Vision Research xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

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

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

[a] Neuroscience Graduate Program, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089, USA
[b] Department of Computer Science, University of Wisconsin – Milwaukee, PO Box 784, Milwaukee, WI 53211, USA
[c] Department of Computer Science, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089, USA
[d] Department of Psychology, University of Southern California, 3641 Watt Way, Los Angeles, CA 90089, USA

ARTICLE INFO

Article history:
Received 3 July 2014
Received in revised form 22 October 2014
Available online xxxx

Keywords:
Visual attention
Eye movements
Gaze following
Head pose detection
Saliency modeling
Fixation prediction

ABSTRACT

Previous studies have shown that gaze direction of actors in a scene influences eye movements of passive observers during free-viewing (Castelhano, Wieth, & Henderson, 2007; Borji, Parks, & Itti, 2014). 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 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, which we call the Dynamic Weighting of Cues model (DWOC), 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.

1. Introduction

Where do people look during free-viewing of natural scenes? A tremendous amount of research in the cognitive and visual neuroscience 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, Frady, & Koch, 2009; Humphrey & Underwood, 2010; Judd et al., 2009), text (Wang & Pomplun, 2012), object center prior (Nuthmann & Henderson, 2010), image center prior (Tatler, 2007), semantic object distance (Hwang, Wang, & Pomplun, 2011), scene global context (Torralba et al., 2006), emotions (Subramanian et al., 2014), memory (Carmi & Itti, 2006; Droll et al., 2005), gaze following (Borji, Parks, & Itti, 2014; Castelhano, Wieth, & Henderson, 2007), culture (Chua, Boland, & Nisbett, 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, Sihite, & Itti, 2013; Einhäuser, Spain, & Perona, 2008; Hwang, Wang, Pomplun, 2011; Nuthmann & Henderson, 2010). Also note that, while here we focus on free-viewing tasks, some of these factors also play a role in task-driven visual attention (Ballard, Hayhoe, & Pelz, 1995; Borji & Itti, 2014; Borji, Sihite, & Itti, 2014; Einhäuser, Rutishauser, & Koch, 2008; Land & Hayhoe, 2001; Land & Lee, 1994; Triesch et al., 2003; Yarbus, 1967).

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. Understanding how humans selectively attend to visual information can give us insight into what components of the stimuli modulate this

* Corresponding author at: University of Southern California, Hedco Neuroscience Building, 3641 Watt Way, Los Angeles, CA 90089-2520, USA.
E-mail addresses: danielfp@usc.edu (D. Parks), borji@uwm.edu (A. Borji), itti@pollux.usc.edu (L. Itti).

http://dx.doi.org/10.1016/j.visres.2014.10.027
0042-6989/

attention. Further, we can analyze any differences in this selective attention that occur among populations. Replicating such behavior is also becoming more important recently due to the abundance of visual data, especially in computer vision and robotics.

Several works have looked at subject gaze in task dependent scenarios: toddler joint play (Smith, Yu, & Pereira, 2011), sentence disambiguation (Tanenhaus et al., 1995), making a sandwich (Ballard & Hayhoe, 2009), and gaze following as a social cue to disambiguate objects (Hanna & Brennan, 2007). In all these cases there was an explicit task and/or another agent actively cueing the subject. Here we wish to extract information from the subjects through their eye movements that is relevant to their future eye movement behavior without any explicit task demands.

In studies, perceived gaze facilitates covert attention in the direction of gaze, even when the gaze cue is not informative of the task (Driver et al., 1999; Friesen & Kingstone, 1998). Perceived gaze has also been shown to facilitate overt attention if the gaze is in the same direction as the task direction, and hamper it if the gaze is not (Hietanen, 1999; Langton & Bruce, 1999; Ricciardelli et al., 2002; Kuhn & Kingstone, 2009). This work suggests that gaze following cues are automatically processed, regardless of the current task, and not under the explicit control of the subject. As a result, a complete model of bottom up attention should include gaze cueing as well as the more standard simple features (orientations, color, intensity) and head detections used in current models (Cerf et al., 2007).

1.1. Gaze following in natural scenes

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

Gaze direction has been shown to influence visual attention and eye movements in free-viewing. Birmingham, Bischof, and Kingstone (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, Wieth, and Henderson (2007) showed that observers follow the gaze directions of actors in the scene during the presentation of a slideshow. Borji, Parks, and Itti (2014) 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 the 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 and Ramanan (2012), who proposed a model for head pose detection (only for the yaw direction) in natural scenes (see also Murphy-Chutorian, Doshi, & Trivedi (2007), Weidenbacher et al. (2006), Yücel et al. (2013), Valenti, Sebe, & Gevers (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, Sung, & Venkateswarlu (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, Parks, & Itti, 2014), 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 predicting 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 and Itti (2013).

Saliency models, influenced by the Feature Integration Theory (FIT) of Treisman and 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, Koch, & Niebur, 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, Koch, and Niebur (1998), Ehinger et al. (2009), Cerf, Frady, and Koch (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 and 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 and 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 and Itti (2007) proposed an optimal gains theory

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

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, McAllester, and Ramanan (2008)) and learned a saliency model with a linear support vector machine (SVM). Zhao and 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, non-linear 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 and 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 an SVM with Gaussian radial basis functions (RBF).

Task-driven attention models have been reviewed recently, where they focus on modeling the cognitive agenda involved in task specific behavior (Hayhoe and Ballard, 2014). Task-related modeling is highly effective in situations where there is strong top-down guided behavior (driving, making a sandwich, etc.), and useful anytime it is known or can be estimated. In this paper, we focus on free-viewing behavior, where the cognitive agenda is not known, and in our view is not a strong driver of behavior. Both aspects are highly relevant to human behavior, and in order to be fully understood, they both must be integrated. When a baseball flies at the head of a person making a sandwich, they will look, in spite of a top-down agenda. In the same way, the cognitive agenda is never fully extinguished even in free-viewing conditions.

Our prior work (Borji, Parks, & Itti, 2014) 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. These images were interspersed with images without humans (60 with and 60 without) to mask the purpose of the experiment. We then evaluated gaze following on a Flickr dataset (used here), where we also found a similar effect. 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 head pose, gaze following, and low-level saliency 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.

2. Flickr gaze dataset

We use data from our previous study (Borji, Parks, & Itti, 2014). 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 each had a 5 min 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. Although computer displays are not the same as real world 3D environments, some studies have established similarities between free-viewing setups and real world viewing in addition to showing that saliency is predictive in both conditions Betsch et al. (2004), Marius't Hart et al. (2009).

30 students from the University of Southern California (4 male, 26 female) were shown images for 10 s each with a 5 s gray screen in between images. Stimuli were presented at 60 Hz refresh and at a resolution of 1920×1080 pixels subtending $45.5^\circ \times 31^\circ$ of visual angle. The experimental methods were approved by the University of Southern California's Institutional Review Board (IRB). Fig. 1 shows sample images along with annotated objects and head pose directions.

The instructions given to the subjects was to "watch and enjoy" the pictures in a free-viewing paradigm. Note that like any free-viewing experiment, although no task was given, we cannot claim to know the cognitive agenda of the subjects. Please refer to Borji, Parks, and Itti (2014) 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 3 fixations land on a head, and less than 1 in 6 fixations leave a head region. This places an upper bound on how much improvement head related information can provide in improving fixation prediction.

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. Fig. 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.

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

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

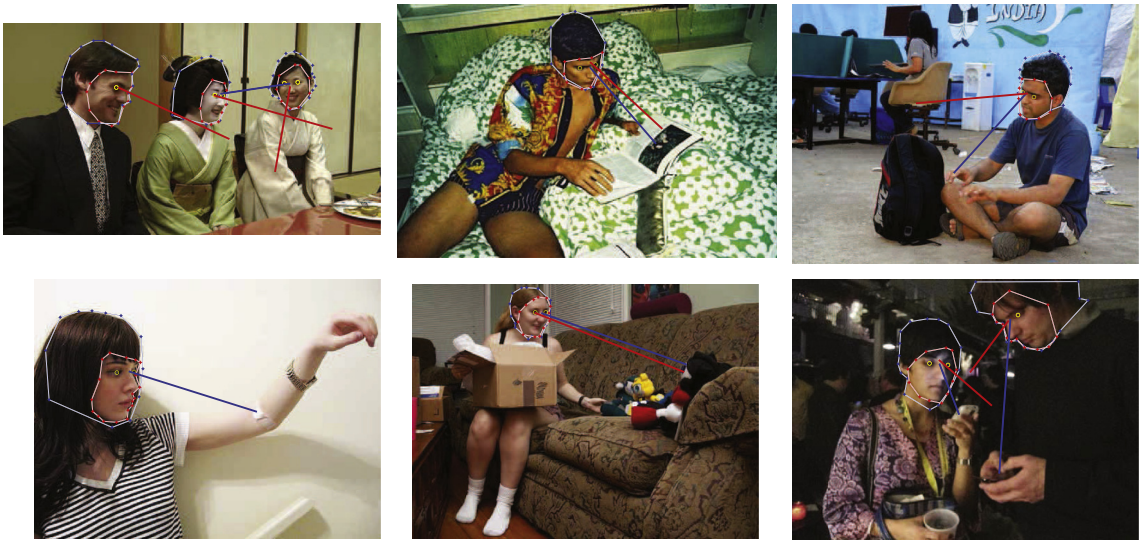

Fig. 1. Sample images from our dataset along with annotated object boundaries and head pose annotations. The face regions (red), whole head (blue), eyes (yellow), head pose 2D angle (red), and final eye gaze 2D angle (blue) are labeled.

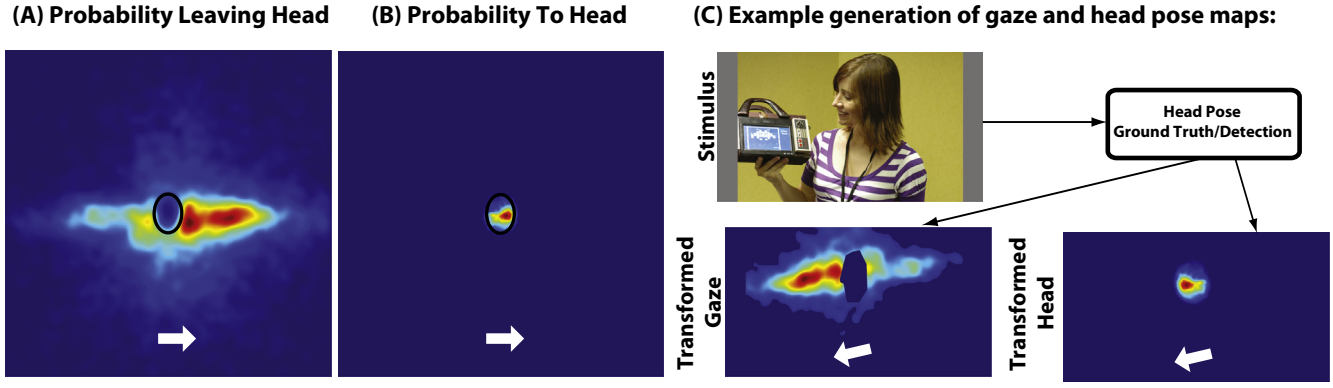

Fig. 2. (A) Average fixation probability after leaving a normalized head over the whole dataset, where all head poses were rotated to point to the right (white arrow), and sized to a nominal head size (black oval). (B) Average fixation probability for all head fixations over the whole dataset, where all head poses were rotated to point to the right and size to the nominal head size. (C) An example image which had head pose detections applied. The gaze following and head region maps are then translated, scaled, and rotated according to the head position, head size and head pose image angle, respectively. This is done for each head detection.

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 Fig. 2B, the head fixations had a high density in the eye region, with low density in the hair and chin regions. Once these maps were learned over the training set, they were able to be used in testing. For each head detection, during testing, each map is translated and scaled based on the position and size of the head pose detection. The maps are then rotated about the center of the head based on the head pose angle as shown in Fig. 2C. Pajak and Nuthmann (2013) showed that during saccade targeting for scene perception, the centers of objects were preferred, and this scaled with object size. Since objects near a person's face are likely to be the same distance from the camera as the person, we felt that we should scale the saccade vectors based on the size of the head to control for when faraway humans were looking at faraway objects and nearby humans were looking at nearby objects.

As we can see in the gaze following heat map, saccades leaving the head seem to follow the head pose direction, albeit in a weak manner. In our prior paper (Borji, Parks, & Itti, 2014), we had another set of subjects ($n = 15$) estimate the gaze direction of the heads in the Flickr dataset used here (by drawing a line), and the standard deviation was 11.77°. This suggests that the diffuseness of the map is due to the weakness of the cue, and not just the uncertainty of the angle estimate.

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., 2013), and because it does not 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 Fig. 2A and B,

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

respectively, to form cue maps for each stimulus image as shown in Fig. 2C. Fig. 3 shows what combined head pose and gaze following maps look like compared to a simple cone model 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 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. From here on, we will refer to our model as the Dynamic Weighting of Cues model (DWOC).

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 Fig. 4.

This proposed model, like the other fixation prediction models that we reference, only predicts the next saccade. The current position is used to determine, either from ground truth annotations or from automatic head detections, the current state of the model (Head or NonHead). Because the fixation space in an HD image is large already, raising it to the *n*th power, where *n* is the number of consecutive saccades to predict, causes the space to quickly grow very large. This makes it infeasible to generate an inter-observer model as a benchmark of performance, due to the inability to adequately sample the space.

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

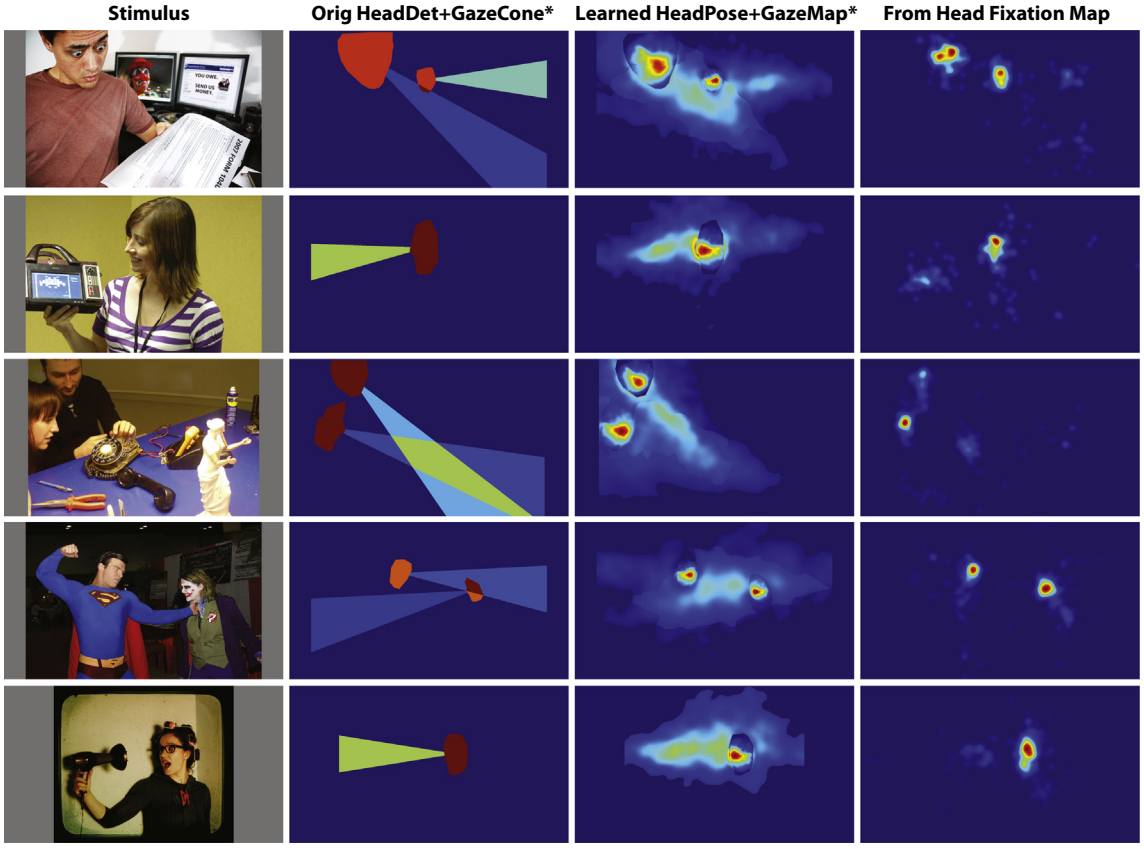

Fig. 3. Sample images comparing the original head detection + gaze, and the learned head pose + gaze model, along with fixations coming from the head. * Note that the maps are given equal weights for the two cues in both cases.

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

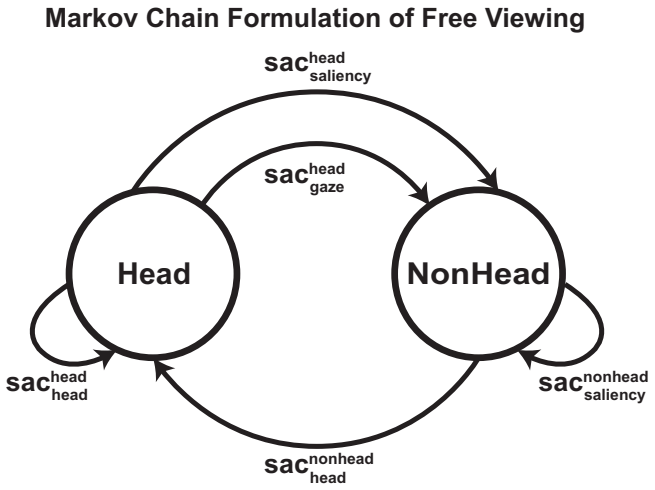

Fig. 4. Discrete-time Markov chain formulation of saccadic eye movement of an observer performing a free-viewing task.

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 Fig. 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. When scoring "to gaze" transitions, since gaze following is defined here to be valid only when leaving a gazing head, just the current head's gaze is used to build the map. For "to head" transitions, the current head (if at a head state) vs. other heads are not differentiated for simplicity.

The transition probabilities try and capture the relative importance that a subject will consider a particular cue, given their current fixation type (head or non-head). Many cognitive processes can be involved in modulating these probabilities, which are not modeled here, and the probabilities are likely to be dependent on the subjects own biases and the dataset. However, they can be extracted without being privy to the task of the subject while still allowing us to utilize the current fixation behavior at a more abstract level then simple spatial position. This strategy could be expanded to other cue types, for instance, text, where the simple act of looking at the text could provide information as to what cues are currently important to a subject. We learned the transition probabilities for several systems involving the following cues: head detections (H), head pose (P), gaze following (G), and saliency (S). The transition probabilities for heads and head pose were considered to be interchangeable, with only the map changing and not the weight. One 3-component model (PSG) was learned, along with several 2-component models (PG, HS, PS).

Fig. 5 shows the learned weights for the HSG (or PSG) model as well as the HS (or PS) model. It shows that 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

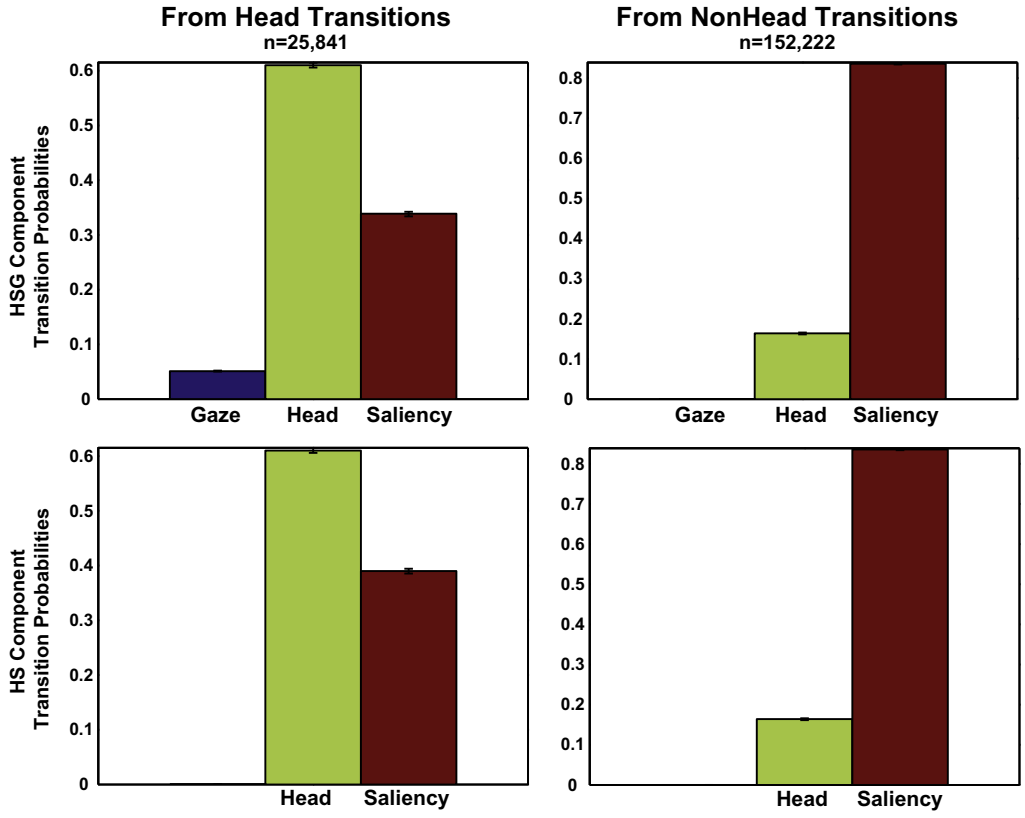

Fig. 5. Transition probabilities when saccading to the head regions, gaze regions, or salient regions when starting from a head region or a non-head region. Note that saliency is much stronger when originating from a non-head region. Head detections (H) and head pose (P) cues are considered to have interchangeable transition probabilities. 5th and 95th percentile confidence intervals are shown across the cross validation folds.

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

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 saliency maps were all made into probability maps (i.e., sum to 1). The transition probabilities for a particular model were used to weigh the respective maps, which were then summed to create the final maps. For instance, maps for the $DWOC_{PSG}$ model are shown in the second column from the right in Fig. 6. This model was weighted using the following equation:

$$DWOC_{PSG}map = p(sac_{head})cue_{head\_pose} + p(sac_{salient})cue_{salient} + p(sac_{gaze})cue_{gaze} \quad (1)$$

For comparison purposes, several simpler 2-component systems were constructed, which used only two of the cues. The $DWOC_{HS}$ baseline system takes head detections and bottom-up saliency into account without head pose or gaze information. The $DWOC_{PG}$ system models how well head pose and gaze can perform alone. The $DWOC_{PS}$ model, is an intermediate model that shows what adding the head pose cue, but not gaze following cue can perform. After the transitions were learned, the maps were combined to form the combined probability maps using the following equations:

$$DWOC_{PG}map = p(sac_{head})cue_{head\_pose} + p(sac_{gaze})cue_{gaze} \quad (2)$$

$$DWOC_{HS}map = p(sac_{head})cue_{head\_det} + p(sac_{salient})cue_{salient} \quad (3)$$

$$DWOC_{PS}map = p(sac_{head})cue_{head\_pose} + p(sac_{salient})cue_{salient} \quad (4)$$

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

4. Annotation based fixation prediction results

In order to assess the performance of the individual components of the model along with various combinations of these components we generated Receiving Operator Characteristic (ROC) curves, and determined the area under the curve (AUC). An AUC value of 0.5 would indicate that the model is predicting fixations at a purely random level. ROC curves are commonly used to show the performance of a binary classifier that has a discrimination threshold. Here the classification is simply whether a point on an image was attended to or ignored. The *x*-axis on the ROC curve represents the false positive rate, and the *y*-axis represents the true positive rate. A perfect system would order the image positions according to the probabilities in the fixation map (e.g. right-most column in Fig. 6). This would result in a ROC curve that rises to 1 immediately, and would have an area of 1 under the curve (AUC). A completely random model would generate an AUC value of 0.5.

The $DWOC_{HS}$ 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, like the full model, $DWOC_{PSG}$. Here we hypothesize that including gaze and head pose information will improve the predictive power of a fixation prediction model. We also compare the different components of the model along with different combinations of components to evaluate the relative importance of these components. Fig. 7A shows the AUC performance for gaze, head detections, head pose, saliency, $DWOC_{PG}$, $DWOC_{HS}$, $DWOC_{PS}$, and $DWOC_{PSG}$ 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 gaze cue is the learned map of gaze following, while "head dets" is a uniform map of the actual head detections. Head pose applies the learned head pose map to each head detection.

The Wilcoxon signed rank test was used in all of the following comparisons. Our initial null hypothesis, that there is no difference between current state of the art ($DWOC_{HS}$) and our model ($DWOC_{PSG}$), was rejected $DWOC_{HS} < DWOC_{PSG}$ ($p < 0.002$). We also ran post hoc comparisons of adjacent models and components for illustration purposes, which were adjusted using Holm–Bonferonni (Holm, 1979) for multiple comparisons ($n = 8$). They 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

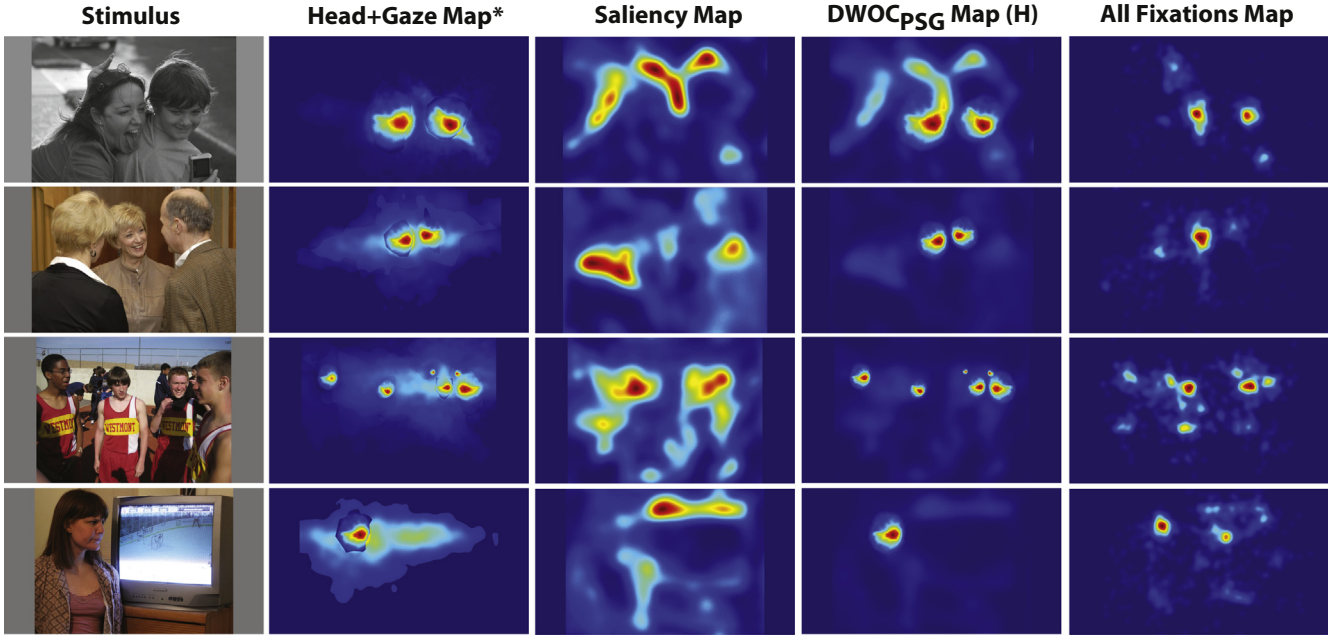

Fig. 6. Sample images with corresponding maps for Head + Gaze, Saliency, and the final $DWOC_{PSG}$ model, along with maps of all fixations. The combined $DWOC_{PSG}$ map is shown with the weights from the Heads state. * Note that the maps in the Head + Gaze column gives equal weights to the cues for illustration purposes, but all DWOC maps, including the $DWOC_{PSG}$ column, use the learned weights.

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

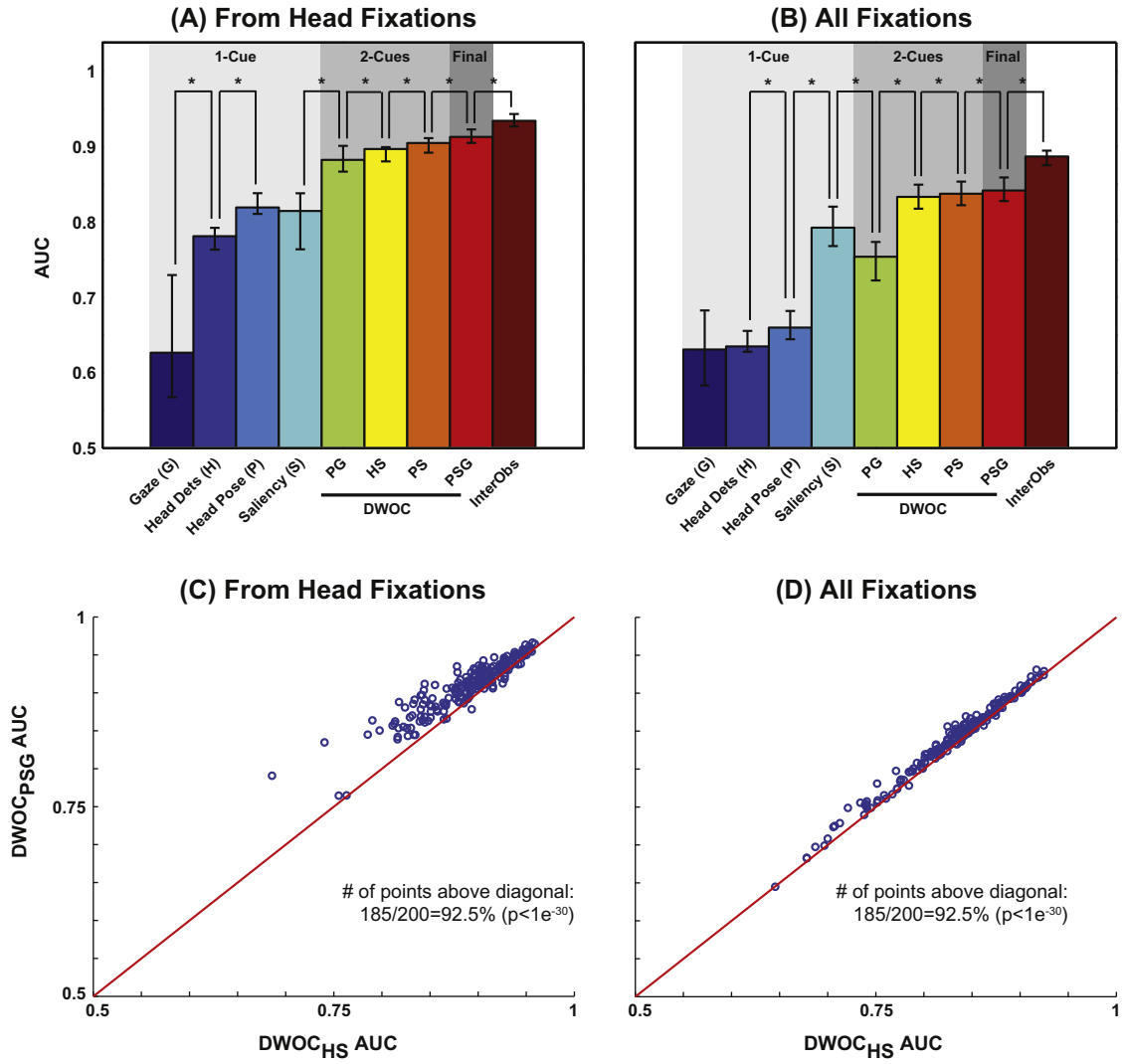

Fig. 7. Performance of each component using ground truth. AUC performance on saccades originating (A) from the head and (B) all saccades. An inter-observer model is shown to provide a ceiling for performance. 5th and 95th percentile confidence intervals are shown. There is also a per image AUC comparison between the $DWOC_{HS}$ model and the $DWOC_{PSG}$ model for (C) from the head and (D) all saccades.

comparisons here the winning model won on every fold. For from head fixations, the following adjacent cues were found to be significant: Gaze < Head Detections ($p < 0.02$), Head Detections < Head Pose ($p < 0.02$), Saliency < $DWOC_{PG}$ ($p < 0.02$), $DWOC_{PG} < DWOC_{HS}$ ($p < 0.02$), $DWOC_{HS} < DWOC_{PS}$ ($p < 0.02$), $DWOC_{PS} < DWOC_{PSG}$ ($p < 0.02$), $DWOC_{PSG}$ < Inter-observer[2] ($p < 0.02$). Note that the $DWOC_{PG}$ head pose and gaze model outperforms bottom up saliency when looking at from head fixations.

There were 184,061 total fixations in the data, with 53,928 or 29.3% 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 Fig. 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. With all fixations, we were again able to show that $DWOC_{HS} < DWOC_{PSG}$ ($p < 0.002$). Adjacent models were compared, post hoc, again with Holm–Bonferroni correction, for all fixations. The $DWOC_{PSG}$ model outperformed the $DWOC_{PS}$ model ($p < 0.02$) which outperformed the $DWOC_{HS}$ model ($p < 0.02$). The following cues that are significantly different: Head Detections < Head Pose ($p < 0.02$), Head Pose < Saliency ($p < 0.02$), $DWOC_{PG}$ < Saliency ($p < 0.02$), $DWOC_{PG} < DWOC_{HS}$ ($p < 0.02$), $DWOC_{HS} < DWOC_{PS}$ ($p < 0.02$), $DWOC_{HS} < DWOC_{PSG}$ ($p < 0.02$), and $DWOC_{PSG}$ < Inter-observer ($p < 0.02$).

As we can see from these results, head pose and gaze $DWOC_{PG}$ works better for from head fixations, while saliency works better for other fixations. This validates the use of different weights for these maps when starting from the head vs. other regions. Head pose by itself also performs better when starting from the head, but does not account for all of the performance improvement.

Fig. 7C and D shows the per image AUC performance of the $DWOC_{HS}$ and $DWOC_{PSG}$ models for from head fixations and all fixations, respectively. Here we see that the pose information improves fixation prediction when fixations are from the head, and this is enough to improve the performance of all fixations ($DWOC_{HS} < DWOC_{PSG}$ ($p < 2e{-}31$) for fixations from the head, and $DWOC_{HS} < DWOC_{PSG}$ ($p < 4e{-}32$) for all fixations). In 185 of the 200 images, AUC performance for both from head fixations and all fixations improved when using head pose and gaze following.

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

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

For from head fixations, this improvement on a per image basis was, on average, 2.1%.

To determine where our combined model and bottom-up saliency diverge, we also plot the performance of Saliency vs. $DWOC_{PSG}$ in Fig. 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 dog, although one could argue that we could have labeled non-human faces as well to account for this. Image 3 shows where the head pose information complements low level saliency because the heads are not found to be salient. In image 4, both Saliency and $DWOC_{PSG}$ are not performing well, and it is interesting to note that both models miss another 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 saliency 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.

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 head pose and gaze 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 Fig. 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 and Ramanan (2012) model (Zhu & Ramanan, 2012) that we used to detect heads and extract head pose only provided a yaw estimate. Fig. 9B shows the basic structure of their model. Local parts of the face (68 landmarks for -45° to 45° frontal faces; 39 landmarks for side views) are detected using mixture of trees part detectors (similar to the detectors used in Felzenszwalb, McAllester, and Ramanan (2008), and the relative locations of the detected parts create deformations in the larger pose model at a certain cost. This model is trained for heads that are at least 80×80 pixels, and given the number of landmarks used, this is a limiting constraint. Further, the yaw angle is limited to 90° to -90°, where 0° is facing the camera, since the landmarks are exclusive to the face and jawline.

After the head pose and landmarks are returned from the original Zhu model and the estimated positions of the 68 or 39 landmarks are generated, the X and Y positions of these landmarks are used to generate the pitch estimate. This is done using a set of random binary ferns (Ozuysal et al., 2010). These ferns make a series of binary comparisons ($n = 3$) and for each possible result of those comparisons a fern value, f_{val} is generated (010, 001, etc., a total of 8 possible values). A histogram of the pitch angle categories, ang_{cat}, (-22.5°, 0°, and 22.5°) is learned over the training data for each f_{val} separately. This means that all of the training data for a fern that has an f_{val} of 010 is put together and the pitch angle

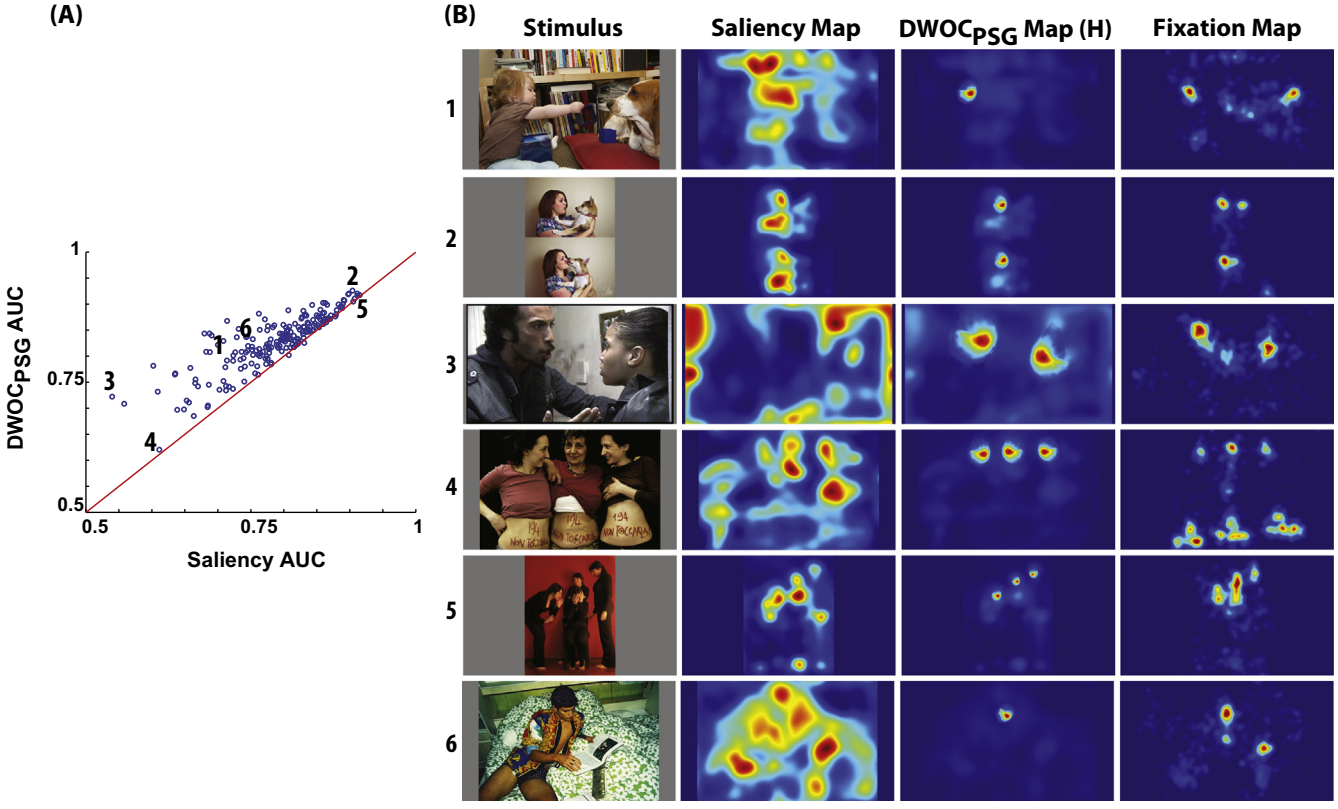

Fig. 8. (A) Scatter plot showing per image AUC performance of saliency vs. the $DWOC_{PSG}$ model. (B) Four sample points on the scatter plot with images and their corresponding maps are shown for illustration. The combined $DWOC_{PSG}$ map is shown using the weights from the Head (H) state.

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

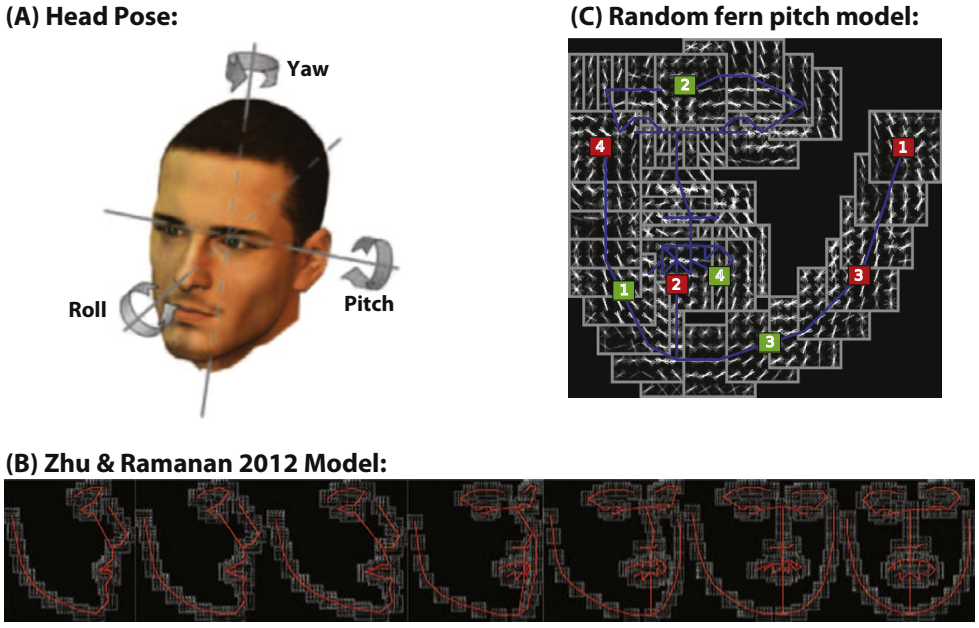

Fig. 9. (A) [Modified from Murphy-Chutorian and Trivedi (2009)] 3D head pose angle of a head can be defined egocentrically using *roll*, *pitch*, and *yaw*. (B) The head pose model (Zhu & Ramanan, 2012) has 146 shared local parts across 13 learned poses, with learned deformation costs (in red) between 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. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

histogram is formed on just that data. During testing, each fern's value is computed and the corresponding learned pitch angle histograms are added together across all of the ferns. The maximum pitch angle bin is then considered to be the pitch angle.

The binary comparisons that are used to compute a fern's value are the *X* or *Y* positions of two randomly chosen landmarks. If you look at Fig. 9C, you can see an example of the binary comparison computation for a single 4-bit fern. Here you can see a set of detected landmarks, each with a particular *X*, *Y* center position. When each fern was initially generated, these landmarks were randomly sampled, where each bit of the fern has a green and a red landmark and an axis (*X* or *Y*). If the axis position of the green landmark is greater than the axis position of the red landmark ($axis_{green} > axis_{red}$), then the fern bit will be 1, and 0 otherwise. This is done for all n bits of the fern and results in a particular binary value, such as 0110 for a 4-bit fern, referred to here as f_{val}.

During testing, the histogram for each calculated 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°, and −22.5°), and was only trained on heads that were 45° to −45°.

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 Fig. 10. The system generates a yaw and pitch angle in addition to a confidence score. Because discrete objects like heads have an unknown integer number of appearances in a given image, a signal detection score like ROC is not as useful. Precision and recall are commonly used when discussing the accuracy in which an object detector performs, for example when measuring the performance on ImageNet, a large scale object recognition competition (Russakovsky et al., 2014). Precision is the rate at which true positives are identified relative to the total number of positives declared by a model. Recall is the rate at which true positives are identified relative to the total number of actual objects present.[3] Here we use the F_1 score to evaluate performance, which is the harmonic mean of precision and recall to report performance, which is useful when averaging rates. F_1 is defined to be: $F_1 = 2PR/(P + R)$, and is bounded by 0 and 1, inclusively.

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°, and −60°) for the yaw angle, and up, level, and down (22.5°, 0°, and −22.5°) for the pitch angle. Therefore, chance for both angles is 33.3%. The performance of the system over the 10 folds of the Flickr set is shown in Table 1.

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 provided head polygons and 2D pose angles, which were 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. Again the Wilcoxon signed rank test was used for all comparisons.

Fig. 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

[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.

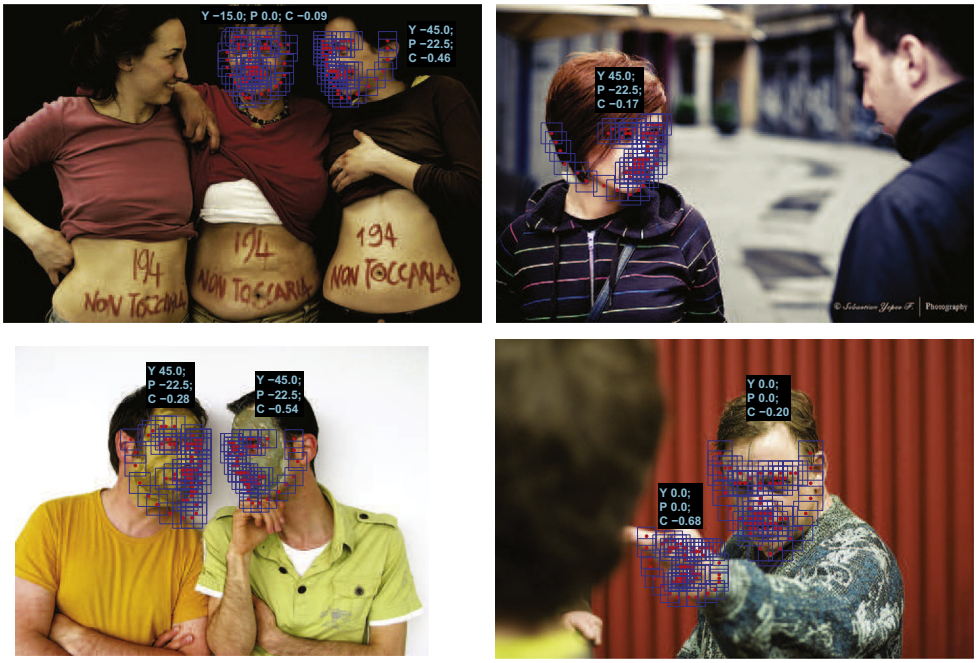

Fig. 10. Sample head pose model detections. *Y* is yaw angle with 0° pointing out of the image and positive moving to the left from the head's perspective. *P* is the pitch angle with 0° being level and positive looking up. Both are in degrees. The confidence of the detection is shown with C, and is 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
Head detection	0.75	0.05	–
Yaw detection	0.82	0.07	0.33
Pitch detection	0.56	0.07	0.33

data. For from head fixations, our initial null hypothesis was again rejected, $DWOC_{HS} < DWOC_{PSG}$ ($p < 0.002$). Adjacent models and components were again compared using Holm–Bonferonni for multiple comparisons ($n = 8$). Significance was found for the following: Head Detections < Gaze ($p < 0.02$), Head Detections < Head Pose ($p < 0.02$), Saliency < $DWOC_{PG}$ ($p < 0.02$), $DWOC_{PG} < DWOC_{HS}$ ($p < 0.02$), $DWOC_{HS} < DWOC_{PS}$ ($p < 0.02$), $DWOC_{PS} < DWOC_{PSG}$ ($p < 0.02$), and $DWOC_{PSG}$ < Inter-observer ($p < 0.02$).

When looking at all fixations, as in Fig. 11B, the difference is again smaller, and the saliency model alone accounts for most of the performance. Still, however, the differences were significant. For all fixations, the $DWOC_{PSG}$ model outperforms the $DWOC_{PS}$ model ($p < 0.02$) which outperforms the $DWOC_{HS}$ model ($p < 0.02$). The adjacent cues that were found to be significantly different: Head Detections < Head Pose ($p < 0.02$), Head Detections < Head Pose ($p < 0.02$), Head Pose < Saliency ($p < 0.02$), $DWOC_{PG}$< Saliency ($p < 0.02$), $DWOC_{PG} < DWOC_{HS}$ ($p < 0.02$), $DWOC_{HS} < DWOC_{PS}$ ($p < 0.02$), $DWOC_{HS} < DWOC_{PSG}$ ($p < 0.02$), and $DWOC_{PSG}$ < Inter-observer ($p < 0.02$). The drop in our combined

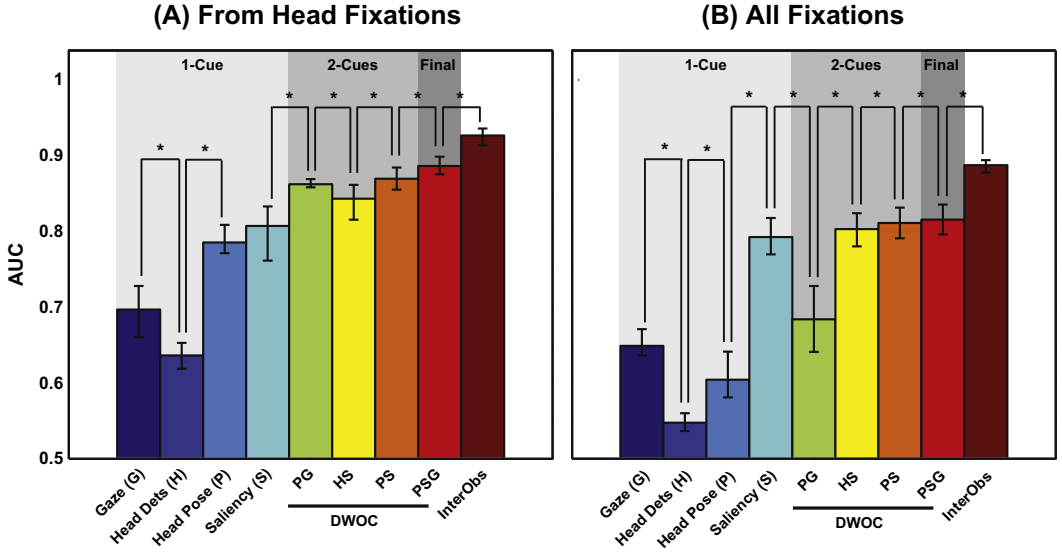

Fig. 11. Performance of each component using head pose detections. AUC performance on saccades originating (A) from the head and (B) all saccades. An inter-observer model is shown to provide a ceiling for performance. 5th and 95th percentile confidence intervals are shown.

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

$DWOC_{PSG}$ model AUC for from head fixations from a mean of 0.913 to 0.886 when changing from ground truth head pose angles in Section 4 to using the automatic head pose estimates in Section 6 shows the deficit due to detection performance.

7. Discussion

We showed that head pose and gaze following information alone 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.

In our prior work (Borji, Parks, & Itti, 2014), it was shown that when the direction of other faces were incongruent with the gaze direction, gaze following was weaker. It was also weaker, but less so, when no other faces were present. This implies that gaze following is especially useful when predicting fixations in social scenes. There are several other factors that we did not model that could further contribute to fixation prediction. Text detection in the wild (e.g., (Meng & Song, 2012)) could extract a cue that has been shown to be useful in fixation prediction (Cerf, Frady, & Koch, 2009). The presumed semantic relevance of an object to an actor could also potentially be used. For instance, if an actor was holding a knife, a jar of peanut butter would likely be more semantically relevant than a book. The facial expression of a face could also differentially drive fixations and could be evaluated. Alternatively, threatening objects (gun, knife, etc.) and high-value objects (money, jewelry) could be evaluated. Quantifying these concepts would be difficult, but could represent some portion of the remaining performance.

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, Asada, & Hosoda (2002), Breazeal & Scassellati (2002), Bakeman & Adamson (1984)), determining the attention levels of a driver (e.g., Murphy-Chutorian, Doshi, & Trivedi, 2007), and enriching e-learning systems (e.g., Asteriadis, Karpouzis & Kollias, 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), and autism (Fletcher-Watson et al., 2009; Klin et al., 2009)). 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 to detect eye contact preference. As another example, Alghowinem et al. (2013) demonstrated that eye movement can be used as a means of depression detection. Several works have built models to distinguish patient populations from controls such as visual agnosics (Foulsham et al., 2009) as well as ADHD, FASD, and Parkinson's disease (Tseng et al., 2012).

Our proposed model can be used for distinguishing patient populations by looking at how each group differs in the relative weighting of the component cues. For instance, one might speculate that autistic individuals might exhibit significantly different transition probabilities than the control participants in the present study. Further, we can tailor our model to these specific populations to better predict each population's eye movements. The model can also be used to optimize the stimuli presented to the different populations. We look forward to addressing these areas in our future work.

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 two states (head and non-head) with different cue priorities. 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. In many cases, the cognitive agenda and current drives of a subject are not known. However, the learned weights of our model can give us insight into the cognitive biases of subjects when analyzing the differences between patient groups, cultures, and genders. Other cues that were not addressed here, such as text and motion, can also be integrated into this model to further enhance fixation prediction performance.

Acknowledgments

This work was supported by the National Science Foundation (Grant No. CCF-1317433), the Army Research Office (W911NF-12-1-0433), and the Office of Naval Research (N00014-13-1-0563). 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 Flickr dataset. We wish to thank reviewers for their valuable comments.

References

Alghowinem, S., Goecke, R., Wagner, M., Parkerx, G., & Breakspear, M. (2013). Head pose and movement analysis as an indicator of depression. In *Humaine association conference on Affective Computing and Intelligent Interaction (ACII), 2013* (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*, 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. H., & Hayhoe, M. M. (2009). Modelling the role of task in the control of gaze. *Visual Cognition, 17*, 1185–1204.

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*, 379–398.

Betsch, B. Y., Einhäuser, W., Körding, K. P., & König, P. (2004). The world from a cats perspective–statistics of natural videos. *Biological Cybernetics, 90*, 41–50.

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 *2012 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 438–445). IEEE.

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

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on 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. (2014). Complementary effects of gaze direction and early saliency in guiding fixations during free-viewing. *Journal of Vision, xx*, xx.

Borji, A., Sihite, D. N., & Itti, L. (2013). 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., Sihite, D., & Itti, L. (2014). What/where to look next? Modeling top-down visual attention in complex interactive environments. *IEEE Transactions on Systems, Man, and Cybernetics, Part A–Systems and Humans*.

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

Breazeal, C., & Scassellati, B. (2002). Challenges in building robots that imitate people. *Imitation in Animals and Artifacts*, 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., Harel, J., Einhäuser, W., & Koch, C. (2007). Predicting human gaze using low-level saliency combined with face detection.

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*.

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.

Driver, J., Davis, G., Ricciardelli, P., Kidd, P., Maxwell, E., & Baron-Cohen, S. (1999). Gaze perception triggers reflexive visuospatial orienting. *Visual Cognition, 6*, 509–540.

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.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency, 8(18), 1–26. doi: 10.1167/8.14.18.

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 *IEEE conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008* (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., Barton, J. J., Kingstone, A., Dewhurst, R., & Underwood, G. (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., et al. (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.

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review, 5*, 490–495.

Friston, K., Tononi, G., Reeke, G., Jr., 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.

Hanna, J. E., & Brennan, S. E. (2007). Speakers eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*, 596–615.

Hayhoe, M., & Ballard, D. (2014). Modeling task control of eye movements. *Current Biology, 24*, R622–R628.

Hietanen, J. K. (1999). Does your gaze direction and head orientation shift my visual attention? *Neuroreport, 10*, 3443–3447.

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.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 65–70.

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 *2009 IEEE 12th international conference on Computer Vision* (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.

Kuhn, G., & Kingstone, A. (2009). Look away! Eyes and arrows engage oculomotor responses automatically. *Attention, Perception, & Psychophysics, 71*, 314–327.

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.

Langton, S. R., & Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Visual Cognition, 6*, 541–567.

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

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*, 1–15.

Marius't Hart, B., Vockeroth, J., Schumann, F., Bartl, K., Schneider, E., Koenig, P., et al. (2009). Gaze allocation in natural stimuli: Comparing free exploration to head-fixed viewing conditions. *Visual Cognition, 17*, 1132–1158.

Meng, Q., & Song, Y. (2012). Text detection in natural scenes with salient region. In *2012 10th IAPR international workshop on Document Analysis Systems (DAS)* (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 *1994 IEEE computer society conference on Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94.* (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* (pp. 709–714). IEEE.

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

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027

Nagai, Y., Asada, M., & Hosoda, K. (2002). A developmental approach accelerates learning of joint attention. In *The 2nd international conference on development and learning, 2002. proceedings* (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. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 32*, 448–461.

Pajak, M., & Nuthmann, A. (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. *Journal of Vision, 13*, 2.

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. *IEEE ASSP Magazine, 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.

Ricciardelli, P., Bricolo, E., Aglioti, S. M., & Chelazzi, L. (2002). My eyes want to look where your eyes are looking: Exploring the tendency to imitate another individual's gaze. *NeuroReport: For Rapid Communication of Neuroscience Research, 13*, 2259–2264. http://dx.doi.org/10.1097/00001756-200212030-00018.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., & Ma, S., et al. (2014). Imagenet large scale visual recognition challenge. arXiv preprint arXiv:1409.0575,.

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.

Smith, L. B., Yu, C., & Pereira, A. F. (2011). Not your mothers view: The dynamics of toddler visual experience. *Developmental Science, 14*, 9–17.

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

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science, 268*, 1632–1634.

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. *IEEE Transactions on Image Processing, 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. *Proceedings of the 2001 IEEE computer society conference on Computer Vision and Pattern Recognition, 2001. CVPR 2001* (Vol. 1). IEEE, pp. I–511.

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 *Ninth IEEE international conference on computer vision, 2003. Proceedings* (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., et al. (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. *IEEE Transactions on Cybernetics, 43*, 829–842. http://dx.doi.org/10.1109/TSMCB.2012.2216979.

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 *2012 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2879–2886). IEEE.

Please cite this article in press as: Parks, D., et al. Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes. *Vision Research* (2014), http://dx.doi.org/10.1016/j.visres.2014.10.027