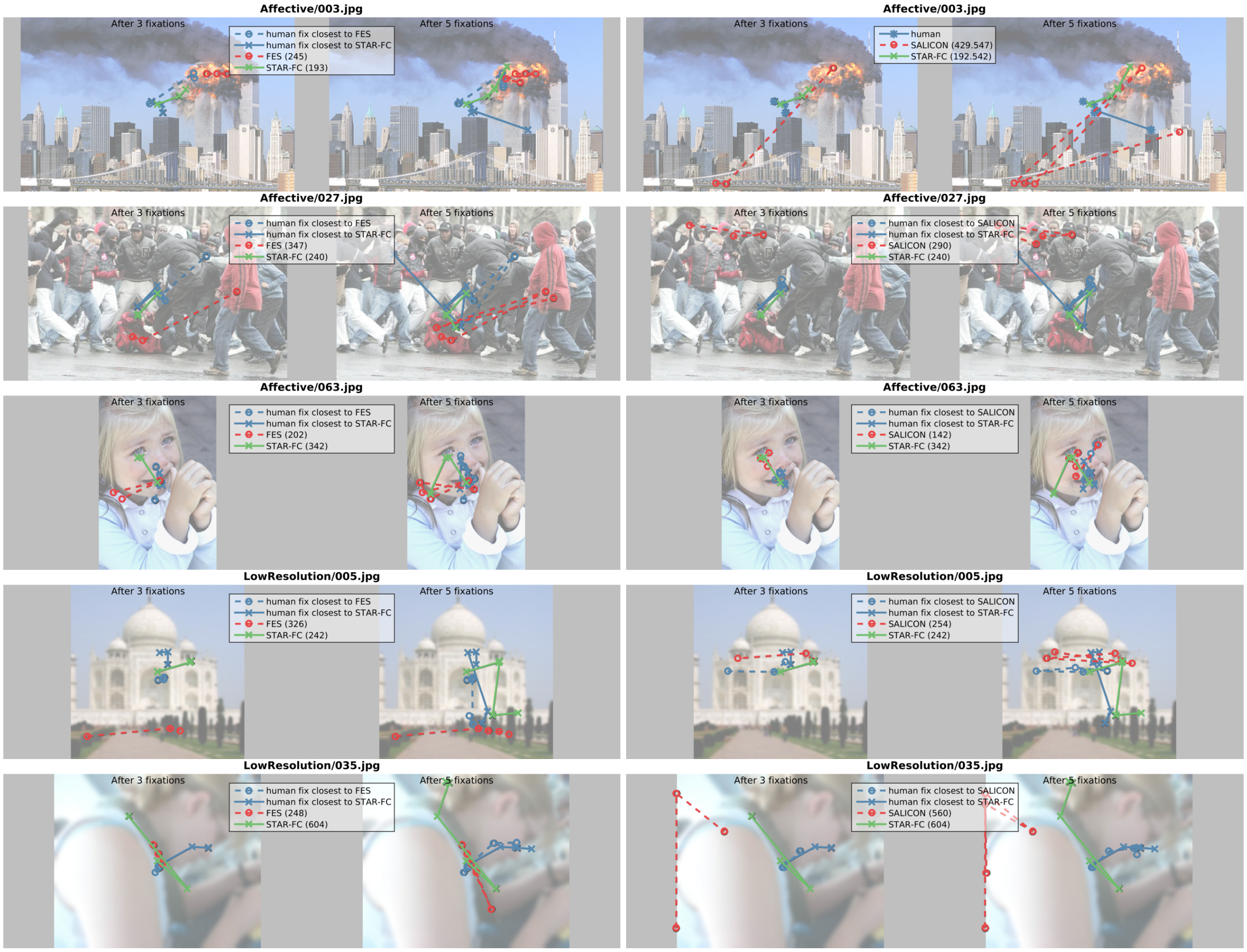


Figure 7: Examples of fixations predicted by FES (left column) and SALICON (right column) compared to the proposed STAR-FC model (with AIM 21infomax basis and MCA blending strategy). Blue lines represent the closest human fixations to each of the compared algorithms and numbers in parentheses indicate the corresponding Euclidean distance.

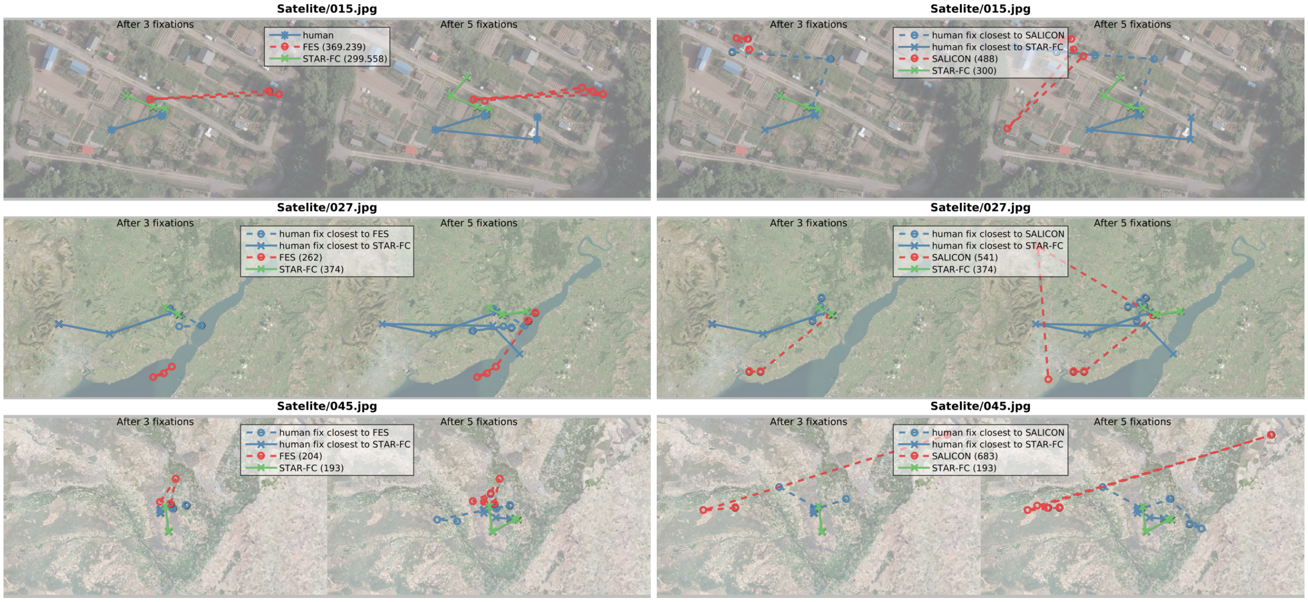


Figure 8: Examples of fixations predicted by FES (left column) and SALICON (right column) compared to the proposed STAR-FC model (with AIM 21infomax basis and MCA blending strategy). Blue lines represent the closest human fixations to each of the compared algorithms and numbers in parentheses indicate the corresponding Euclidean distance.

References

[1] A. Borji and L. Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581. 1

[2] W. S. Geisler and J. S. Perry. Real-time foveated multiresolution system for low-bandwidth video communication. In *Photonics West'98 Electronic Imaging*, pages 294–305. International Society for Optics and Photonics, 1998. 1

[3] J. K. Tsotsos, I. Kotseruba, and C. Wloka. A focus on selection for fixation. *Journal of Eye Movement Research*, 9:1–34, 2016. 1

[4] A. B. Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision*, 14(7):1–17, 2014. 1