Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes

Daniel Parks[a], Ali Borji[b], Laurent Itti[b,a,c]

[a]*Neuroscience Graduate Program, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089*
[b]*Department of Computer Science, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089*
[c]*Department of Psychology, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089*

Abstract

Previous studies have shown that gaze direction of actors in a scene influences eye movements of passive observers during free-viewing (Castelhano et al., 2007; Borji et al., 2014a). However, no computational model has been proposed to combine bottom-up saliency with actor's head pose and gaze direction for predicting where observers look. Here, we first learn probability maps that predict fixations leaving head regions (gaze following fixations), as well as fixations on head regions (head fixations), both dependent on the actor's head size and pose angle. We then learn a Bayesian combination of gaze following, head region, and bottom-up saliency maps with a Markov chain composed of head region and non-head region states. This simple structure allows us to inspect the model and make comments about the nature of eye movements originating from heads as opposed to other regions. Here, we assume perfect knowledge of actor head pose direction (from an oracle). The combined model explains observers' fixations significantly better than each of the constituent components. Finally, in a fully automatic combined model, we replace the oracle head pose direction data with detections from a computer vision model of head pose. Using these (imperfect) automated detections, we again find that the combined model significantly outperforms its individual components. Our work extends the engineering and scientific applications of saliency models and helps better understand mechanisms of visual attention in free-viewing and daily-life tasks.

Keywords: visual attention, eye movements, gaze direction, gaze following, head pose detection, bottom-up saliency, saliency modeling, fixation prediction, free viewing

1. Introduction

Where do people look during free viewing of natural scenes? A tremendous amount of research in the cognitive and computer vision communities has addressed this question. Two types of cues are believed to influence eye movements in this task: 1) low-level image features (a.k.a., bottom-up saliency) such as contrast, edge content, intensity bispectra, color, motion, symmetry, surprise, and 2) semantic high-level features (i.e., object and semantic level) such as faces and people (Cerf et al., 2009; Judd et al., 2009; Humphrey & Underwood, 2010), text (Wang & Pomplun, 2012), object center prior (Nuthmann & Henderson, 2010), image center prior (Tatler, 2007), semantic object distance (Hwang et

Email addresses: danielfp@usc.edu (Daniel Parks), borji@usc.edu (Ali Borji), itti@pollux.usc.edu (Laurent Itti)

al., 2011), scene global context (Torralba et al., 2006), emotions (Subramanian et al., 2014), memory (Droll et al., 2005; Carmi & Itti, 2006), gaze following (Castelhano et al., 2007; Borji et al., 2014a), culture (Chua et al., 2005), and survival-related features such as food, sex, danger, pleasure, and pain (Friston et al., 1994; Shen & Itti, 2012). Some of these features are well established while others need further investigation, such as semantic object distance and object center bias (Borji et al., 2013a; Einhäuser et al., 2008; Nuthmann & Henderson, 2010; Hwang et al., 2011). Also note that, while here we focus on free-viewing tasks, some of these factors also play a role in task-driven visual attention (Yarbus, 1967; Land & Lee, 1994; Ballard et al., 1995; Land & Hayhoe, 2001; Triesch et al., 2003; Einhäuser et al., 2008; Borji et al., 2014b; Borji & Itti, 2014).

Eye movements are proxies for overt visual attention which help us understand how humans allocate their focus to constrain the enormous quantity of observable visual data. Replicating such behavior is becoming more and more important recently due to the abundance of visual data, especially in computer vision and robotics. The overarching goal of machine attention would be to selectively process important and relevant information while filtering out redundant or irrelevant information. This capability has many applications such as image/video compression, object recognition, tracking, human-computer interaction (HCI), and photo collage (see Borji & Itti (2013) for a review).

1.1. Gaze following in natural scenes

Gaze direction plays an important role in joint/shared attention (Bock et al., 2008; Emery, 2000; Triesch et al., 2006), in learning in children and in the developmental phase (Hoffman et al., 2006; Okumura et al., 2013; Baldwin, 1995), and in daily social interactions (social learning, collaboration, coordination, and communication) in humans and some animals (e.g., Bock et al. (2008); Emery (2000); Shepherd et al. (2006); Kobayashi & Kohshima (1997)). Gaze following is also linked to the understanding and interpreting of the intentions of others (Premack & Woodruff, 1978; Baron-Cohen et al., 1995).

Gaze direction has been shown to influence visual attention and eye movements in free-viewing. Birmingham et al. (2009) showed that saliency does not account for fixations to eyes within social scenes. Instead, it appears that observers' fixations are driven largely by their default interest in social information. Fletcher-Watson et al. (2008) showed that observers are more likely to attend towards regions (and for longer times) that are being looked at by the central figure in the display. Castelhano et al. (2007) showed that observers follow the gaze directions of actors in the scene during the presentation of a slideshow. Borji et al. (2014a) studied the interaction of gaze direction and bottom-up saliency, and showed that gaze direction causally influences eye movements. They found that people follow the gaze direction even in the presence of other salient objects or people in a scene. They also predicted location of fixations that leave a person's head, using a simple cone centered about the gaze. They did not, however, take into account the effect of head pose and head fixations, propose a combined model with learned maps, or apply their model to all saccades in a scene, which are all considered here.

Effective gaze direction is a function of two factors: 1) head pose and 2) eye gaze direction. Thus, a full gaze direction model should be able to detect these two components. The first factor has been partially addressed by Zhu & Ramanan (2012), who proposed a model for head pose detection (only for the yaw direction) in natural scenes (see

also Murphy-Chutorian et al. (2007); Weidenbacher et al. (2006); Yücel et al. (2013); Valenti et al. (2012)). Here, we augment this model to also detect the head's pitch angle. The second factor has been relatively less explored (e.g., Wang et al. (2003)) due to complications such as individuality of eyes, occlusion, variability in scale, location, and lighting conditions. When a clear, high-resolution, live view of the eye is available, gaze estimation models have been used in applications such as evaluating user experience, monitoring and enhancing reading behavior, and adaptive displays (e.g., Wood & Bulling (2014)). However, this is not true for one-shot estimations of eye gaze in single images of persons in complex scenes.

In our prior paper (Borji et al., 2014a), we showed that human accuracy of gaze estimation did not fall dramatically when the eyes were removed from the face. This is not because the eye gaze is not important, but simply that the head pose is correlated with the final gaze angle, at least for the images in the Flickr dataset (see next section). Since we investigated a head pose detection algorithm, we optimized our model around head pose instead of final eye gaze. This seems reasonable for the head fixations, as they would seem to be more dependent on the locations of the facial features than the precise angle of the eyes. For gaze direction, we implicitly estimate the final eye gaze uncertainty given the head pose, to compensate for this. As a result of this decision, the difference in our results using ground truth and detections should reflect only the imperfections in the computer vision head pose detection system.

1.2. Learning-based fixation prediction models

In this paper, we attempt to model the influence of gaze following and fixations on human heads by learning spatial probability maps along with weights for these effects relative to bottom up saliency. We provide a short review of the many fixation prediction models that involve learning weights or filters. For a more detailed review, see Borji & Itti (2013).

Saliency models, influenced by the Feature Integration Theory (FIT) of Treisman & Gelade (1980), typically first extract a set of visual features such as contrast, edge content, intensity, and color for a given image (Koch & Ullman, 1985; Milanese et al., 1994; Itti et al., 1998). Second, they apply a spatial competition mechanism via a center-surround operation (e.g., using Difference of Gaussian filters) to quantify conspicuity in a particular feature dimension. Third, they linearly (often, with equal weights) integrate conspicuity maps to generate a scalar master saliency map (e.g., Treisman & Gelade (1980); Koch & Ullman (1985); Itti et al. (1998); Ehinger et al. (2009); Cerf et al. (2009); Borji & Itti (2012)), which is a map of the scalar quantity of saliency at every location in the visual field. Finally, a Winner-Take-All (WTA) neural network identifies the most salient region and inhibits this region (i.e., a decaying property) allowing other regions to become more salient in the next time step (i.e., to simulate shifting mechanism of attention). Since there are several design modeling parameters (e.g., the number and type of features, the shape and size of the filters, the choice of feature weights and normalization schemes, etc), some models have proposed various ways to learn these parameters. For example, Itti & Koch (2001) suggested to normalize the feature maps based on map distributions before linearly integrating them.

Instead of linear combination with equal weights, "max" (Li, 2002), or *maximum a posteriori* "MAP" (Vincent et al., 2009) type of integration, some models learn weights for different channels from a set of training data. For example,

Itti & Koch (2001) weighted different feature maps according to their differential level of activation within compared to outside manually-outlined objects of interest in a training set (e.g., traffic signs). Navalpakkam & Itti (2007) proposed an optimal gains theory that weights feature maps according to their target-to-distractor signal-to-noise ratio, and applied it to search for objects in real scenes. Judd et al. (2009) used low-level image features, a mid-level horizon detector, and two high-level object detectors (faces using Viola & Jones (2001) and humans using Felzenszwalb et al. (2008)) and learned a saliency model with liblinear SVM. Following Judd et al., Zhao & Koch (2011) learned feature weights using constrained linear regression and showed enhanced results on different datasets using different sets of weights. Later, Borji (2012) proposed an AdaBoost (Freund & Schapire, 1997) based model to approach feature selection, thresholding, weight assignment, and integration in a principled, nonlinear learning framework. The AdaBoost-based method combines a series of base classifiers to model the complex input data.

Some models directly learn a mapping from image patches to fixated locations. Following Reinagel & Zador (1999) who proposed that fixated patches have different statistics than random patches, Kienzle et al. (2007) learned a mapping from patch content to whether it should be fixated or not (i.e., +1 for fixated and -1 for not fixated). They learned a completely parameter-free model directly from raw data using a support vector machine with Gaussian radial basis functions (RBF).

Our prior work (Borji et al., 2014a) showed a causal influence of actor gaze onto observer saccades that leave the actor's head, by comparing observer saccade distributions between pairs of photographs that only differed by the actor's gaze direction. Human faces have already been shown to be important in predicting eye movements (Cerf et al., 2007), and low-level features have formed the basis of many saliency models. We investigate how combining these cues in real-world scenes involving human actors can produce state-of-the-art fixation prediction results.

1.3. Contributions

We introduce three main contributions. First, we learn a gaze following probability map using saccades that leave the head of an actor in a scene, along with a probability map of saccades that land on an actor's head (both computed relative to the actor's head pose angle), which combine to predict fixations better than our original head annotation and gaze cone model. Second, we learn a combined model of bottom-up saliency, head pose, and gaze following (first, assuming an oracle for actor head pose direction) which performs better than saliency and head detection maps, and better than the current state of the art without head pose or gaze information. Third, we construct a computer vision model to make the steps in the second contribution fully automatic. Even with the additional noise of the automated detections, our model still outperforms other models, and allows our model to be used out of the box. In section 4, we use ground truth head pose direction to build and test the first version of the model, and then in section 6 we replace the ground truth with the output of a computer vision model.

Figure 1: Sample images from our dataset along with annotated object boundaries and head pose annotations. The face regions (red), whole head (blue), eyes (yellow), head pose 2D angle (green), and final eye gaze 2D angle (blue) are labeled.

2. Flickr Gaze Dataset

We use data from our previous study (Borji et al., 2014a). Stimuli consisted of a set of 200 color photographs collected mostly from the Flickr web site[1]. Photographs span a variety of topics and locations such as indoor, outdoor, social interactions, object manipulations, instrument playing, athletics, reading, and dancing. We chose images in which at least one of the faces was large enough, visible, and was gazing at something visible in the scene (another person or an object). Images were shown to observers in two sessions with 100 images each. Observers had 5 minutes break in between two sessions. The eye tracker was re-calibrated before the second session. We manually annotated heads, faces, eyes, and head pose directions for all 200 images.

A total of 30 students from the University of Southern California (4 male, 26 female) were shown images for 10 seconds each with a 5 second gray screen in between two consecutive images. Stimuli were presented at 60Hz refresh and at a resolution of 1920×1080 pixels subtending $45.5^\circ \times 31^\circ$ of visual angle. Figure 1 shows sample images along with annotated objects and head pose directions. Please refer to Borji et al. (2014a) for more details on this dataset. This dataset was randomly divided into 10 folds (n=20 images) for cross validation, where each fold was considered the test set, and the remaining 9 groups were treated as the training set. While this dataset has faces in every image, they are situated in real, complex scenes, where less than 1 in 5 fixations go to the head regions. This places an upper bound on how much improvement head related information can provide in improving fixation prediction.

[1] http://www.flickr.com/. Some images were also borrowed from the AFW test dataset (Zhu & Ramanan, 2012).

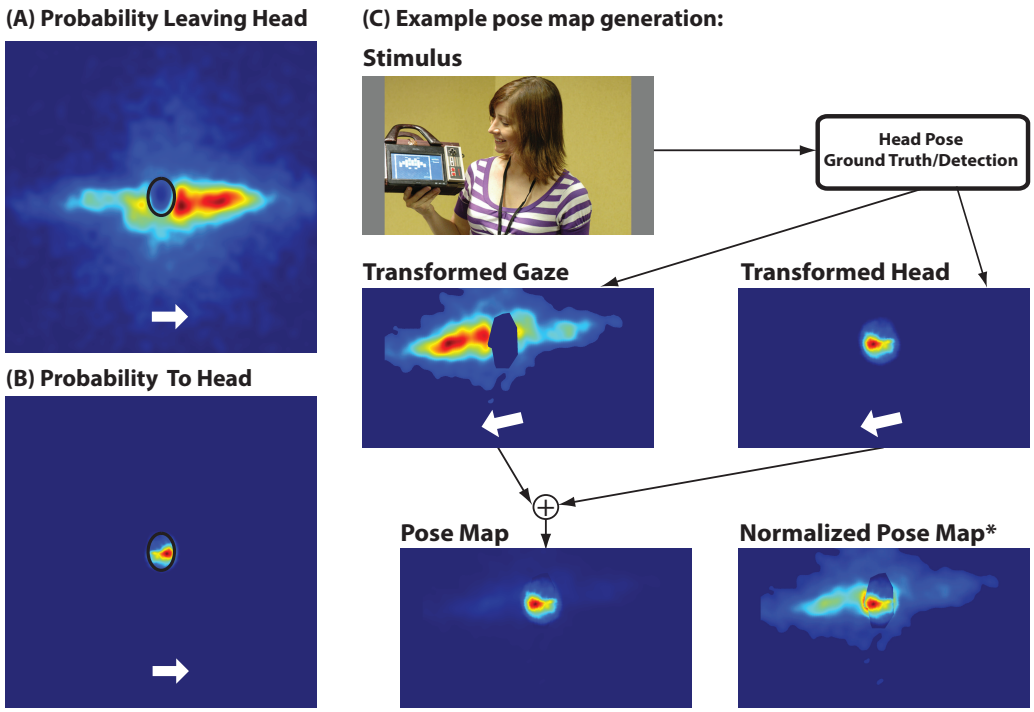

Figure 2: (A) Average fixation probability after leaving a normalized head over the whole dataset, where all head poses were rotated to point to the right (white arrow), and sized to a nominal head size (black oval). (B) Average fixation probability for all head fixations over the whole dataset, where all head poses were rotated to point to the right and size to the nominal head size. (C) These maps are then combined to form a pose map representing the fixation probability due to a particular head pose. The combined pose map is hard to see because of the relative energy in both maps. To better show the combined pose map, the head pose and gaze maps are individually normalized by their max value as shown in the normalized pose map. Normalize pose maps are shown in other figures for illustration purposes only, and are signified with a (*).

3. Head Pose Integrated Model

3.1. Learned Head Pose and Gaze Following Maps

In trying to improve upon our prior model of fixations leaving the head, we created a probability map of those fixations over the training set. First the 2D rotation angle between the current head pose angle and a reference head pose angle on the image plane was determined, and all saccade vectors leaving the head were rotated by that angle. The saccade vectors were then normalized by the size of the head. Figure 2A shows the probability map (i.e., average gaze following map) extracted from the training set. Note that since this is normalized by head pose and not final eye gaze, the uncertainty of final eye gaze given head pose is implicitly taken into account in this probability map.

We implicitly assumed in our prior work that all head fixations could be explained by the head detections alone, and then focused on fixations leaving the head by using a model of gaze. However, for this work, we decided to also learn fixation probabilities within the head, as the head pose angle would also presumably bias head fixations as well. Again, the saccade vectors were rotated based on the head pose angle, and then the length of the vectors was normalized by the head size. As shown in Figure 2B, the head fixations had a high density in the eye region, with low density in the hair and chin regions. For each head in the image, the gaze and head probability maps can linearly combine with weights to form what we refer to as the "pose map" as shown in Figure 2C.

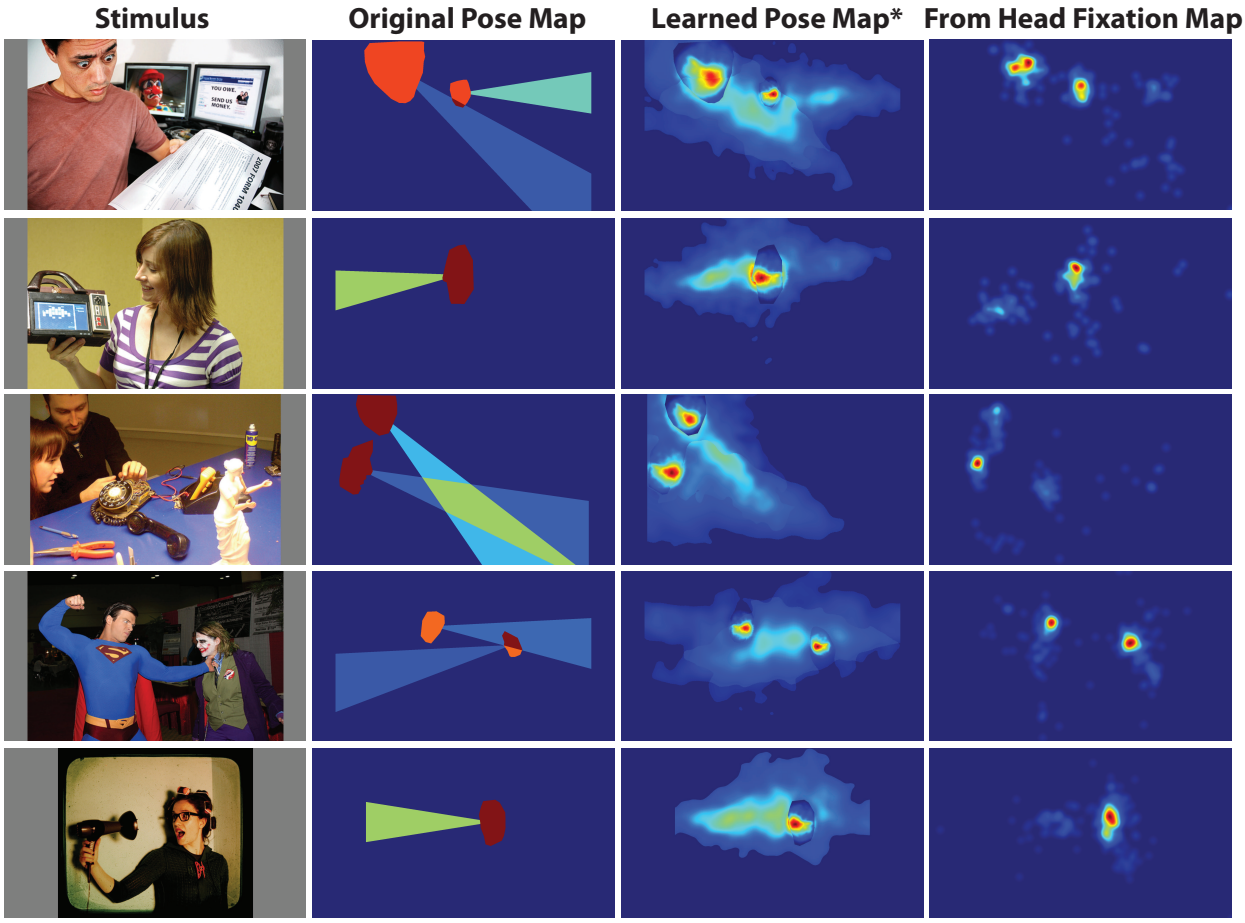

Figure 3: Sample images comparing the original pose, the learned pose model, and fixations coming from the head. *Note that the learned pose map is normalized so that both the head pose and the gaze following components can be seen clearly for illustration purposes.

3.2. Integration with Saliency

The model uses three components: *the gaze following probability map, the head probability map, and the saliency map*. The saliency map was generated using the adaptive whitening saliency (AWS) algorithm (Garcia-Diaz et al., 2012). AWS was chosen as the saliency benchmark due to its performance in a recent review of saliency models (Borji et al., 2013b), and because it doesn't try to model higher level concepts, such as human faces or objects, and so is suitable in our view for serving as a proxy of low-level attention. The gaze following probability and head probability maps both used annotated head regions and head pose angles to place, scale, and rotate the learned probability maps from Figure 2A and Figure 2B, respectively, which are then combined to form pose maps. Figure 6 shows what these combined pose maps look like for some sample images. Since both human faces and gaze direction have been shown to influence eye movements, we need a means to integrate these higher-level cues with saliency.

Ever since Yarbus (1967), it has been argued that the scan path of saccades and not just the collection of fixations is important to understanding eye movements, however, most saliency models have focused on fixed maps or an inhibition scheme to predict a series of saccades. It was already known from our previous work that head fixations were

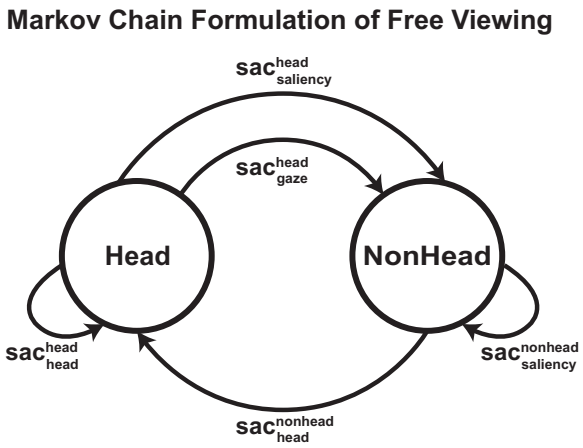

Figure 4: Discrete-time Markov chain formulation of saccadic eye movement of an observer performing a free viewing task.

frequently followed by other head fixations, and non-head fixations were frequently followed by non-head fixations, suggesting to us that this affinity might reflect a change in implicit task on the part of the observer. As a result, we propose that a gaze contingent model, where the last fixation location is provided, is more suitable in the quest to achieve parity with the predictive power of other human observer's fixations (i.e. an inter-observer model). This inherent temporal structure can be captured by a discrete time Markov chain formulation (Norris, 1998) with two states,*head* (H) and *nonhead* (N). Unlike an implicit hidden Markov model (Rabiner & Juang, 1986), the states in this model are known and the meaning of the weights is easily understood. In addition to offering predictive power, this simple structure allows us to obtain an understanding of how top-down and bottom-up cues are combined.

For the rest of the paper, we will use the following notation for saccades: $sac^{origin}_{destination}$, and we will call saccades originating from the head, sac^{head}, and from other regions, $sac^{nonhead}$. From these two regions, there is a certain probability that subjects will saccade to a head $p(sac_{head})$, to a point gazed at by an actor in the scene (i.e., following the actor's gaze) $p(sac_{gaze})$, or to a salient point $p(sac_{salient})$. These probabilities express how likely an observer is to follow each particular cue. The possible transitions are shown in Figure 4.

3.3. Learned Transition Probabilities

To integrate head, gaze, and saliency information, we estimated the probability with which a subject transitions between regions of the image, namely head regions and other regions. For tractability of calculation, we treat the possible end point maps as non-overlapping and exhaustive sets: $p(sac_{all}) = p(sac_{head}) + p(sac_{gaze}) + p(sac_{salient}) = 1$. Of course, these are not completely disparate sets with no overlap, nor are they likely to be exhaustive, which is a source of error in our model. We could model the conjunctions of each of these components as well, but it would be hard to identify what proportion of each saccade is due to each component. These transition probabilities were learned in the training set separately for transitions from the head $p(sac^{head}_{all})$ and not the head $p(sac^{nonhead}_{all})$. This was to distinguish fixations that are presumably more gaze focused from the rest of the fixations. The transition probabilities were considered to be weights on the corresponding maps for each component. The transition probabilities are shown

in Figure 5. Saccades going to any head was scored as a head transition. For saccades leaving a head, the value of the gaze following map at the saccade destination was compared to the value of the saliency map at that point. If the location in the gaze map was higher, then it was considered a gaze transition. In all other cases, it was scored as a saliency transition.

We learned the transition probabilities for two systems, *with* and *without* the sac_{gaze}^{head} component: *the 3 component* and *the 2 component model*, respectively. As seen in Figure 5, gaze and head information are stronger during from head saccades, while saliency dominated in the non-head case. This is intuitive, as head and gaze seem more relevant when already looking at a head. It is also useful to note that gaze following is much less of a factor than head or saliency information, and is more of a second-order effect. Thus, only a small gain in fixation prediction accuracy is likely to result from taking gaze into consideration in the final model; however, with the maturity of the field of eye movement prediction, it is likely that more second-order effects such as this one will need to be addressed to further improve upon the already very good fixation prediction abilities of state-of-the-art models

The gaze following, head pose, and AWS maps were all made into probability maps (i.e., sum to 1). The transition probabilities for the 3 component model were used to weigh the respective maps, which were then summed to create the final maps shown in the second column from the right in Figure 6 for what we call the Pose+AWS model:

$$map_{pose+AWS} = p(sac_{head})map_{head\_pose} + p(sac_{gaze})map_{gaze} + p(sac_{salient})map_{salient} \quad (1)$$

For a baseline system that takes head detections and bottom-up saliency into account without head pose or gaze information, we constructed the Head+AWS model, which uses the 2 component model to weigh the head detections along with the bottom up saliency and multiplied by these learned transitions, to create a final combined probability map:

$$map_{head+AWS} = p(sac_{head})map_{head} + p(sac_{salient})map_{salient} \quad (2)$$

For both of these models, the weights change depending on the origin of the saccade (sac^{head} or $sac^{nonhead}$).

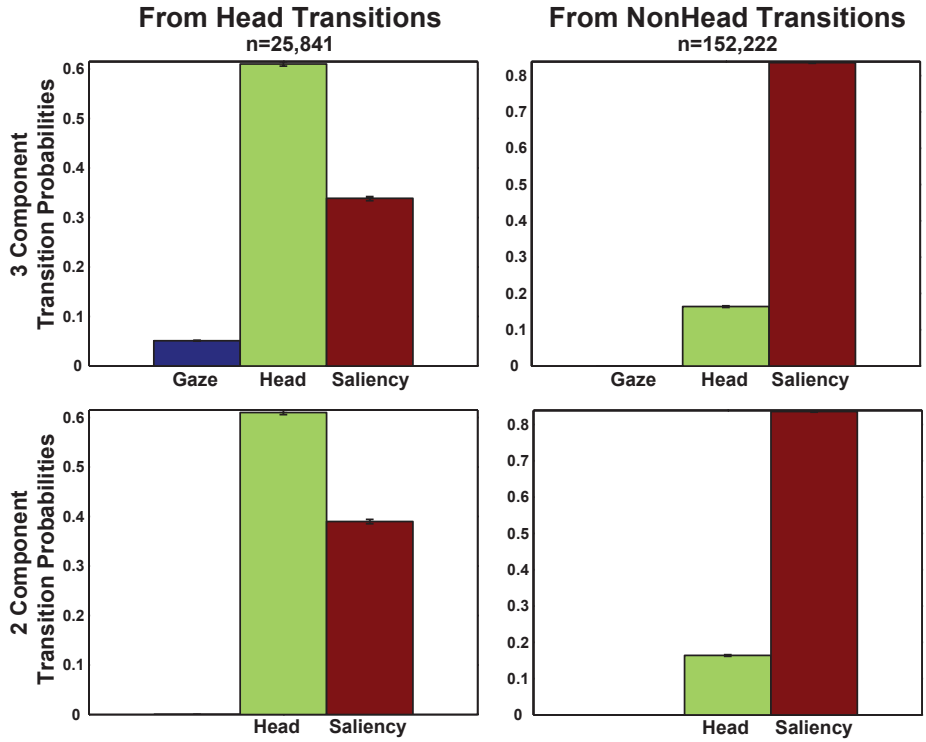

Figure 5: Transition probabilities when saccading to the head regions, gaze regions, or salient regions when starting from a head region or a non-head region. Note that saliency is much stronger when originating from a non-head region. 5th and 95th percentile confidence intervals are shown across the cross validation folds.

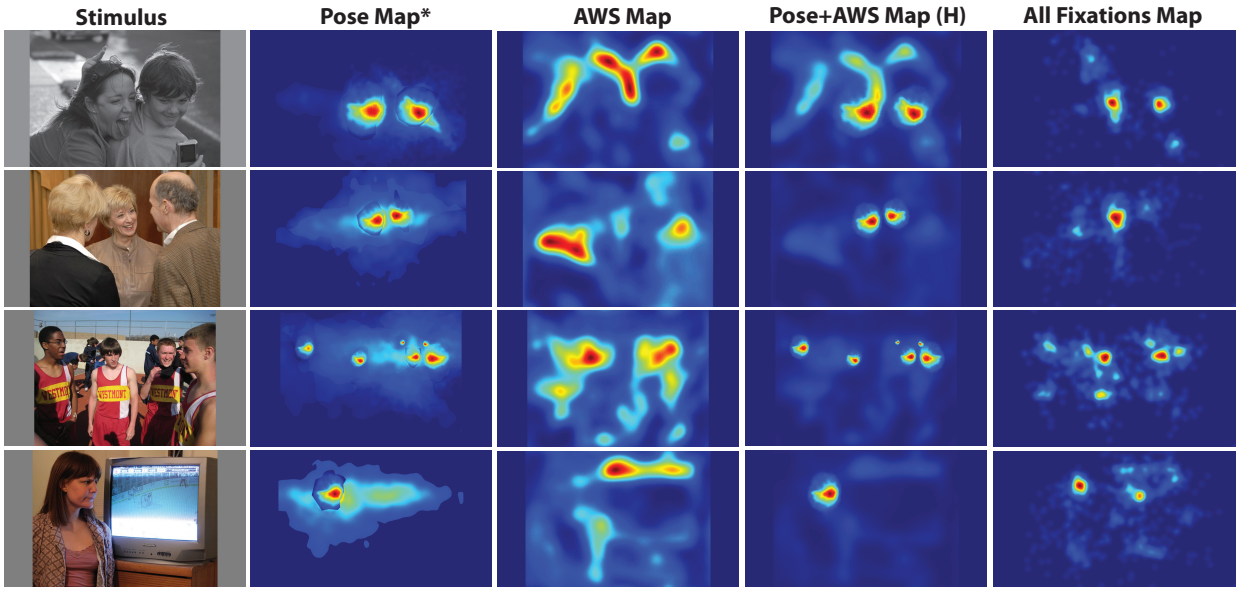

Figure 6: Sample images with corresponding maps for Pose, AWS, and the final Pose+AWS model, along with maps of all fixations. The combined Pose+AWS map is shown with the weights from the Heads (H) state. *Note that the maps in the Pose column are normalized for illustration purposes, but the final Pose+AWS column uses the original Pose map.

4. Annotation Based Fixation Prediction Results

Figure 7A shows the AUC performance for heads, pose, AWS, head+AWS, and pose+AWS as well as an inter-observer model for from head fixations. 5th and 95th percentile confidence intervals over the cross folds are shown for all AUC data. The pose model includes the combined learned maps of head pose and gaze following, while the head is just the actual head detections. The head+AWS model can be viewed as an updated version of Cerf's saliency model with heads Cerf et al. (2007), and provides our baseline of a state of the art system that includes head detection, but does not include head pose or gaze following information.

The Wilcoxon signed rank test was used in all of the following comparisons, which have the same p-value because the signed rank significance value is determined by the number of comparisons and the number of times that each comparison has a certain rank, and for all model comparisons here the winning model won on every fold. For from head fixations: Head<Pose($p<0.002$), AWS<Pose ($p<0.002$), Pose<Head+AWS ($p<0.002$), AWS<Head+AWS ($p<0.002$), Head+AWS<Pose+AWS ($p<0.002$), Pose+AWS<Inter-observer[2] ($p<0.002$). Note that the pose model by itself outperforms bottom up saliency when looking at from head fixations.

There were 178,063 total fixations in the data, with 25,841 or 14.5% of the fixations originating from a head. As a result, the effect of the improvement due to the from head fixations is muted in the overall data. When looking at all fixations, as in Figure 7B, the difference is much smaller, and the AWS saliency model does markedly better by itself. It is also interesting to note that the inter-observer model is more predictive for the saccades originating from the head, implying that they are more stereotyped than the remaining saccades. For all fixations, the pose+AWS model outperforms the head+AWS model by a small but significant ($p<0.002$) amount. The different models are significantly different: Head<Pose ($p<0.002$), Pose<AWS ($p<0.002$), AWS<Head+AWS ($p<0.002$), Pose+AWS<Head+AWS ($p<0.002$), and Pose+AWS<Inter-observer ($p<0.002$).

As we can see from these results, pose works better for from head fixations, while saliency works better for other fixations. This validates the different weights associated with these maps when starting from the head versus other regions. Heads by themselves also perform better when starting from the head, but do not account for all of the performance improvement.

Figures 7C and 7D show the per image AUC performance of the Head+AWS and Pose+AWS models for from head fixations and all fixations, respectively. Looking at the data this way, the pose information improves fixation prediction when fixations are from the head, and this is enough to improve the performance of all fixations (Head+AWS<Pose+AWS ($p<2e\text{-}31$) for fixations from the head, and Head+AWS<Pose+AWS ($p<4e\text{-}32$) for all fixations).

To determine where our combined model and bottom-up saliency diverge, we also plot the performance of AWS vs Pose+AWS in Figure 8 on a per image basis. Image 1 is one where the saliency map does not pick up saccades in the direction of gaze. Image 2 shows where the pose model is aided by the gaze following from the human to the

[2] A map built from fixations of other observers on the same image seen by an observer.

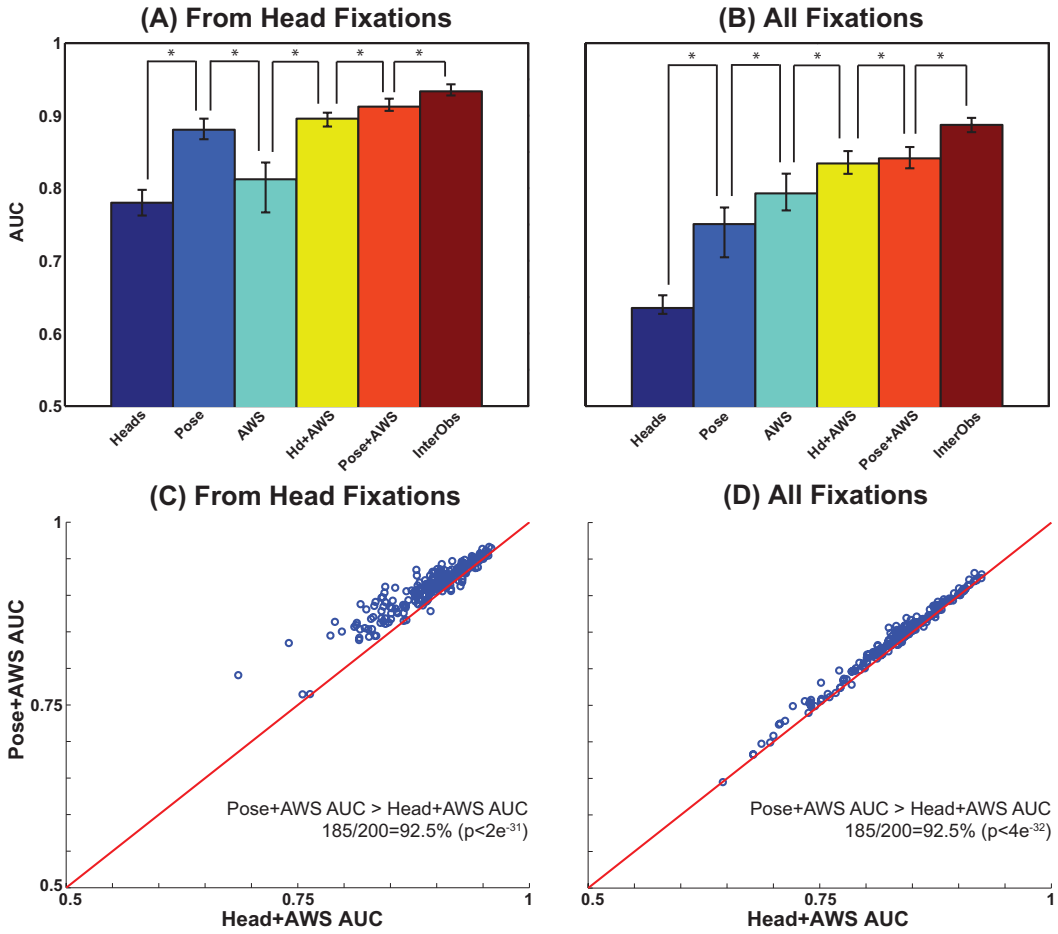

Figure 7: Performance of each component using ground truth. AUC performance on saccades originating (A) from the head and (B) all saccades. An inter-observer model is shown to provide a ceiling for performance. 5th and 95th percentile confidence intervals are shown. There is also a per image AUC comparison between the Head+AWS model and the Pose+AWS model for (C) from the head and (D) all saccades.

dog, although one could argue that we could have labeled non-human faces as well to account for this. Image 3 shows where the pose information greatly outperforms low level saliency because the heads are not found to be salient. In image 4, both AWS and Pose+AWS are not performing well, and it is interesting to note that both models miss another high level cue, text on the stomachs of the women. Image 5 is one where both models are performing well, as the heads are salient already given low level statistical information. Image 6 is one where the AWS model picks up the major attended areas, but also many more unattended regions. Overall, the combined model provides a small but consistent improvement over all fixations.

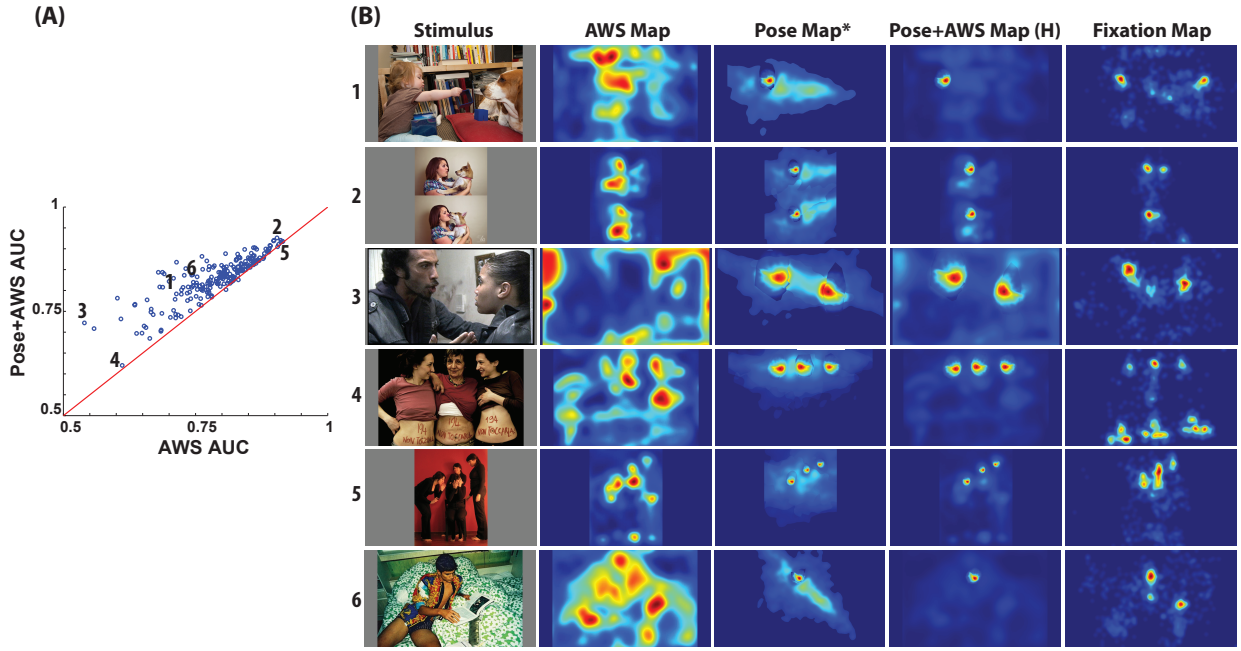

Figure 8: (A) Scatter plot showing per image AUC performance of AWS saliency versus the Pose+AWS model. (B) Four sample points on the scatter plot with images and their corresponding maps are shown for illustration. The combined Pose+AWS map is shown using the weights from the Heads (H) state. *Note that the maps in the Pose column maps are normalized for illustration purposes, but the final Pose+AWS column is shown as used by the model.

5. Head Pose Detection Model

We established that given the ground truth head pose, head pose and gaze following information can improve state of the art saliency algorithms. We also found that saccades originating from the head were better predicted by this pose information than saccades originating elsewhere, which were better predicted by low-level saliency. However, in order to improve applicability for the model, it is useful to remove the need for manual annotations, which are expensive and time consuming. As before with the annotations, we need the detection model to generate head pose polygons and image plane head pose angles for each head to generate the maps for an image.

When determining the 3D angle to which a head is oriented, we defined the pose angles egocentrically, as shown in Figure 9A. The yaw and pitch of the head are important for determining the 2D image angle. The roll of the head simply rotates the head about the 2D image angle formed by the yaw and the pitch. Although large roll angles will drastically change the view of the head pose and can make it difficult to accurately extract the correct yaw and pitch angles, this effect is ignored. The Zhu & Ramanan model (2012) that we used to detect heads and extract head pose only provided a yaw estimate. Figure 9B shows the basic structure of their model. Local parts of the face are detected using mixture of trees part detectors (similar to the detectors used in Felzenszwalb et al. (2008), and the relative locations of the detected parts create deformations in the larger pose model at a certain cost.

To add a pitch estimate, a set of random binary ferns (Ozuysal et al., 2010) were learned over the part detections. Each fern randomly samples a fixed set (n=3) of binary comparisons. Randomly selected head model parts were compared in either their X or Y relative positions as shown in Figure 9C with a binary 1 if (green)>(red) and 0

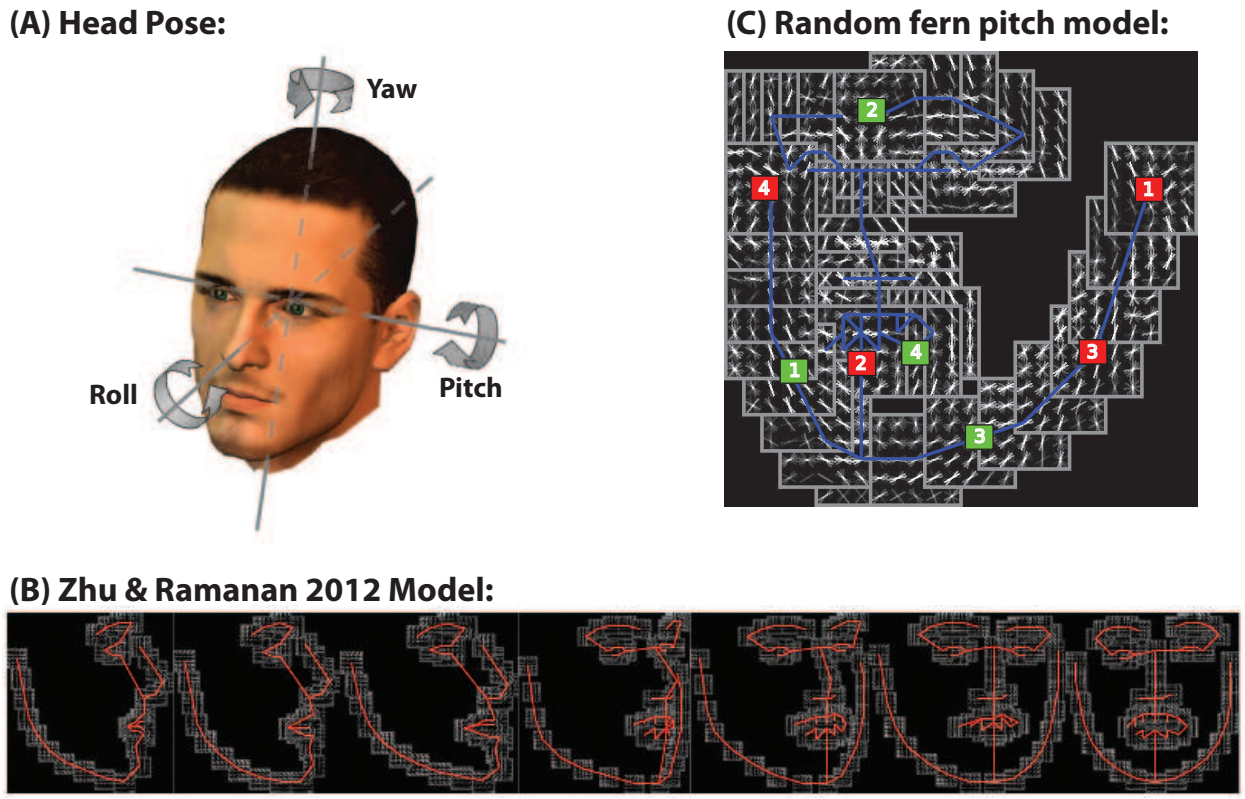

Figure 9: (A) [Modified from Murphy-Chutorian & Trivedi (2009)] 3D head pose angle of a head can be defined egocentrically using $roll$, $pitch$, and yaw. (B) The head pose model Zhu & Ramanan (2012) has 146 shared local parts across 13 learned poses, with learned deformation costs (in red) between neighboring parts. The model only defines head pose in the yaw angle. (C) Sample fern comparison for each bit. The relative X or Y position of the green part, g, is compared with the corresponding position of the red part, r, with $g > r = 1$ and 0 otherwise.

otherwise. This is done for a fixed set of comparisons, n, which is the size of the fern in bits. These fern bits form a value f_{val}, (e.g., 10 = 3 for a 2-bit fern).

Over the training images, a histogram is created for each possible value of the fern, f_{val} (a 2 bit fern has $2^2 = 4$ possible values). Each histogram has bins for each possible ground truth pitch angle category ang_{cat}. Each time a model evaluates to a particular fern value f_{val}, the histogram for the fern value, f_{val} is selected, and the ground truth angle bin ang_{cat} is incremented. During testing, the histogram for that fern value f_{val} is extracted, and these proportions are summed across all ferns (n=500). The angle bin with the maximum value is then used to predict the pitch angle. The random ferns were trained using unused images in Zhu & Ramanan's AFW dataset ((Zhu & Ramanan, 2012) (n=185 images)) with 3 ang_{cat} categories (22.5°, 0°, & -22.5°).

The combined yaw and pitch angle estimates provided the 3D head pose angle. These 3D head pose angles were first converted to 2D image head pose angles using a simple orthographic projection. Alternatively, the camera parameters can be learned, and a perspective projection can be used to gain a more accurate angle estimate. The X and Y extrema of the center points of all of the detected local parts provided the bounding polygon for the head as well. With the 2D head pose angle and head polygon, the system was then run in the same manner as using ground truth.

Sample detections of the head pose estimation system are shown in Figure 10. The system generates a yaw and

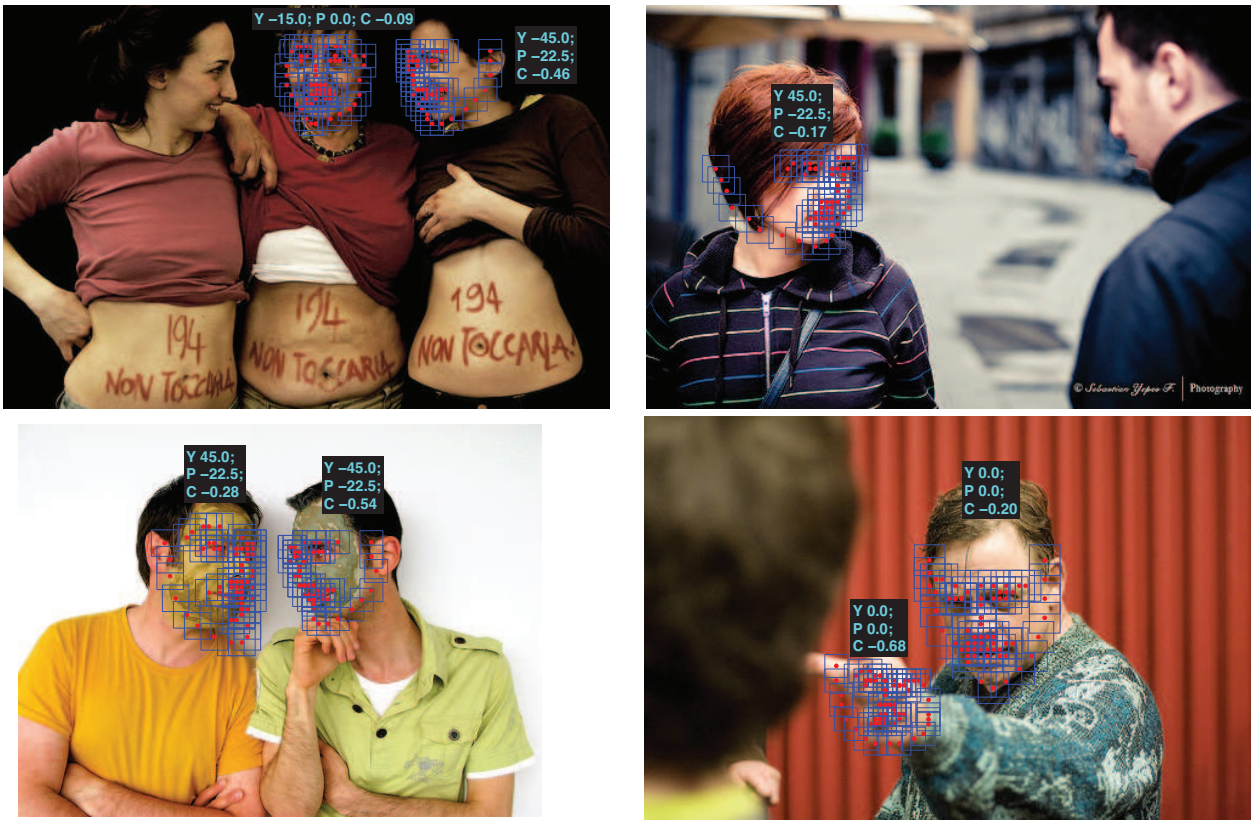

Figure 10: Sample head pose model detections. Y is yaw angle with 0°pointing out of the image and positive moving to the left from the head's perspective. P is the pitch angle with 0°being level and positive looking up. Both are in degrees. The confidence of the detection is shown with C, and is an unbounded number, with higher values being more confident. The detector was thresholded at -0.75.

Table 1: Head Pose Detection: Component Performance

	Mean F_1	**Std F_1**	**Chance**
Face Detection	0.75	0.05	-
Yaw Detection	0.82	0.07	0.33
Pitch Detection	0.56	0.07	0.33

pitch angle in addition to a confidence score. The performance of the system over the 10 folds of the Flickr set is shown in Figure 1. F_1 is simply the harmonic mean of precision P and recall R: $F_1 = 2PR/(P+R)$, and is bounded by 0 and 1, inclusively[3]. With this metric, precision and recall are equally weighted in importance, and a value of 1 indicates perfect recall and precision. Note that the head detection F_1 performance has no real chance lower bound, but that the angle detections were only scored on correct head detections with coarse bins of left, straight, and right (60°, 0°,& -60°) for the yaw angle, and up, level, and down (22.5°, 0°, & -22.5 °) for the pitch angle. Therefore, chance for both angles is 33.3%.

[3]Precision is defined as $P = TP/(TP+FP)$, and recall is defined as $R = TP/(TP+FN)$, where *TP* is True Positive, *FP* is False Positive, and *FN* is False Negative

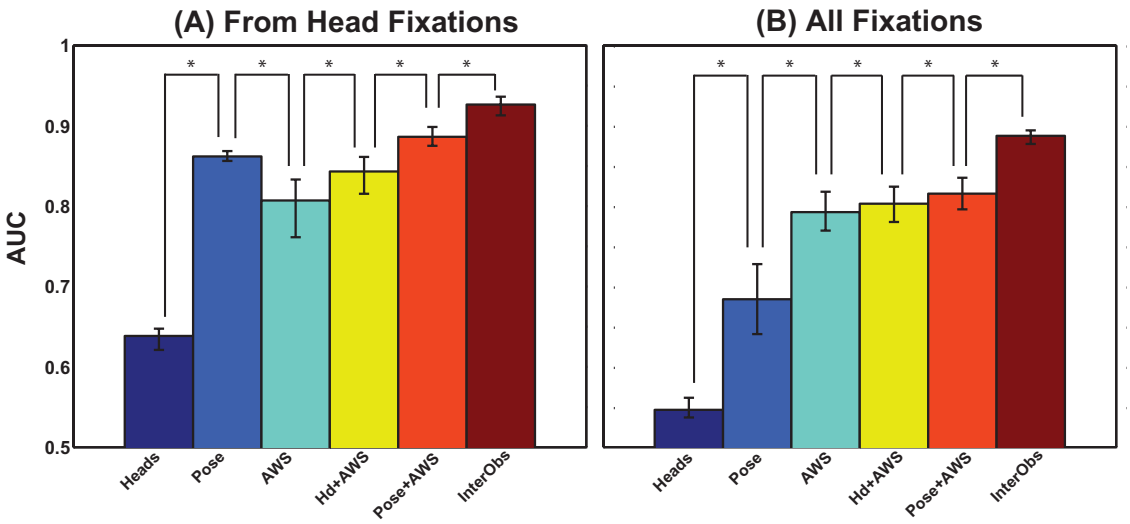

Figure 11: Performance of each component using head pose detections. AUC performance on saccades originating (A) from the head and (B) all saccades. An inter-observer model is shown to provide a ceiling for performance. 5th and 95th percentile confidence intervals are shown.

6. Detection Based Fixation Prediction Results

Using the same cross fold sets as with the annotations, the head pose model was run over the images in the test fold, and this provides head polygons and 2D pose angles, which are used in the same manner as the ground truth head pose: head, head pose, and gaze following per head maps were rotated by the pose angle, scaled relative to the actual head detection (based on the mean scale over image dimensions), and applied to the appropriate map.

Figure 11A shows the AUC performance when using head pose detections from the model instead of from the ground truth annotations when predicting only fixations from the head. Again, 5th and 95th percentile confidence intervals are shown for all AUC data. For from head fixations, significance was determined using the Wilcoxon signed rank test: Head<Pose($p<0.002$), AWS<Pose ($p<0.002$), Pose<Head+AWS ($p<0.002$), AWS<Head+AWS ($p<0.002$), Head+AWS<Pose+AWS ($p<0.002$), Pose+AWS<Inter-observer ($p<0.002$).

When looking at all fixations, as in Figure 11B, the difference is again smaller, and the AWS saliency model alone accounts for most of the performance. Still, however, the differences were significant. For all fixations, the pose+AWS model outperforms the head+AWS model by a small but significant ($p<0.002$) amount. The different models are significantly different: Head<Pose ($p<0.002$), Pose<AWS ($p<0.002$), AWS<Head+AWS ($p<0.002$), Pose+AWS<Head+AWS ($p<0.002$), and Pose+AWS<Inter-observer ($p<0.002$).

7. Discussion

We showed that head pose and gaze following information can contribute to fixation prediction, especially for saccades originating from head regions, where it outperforms purely bottom up saliency. Head pose detections were shown to be a fairly good proxy of the ground truth, as well. The learned gaze map performed better than the cone map used in previous work, but it is probably limited to images and video, since an environment where the observer

and the actor are both present would have no image plane limitation (e.g. the observer could see an actor looking off in the distance, and could turn around to determine if the gazed at entity is behind them).

While most saliency models use a static map, our results validate our approach where a different combination of cues is used with different weights depending on the origin of a saccade. We believe that this kind of gaze-contingent modeling is a promising direction to further bridge the now relatively small gap that remains between saliency models and inter-observer predictions. There are, however, opportunities for further improvement. Integrating final eye gaze direction should help, as the head pose does not always match the final eye gaze direction. This is actually implicitly learned in the gaze following map, since we only aligned the data for head pose and not for final eye gaze. However, the averaging that takes place is likely to reduce performance. Also, note that automatic reliable eye direction detection models do not yet exist in computer vision especially over complex natural scenes, due to high variability of eyes in scenes. However, our model can be easily extended by adding the final gaze direction results of more accurate models in the future. Text detection in the wild (e.g., (Meng & Song, 2012)), could also improve model performance (Cerf et al., 2009).

Another area of potential improvement is the lack of 3D scene information. Being able to extract a depth map for the heads in the image along with the 3D angle and camera information would allow a richer estimation of which head poses are likely to be attending to objects or locations in the field of view versus outside of it. To be really useful, of course, the depth map should be extracted automatically. Another area of improvement is better object segmentation, since gaze estimation has been shown to be biased towards the presence of objects (Lobmaier et al., 2006), and the areas of the gaze probability map could be modulated by the presence of segmented objects in the scene to further improve the accuracy of the prediction. Saliency is not a good proxy for this, since it would be unable to extract gaze following saccades to non salient objects.

Improving the accuracy of gaze direction prediction and saliency models can be useful in several engineering applications, for example in computer vision (e.g., action recognition, scene understanding in videos (Marin-Jimenez et al., 2014), reading intentions of people in scenes (Yun et al., 2013), attentive user interfaces), human-computer and human-robot interaction (e.g., Hoffman et al. (2006); Lungarella et al. (2003); Nagai et al. (2002); Breazeal & Scassellati (2002); Bakeman & Adamson (1984)), determining the attention levels of a driver (e.g., Murphy-Chutorian et al. (2007)), and enriching e-learning systems (e.g., Asteriadis et al. (2013)). Also such models can be useful for scientific research to study psychological disorders and diagnose patients with mental illness (e.g., anxiety and depression (Compton, 2003; Horley et al., 2004; Kupfer & Foster, 1972), schizophrenia (Franck et al., 2002; Langton, 2000), autism (Klin et al., 2009; Fletcher-Watson et al., 2009)). Our proposed model can be used for constructing signatures for distinguishing patients from healthy normal people and conducting neuropsychological and clinical research (similar to Foulsham et al. (2009); Tseng et al. (2012)). Similarly, eye movements help scene understanding tasks (e.g., Mathe & Sminchisescu (2012)). Eye movements can also be used to help build assistive technology for the purpose of autism diagnosis and monitoring of social development, as was done by Ye et al. (2012) where they were used detect eye contact preference. As another example, Alghowinem et al. (2013) demonstrated that eye movement can be used as a means of depression detection.

8. Conclusions

We proposed a combined head pose estimation and low level saliency model that outperforms other models that do not take head pose and gaze following information into account. This information has already been shown to causally predict eye fixations, and is a well known element of human social understanding. Automatic head pose estimation from a single image was incorporated in the model, and allows the system to be run directly.

Our model formulates human saccade movement as a two state Markov chain that can be viewed as a dichotomy between a top-down driven (head) state and a bottom-up drive saliency (non-head) state. It extracts transition probabilities between these two states automatically, and the learned weights show a preference for heads and gaze related fixations when originating on a head, while being more saliency driven when originating elsewhere. This is intuitive, and we see this as a step towards a more dynamic understanding of eye movement behavior of subjects when free viewing natural scenes beyond fixed maps of fixation predictions.

Acknowledgments

This work was supported by the National Science Foundation (grant number CMMI-1235539), the Army Research Office (W911NF-11-1-0046 and W911NF-12-1-0433), and U.S. Army (W81XWH-10-2-0076). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof. Thanks to Boris Schauerte for his help collecting the stimulus set for our second experiment. We wish to thank reviewers for valuable comments.

Commercial relationships: none.

Corresponding author: Daniel Parks.
Email: danielfp@usc.edu.
Address: University of Southern California, Hedco Neuroscience Building, 3641 Watt Way, Los Angeles, CA 90089-2520, USA

References

Alghowinem, S., Goecke, R., Wagner, M., Parkerx, G., & Breakspear, M. (2013). Head pose and movement analysis as an indicator of depression. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on* (pp. 283–288). IEEE.

Asteriadis, S., Karpouzis, K., & Kollias, S. (2013). Visual focus of attention in non-calibrated environments using gaze estimation. *International Journal of Computer Vision*, (pp. 1–24).

Bakeman, R., & Adamson, L. B. (1984). Coordinating attention to people and objects in mother–infant and peer–infant interaction. *Child development*, .

Baldwin, D. A. (1995). Understanding the link between joint attention and language., .

Ballard, D., Hayhoe, M., & Pelz, J. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience.*, *7*, 66–80.

Baron-Cohen, S., Campbell, R., Karmiloff-Smith, A., Grant, J., & Walker, J. (1995). Are children with autism blind to the mentalistic significance of the eyes? *British Journal of Developmental Psychology*, *13*, 379398.

Birmingham, E., Bischof, W. F., & Kingstone, A. (2009). Saliency does not account for fixations to eyes within social scenes. *Vision research*, *49*, 2992–3000.

Bock, S., Dicke, P., & Thier, P. (2008). How precise is gaze following in humans? *Vision Research.*, *48*, 946–957.

Borji, A. (2012). Boosting bottom-up and top-down visual features for saliency estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 438–445). IEEE.

Borji, A., & Itti, L. (2012). Exploiting local and global patch rarities for saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 478–485).

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, *35*, 185–207.

Borji, A., & Itti, L. (2014). Defending yarbus: Eye movements predict observers' task. *Journal of vision*, .

Borji, A., Parks, D., & Itti, L. (2014a). Complementary effects of gaze direction and early saliency in guiding fixations during free-viewing. *Journal of Vision*, *xx*, xx.

Borji, A., Sihite, D., & Itti, L. (2014b). What/where to look next? modeling top-down visual attention in complex interactive environments. *IEEE Trans. on Systems, Man, and Cybernetics, PART A-SYSTEMS AND HUMANS*, .

Borji, A., Sihite, D. N., & Itti, L. (2013a). Objects do not predict fixations better than early saliency: A re-analysis of einhäuser et al.'s data. *Journal of vision*, *13*, 18.

Borji, A., Tavakoli, H. R., Sihite, D. N., & Itti, L. (2013b). Analysis of scores, datasets, and models in visual saliency prediction. In *Computer Vision (ICCV), 2013 IEEE International Conference on* (pp. 921–928). IEEE.

Breazeal, C., & Scassellati, B. (2002). Challenges in building robots that imitate people. *Imitation in animals and artifacts*, (p. 363).

Carmi, R., & Itti, L. (2006). The role of memory in guiding attention during natural vision. *Journal of Vision*, *6*, 4.

Castelhano, M. S., Wieth, M., & Henderson, J. M. (2007). I see what you see: Eye movements in real-world scenes are affected by perceived direction of gaze. In *Attention in cognitive systems. Theories and systems from an interdisciplinary viewpoint* (pp. 251–262). Springer.

Cerf, M., Frady, E. P., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*.

Cerf, M., Harel, J., Einhäuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection.

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 12629–12633.

Compton, R. J. (2003). The interface between emotion and attention: A review of evidence from psychology and neuroscience. *Behavioral and cognitive neuroscience reviews*, *2*, 115–129.

Droll, J. A., Hayhoe, M. M., Triesch, J., & Sullivan, B. T. (2005). Task Demands Control Acquisition and Storage of Visual Information. *Journal of Experimental Psychology Human Perception and Performance*, *31*, 1416–1438.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, *17*, 945–978.

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*, 2.

Einhűuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency, . *8*, 18, 1–26.

Emery, N. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience & Biobehavioral Reviews*, *24*, 581–604.

Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A discriminatively trained, multiscale, deformable part model. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (pp. 1–8). IEEE.

Fletcher-Watson, S., Findlay, J. M., Leekam, S. R., & Benson, V. (2008). Rapid detection of person information in a naturalistic scene. *Perception*, *37*, 571.

Fletcher-Watson, S., Leekam, S., Benson, V., Frank, M., & Findlay, J. (2009). Eye-movements reveal attention to social information in autism spectrum disorder. *Neuropsychologia*, *47*, 248–257.

Foulsham, T., J.J., B., Kingstone, A., Dewhurst, R., & G., U. (2009). Fixation and saliency during search of natural scenes: the case of visual agnosia. *Neuropsychologia*, *47*, 1994–2003.

Franck, N., Montoute, T., Labruyère, N., Tiberghien, G., Marie-Cardine, M., Daléry, J., d'Amato, T., & Georgieff, N. (2002). Gaze direction determination in schizophrenia. *Schizophrenia research*, *56*, 225–234.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*, 119–139.

Friston, K., Tononi, G., Reeke Jr, G., Sporns, O., & Edelman, G. M. (1994). Value-dependent selection in the brain: simulation in a synthetic neural model. *Neuroscience*, *59*, 229–243.

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, *30*, 51–64.

Hoffman, M. W., Grimes, D. B., Shon, A. P., & Rao, R. P. N. (2006). A probabilistic model of gaze imitation and shared attention. *Neural Networks*, *19*, 299–310.

Horley, K., Williams, L. M., Gonsalvez, C., & Gordon, E. (2004). Face to face: visual scanpath evidence for abnormal processing of facial expressions in social phobia. *Psychiatry research*, *127*, 43–53.

Humphrey, K., & Underwood, G. (2010). The potency of people in pictures: Evidence from sequences of eye fixations. *Journal of Vision*, *10*.

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision research*, *51*, 1192–1205.

Itti, L., & Koch, C. (2001). Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging*, *10*, 161–169.

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

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on* (pp. 2106–2113). IEEE.

Kienzle, W., Wichmann, F. A., Schölkopf, B., & Franz, M. O. (2007). A nonparametric approach to bottom-up visual saliency. *Advances in neural information processing systems*, *19*, 689.

Klin, A., Lin, D., Gorrindo, P., Ramsay, G., & Jones, W. (2009). Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*, *459*.

Kobayashi, H., & Kohshima, S. (1997). Unique morphology of the human eye. *Nature*, *387*, 767–768.

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*, 219–227.

Kupfer, D., & Foster, F. G. (1972). Interval between onset of sleep and rapid-eye-movement sleep as an indicator of depression. *The Lancet*, *300*, 684–686.

Land, M. F., & Hayhoe, M. (2001). In what ways do eye movements contribute to everyday activities? *Vision research*, *41*, 3559–3565.

Land, M. F., & Lee, D. N. (1994). Where we look when we steer. *Nature*, *369*, 742–744.

Langton, S. R. (2000). The mutual influence of gaze and head orientation in the analysis of social attention direction. *The Quarterly Journal of Experimental Psychology: Section A*, *53*, 825–845.

Li, Z. (2002). A saliency map in primary visual cortex. *Trends in cognitive sciences*, *6*, 9–16.

Lobmaier, J. S., Fischer, M. H., & Schwaninger, A. (2006). Objects capture perceived gaze direction. *Experimental Psychology*, *53*, 117.

Lungarella, M., Metta, G., Pfeifer, R., & Sandini, G. (2003). Developmental robotics: a survey. *Connection Science*, *15*, 151–190.

Marin-Jimenez, M., Zisserman, A., Eichner, M., & Ferrari, V. (2014). Detecting people looking at each other in videos. *International Journal of Computer Vision*, (pp. 1–15).

Mathe, S., & Sminchisescu, C. (2012). Dynamic eye movement datasets and learnt saliency models for visual action recognition. In *Computer Vision–ECCV 2012* (pp. 842–856). Springer.

Meng, Q., & Song, Y. (2012). Text detection in natural scenes with salient region. In *Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on* (pp. 384–388). IEEE.

Milanese, R., Wechsler, H., Gill, S., Bost, J.-M., & Pun, T. (1994). Integration of bottom-up and top-down cues for visual attention using non-linear relaxation. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on* (pp. 781–785). IEEE.

Murphy-Chutorian, E., Doshi, A., & Trivedi, M. M. (2007). Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE* (pp. 709–714). IEEE.

Murphy-Chutorian, E., & Trivedi, M. M. (2009). Head pose estimation in computer vision: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *31*, 607–626.

Nagai, Y., Asada, M., & Hosoda, K. (2002). A developmental approach accelerates learning of joint attention. In *Development and Learning, 2002. Proceedings. The 2nd International Conference on* (pp. 277–282). IEEE.

Navalpakkam, V., & Itti, L. (2007). Search goal tunes visual features optimally. *Neuron*, *53*, 605–617.

Norris, J. R. (1998). *Markov chains*. 2008. Cambridge university press.

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of vision*, *10*.

Okumura, Y., Kanakogi, Y., Kanda, T., Ishiguro, H., & Itakura, S. (2013). The power of human gaze on infant learning. *Cognition*, *128*, 127–133.

Ozuysal, M., Calonder, M., Lepetit, V., & Fua, P. (2010). Fast keypoint recognition using random ferns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *32*, 448–461.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, *1*, 515–526.

Rabiner, L., & Juang, B.-H. (1986). An introduction to hidden markov models. *ASSP Magazine, IEEE*, *3*, 4–16.

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*, 341–350.

Shen, J., & Itti, L. (2012). Top-down influences on visual attention during listening are modulated by observer sex. *Vision research*, *65*, 62–76.

Shepherd, S. V., Deaner, R. O., & Platt, M. L. (2006). Social status gates social attention in monkeys. *Current Biology*, *16*, R119–R120.

Subramanian, R., Shankar, D., Sebe, N., & Melcher, D. (2014). Emotion modulates eye movement patterns and subse-quent memory for the gist and details of movie scenes. *Journal of Vision*, .

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*, 4.

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, *113*, 766–786.

Treisman, A., & Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology.*, *12*, 97–136.

Triesch, J., Ballard, D., Hayhoe, M., & Sullivan, B. (2003). What you see is what you need. *Journal of Vision.*, *3*, 86–94.

Triesch, J., Teuscher, C., Deák, G. O., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental science*, *9*, 125–147.

Tseng, P., Cameron, I. G. M., Pari, G., Reynolds, J. N., Munoz, D. P., & Itti, L. (2012). High-throughput classification of clinical populations from natural viewing eye movements. *Journal of Neurology*, .

Valenti, R., Sebe, N., & Gevers, T. (2012). Combining head pose and eye location information for gaze estimation. *Image Processing, IEEE Transactions on*, *21*, 802–815.

Vincent, B. T., Baddeley, R. J., Troscianko, T., & Gilchrist, I. D. (2009). Optimal feature integration in visual search. *Journal of Vision*, *9*, 15.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on* (pp. I–511). IEEE volume 1.

Wang, H.-C., & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of vision*, *12*, 26.

Wang, J., Sung, E., & Venkateswarlu, R. (2003). Eye gaze estimation from a single image of one eye. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on* (pp. 136–143). IEEE.

Weidenbacher, U., Layher, G., Bayerl, P., & Neumann, H. (2006). Detection of head pose and gaze direction for human-computer interaction. In *Perception and Interactive Technologies* (pp. 9–19). Springer.

Wood, E., & Bulling, A. (2014). Eyetab: model-based gaze estimation on unmodified tablet computers. In *Proceedings of the Symposium on Eye Tracking Research and Applications* (pp. 207–210). ACM.

Yarbus, A. (1967). *Eye movements and vision.*. New York: Plenum.

Ye, Z., Li, Y., Fathi, A., Han, Y., Rozga, A., Abowd, G. D., & Rehg, J. M. (2012). Detecting eye contact using wearable eye-tracking glasses. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing* (pp. 699–704). ACM.

Yücel, Z., Salah, A., Meriçli, C., Meriçli, T., Valenti, R., & Gevers, T. (2013). Joint attention by gaze interpolation and saliency. *Cybernetics, IEEE Transactions on*, *43*, 829–842.

Yun, K., Peng, Y., Samaras, D., Zelinsky, G. J., & Berg, T. L. (2013). Exploring the role of gaze behavior and object detection in scene understanding. *Frontiers in Psychology.*, *4*.

Zhao, Q., & Koch, C. (2011). Learning a saliency map using fixated locations in natural scenes. *Journal of vision*, *11*, 9.

Zhu, X., & Ramanan, D. (2012). Face detection, pose estimation, and landmark localization in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on* (pp. 2879–2886). IEEE.