Attention Estimation by Simultaneous Observation of Viewer and View

Anup Doshi and Mohan M. Trivedi
Computer Vision and Robotics Research Lab
University of California, San Diego
La Jolla, CA 92093-0434
{andoshi,mtrivedi}@ucsd.edu

Abstract

We introduce a new approach to analyzing the attentive state of a human subject, given cameras focused on the subject and their environment. In particular, the task of analyzing the focus of attention of a human driver is of primary concern. Up to 80% of automobile crashes are related to driver inattention; thus it is important for an Intelligent Driver Assistance System (IDAS) to be aware of the driver state. We present a new Bayesian paradigm for estimating human attention specifically addressing the problems arising in dynamic situations. The model incorporates vision-based gaze estimation, "top-down"- and "bottom-up"-based visual saliency maps, and cognitive considerations such as inhibition of return and center bias that affect the relationship between gaze and attention. Results demonstrate the validity on real driving data, showing quantitative improvements over systems using only gaze or only saliency, and elucidate the value of such a model for any human-machine interface.

1. Introduction

In prior research associated with detecting human behavior in complex environments such as driving, it has sometimes been assumed that *attention* is indistinguishable from *gaze* [16]. However, early in the 20th century Helmholtz concluded that it is possible to attend to locations in the field of view without resorting to eye movements [5]. More recent research into visual attention has brought out the notion that attention and gaze are non-trivially interdependent. While attention generally must precede gaze shifts [6], a number of cognitive processes and distractions are known to affect the relationship between gaze and attention [16].

Computational models of visual attention, many recent ones based on saliency models [8], look to simulate human attentive mechanisms in order to understand the most meaningful regions of a scene. To estimate the attentive state of a human interacting with a complex, dynamic scene, it becomes necessary to observe the observer, to observe the scene, and to maintain some knowledge of the observer's tasks or cognitive state.

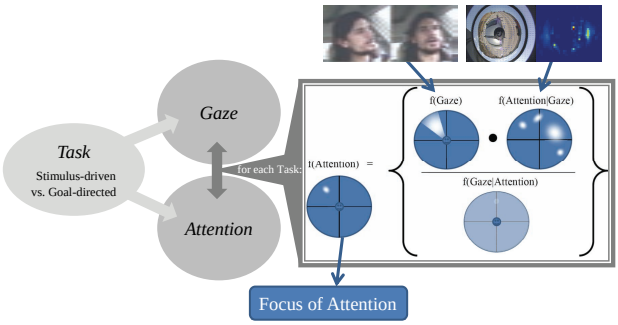

Figure 1. Graphical model relating Gaze, Attention, and Task. This leads to a Task-dependent Bayesian treatment of the joint densities of Gaze location and Attention location. Sample probability distributions for each term are shown below the terms. Vision-based **Gaze** estimation and **Saliency** maps are used to define the numerator terms, and the denominator encodes gaze inhibitors such as "inhibition of return," multi-tasking, and other cognitive distractions affecting Gaze **Orientation given Attention**.

In the following research we introduce a new integrated model to estimate the attentive state of a human subject engaged in a complex task such as driving. We start out by applying a Bayesian treatment to the joint probability distributions of gaze and attention, as shown in Figure 1. A marginalization of the "ongoing task" allows us to break the model down into semantically meaningful and computable terms.

It may be possible to encode many of the real-world cognitive phenomena that affect the relationship between gaze and attention, into the proposed model. This provides an elegant and principled way to estimate the subject's attentional state in dynamic environments, by incorporating vision-based estimates of both gaze and saliency.

The proposed system demonstrates a 45.67% improve-

978-1-4244-7028-0/10/$26.00 ©2010 IEEE

ment in Focus-of-attention estimation over a baseline system based on head pose alone, and a 63.62% improvement over a saliency-based attention estimation system. These results could generalize to any human-machine interface in complex environments where the sensors monitor the subject and surroundings simultaneously.

The remainder of this paper is organized as follows. We review related research and concepts in Section 2. In Section 3, we develop our proposed model and independently consider each relevant concept. The remainder of the paper contains experimental validation in the specific context of highway driving in Section 4, and finally concluding remarks in Section 5.

2. Related Research

While it may not be possible to measure attention directly, prior studies show that we can observe gaze behavior and also identify salient features of the environment to help identify focus-of-attention.

Robust monocular head pose estimation systems have been developed to identify attention patterns [12], though head pose is generally not a precise estimate of true gaze. More precision can be derived from eye gaze detectors; as discussed in Section 3.2, eye gaze tracking in vehicles is quite challenging.

Gaze estimation has helped to demonstrate that drivers' gaze behavior is predictable and task-oriented ("top down"), especially during lane changes [9]. However cognitive and visual ("bottom up") distractions [1] have a significant effect on the gaze patterns of drivers. The proposed model thus incorporates both "top down" and "bottom up" influences in vision-based estimation of gaze.

To determine what parts of a scene draw attention, saliency maps have been developed. Itti and Koch examined the eye glance patterns of human subjects on several scenes and built up a prior of that particular type of scene [8]. However, such "bottom up" saliency maps can not explain the fixations of goal-oriented observers [17].

Ultimately it is the interaction between an observer's goals, and the salient properties of the scene, that guide the observer's attention [19]. More recent versions of saliency maps have thus incorporated "top down" goals [13]. Given the highly dynamic scenes involved in driving, we use motion-based features to build up a "bottom up" saliency map, and combine this with a "top down" map dependent upon the task.

2.1. Focus of Attention in complex environments

Evidence has shown that in some cases behavior is influenced by objects that have not been explicitly attended to [16]. Moreover, "cognitive distractions" which take the subject's mind off the task at hand would inherently imply a de-coupling of gaze and attention. Cognitive psychologists have reported other relevant phenomena which could inhibit the relationship between gaze and attention, including "Inhibition of Return" and "Change blindness" [14]; these are discussed in part in more detail in Section 3.4. It is important to consider these factors in any computational model of attention, and especially in complex environments such as vehicles where these distractions are quite dangerous [10].

Several studies have proposed models to estimate the Focus of Attention in various environments [2, 12]. To the knowledge of the authors, this is the first study in which visual gaze and task-dependent salient targets are dynamically computed, and attentional likelihood is then modeled in a Bayesian manner. This allows a simple yet principled incorporation of many attention-related phenomena, along with noisy measurements of gaze and saliency, to provide a robust estimate of attention. The system quantitatively outperforms baseline attention estimators based either on gaze or saliency alone, as well as naive combinations of the two. In the following section we propose and discuss the framework of this model.

3. Proposed Model

Let G represent the Gaze Estimate, A represent the Focus of Attention Estimate, and T represent the ongoing task. G, A are drawn from continuous 2-D distributions in polar world coordinates (d, θ).

We can consider the joint probability distribution functions of gaze and attention, without assuming anything about the dependencies between them. As we have alluded to, the location of a subject's gaze may or may not indicate that attention is being paid to that particular location. Therefore by Bayes' rule for probability densities,

$$f_A(a) = \frac{f_A(a|G=g) f_G(g)}{f_G(g|A=a)}. \quad (1)$$

For the purposes of this paper T is a discrete random variable drawn from the space of all tasks, $T \in \{T_1, T_2, \ldots T_n\}$. By applying the law of total probability, conditioning on the task T,

$$\begin{aligned} f_A(a) &= \sum_{i=1}^{n} f_A(a|T=T_i) P(T_i) \\ &= \sum_{i=1}^{n} \frac{f_A(a|G=g, T=T_i) f_G(g|T=T_i) P(T_i)}{f_G(g|A=a, T=T_i)} \end{aligned} \quad (2)$$

We can consider each term in Equation 2 separately. First we consider the effect of the tasks on gaze and attention, in Section 3.1. Here we define the "prior" $P(T_i)$ as well as the space of "tasks" for this particular application context.

Each of the other terms are task-dependent; in other words for each task $T_i, i \in \{1...n\}$, there will be a different probability density for those terms. The first term, $f_G(g|T = T_i)$, corresponds to the probability density function of the gaze location. As described in Section 3.2 below, we can encode any uncertainty in the gaze estimate in this term.

We will consider the numerator term $f_A(a|G = g, T = T_i)$, as the likelihood that attention is focused on a particular location given gaze location and task, in Section 3.3. Finally, we will consider the distribution of gaze given attention and task, $f_G(g|A = a, T = T_i)$, in Section 3.4. This section will incorporate such concepts as "center bias" and "inhibition of return" which can affect the link between gaze and attention.

3.1. Influence of Tasks: $P(T_i)$

The particular task which the human is performing has a significant influence on attention [17]. When engaged in a task even in complex scenes and environments gaze follows a predictable pattern, focusing on the most pertinent subjects [9]. It is even evident that gaze can be used to predict whether a driver is distracted from any particular task [1].

Without loss of generality, for the purposes of this study we find it useful to define three "tasks," corresponding to highway driving; other environments may incorporate a different set of tasks into the same model. A primary task T_1 is defined as the task upon which the subject should be focused in most situations; in the case of driving that task involves maintaining the vehicle heading and speed within appropriate boundaries. A secondary task T_2 arises when the driver chooses to change lanes, including the setup to the maneuver and the maneuver itself. Finally, whenever the driver is distracted from either of these tasks, we label that state as T_3.

The prior probability $P(T_i)$ can be defined in a straightforward manner using a data-driven approach. Here for simplicity we assume it is uniform, however we acknowledge that in various environments, certain tasks are more likely than others. It is useful to treat the task as a nuisance variable, since it will be easier to analyze the rest of the equation assuming knowledge of the ongoing task. The direct estimation of task (*e.g.* [11]) could assist the model here by adjusting the prior, however we leave that beyond the scope of this work.

3.2. Gaze Estimation: $f_G(g)$

Monocular *eye gaze* estimation is a difficult problem, in light of occlusions, shadows, and other tough situations. Figure 2 demonstrates some of the tough conditions under which gaze estimation must occur in vehicle-based environments. Monocular *head pose* estimation can prove to be more robust, though less precise than eye gaze estimation. In either case, the gaze estimate will be a noisy approximation of the true gaze.

Figure 2. Sample images showing the difficulty involved in gaze estimation in complex environments such as vehicles.

In this study we use a commercially-available monocular head-pose estimation system. The system outputs the head yaw, synchronized with the rest of the system. For estimating attention in task-oriented behavior, this rough approximation of the true gaze may be sufficient. In primary tasks such as general scanning patterns, research has shown that gaze patterns tend to follow a "center bias" [20] whereby most glances occur toward the center of the field of view, in correspondence with the head pose. On the other hand for secondary tasks involving large visual searches, there has been recent support for the notion that head pose moves with eye gaze, making it easier to measure gaze using just head pose [3]. Nevertheless it is still important to account for gaze movements independent of head pose, especially in a "distracted" state as discussed below.

To deal with uncertainty in the gaze measurement, we take the output $\widehat{\theta}$ of a vision-based gaze tracker, and model uncertainty in the estimate with a 1-D Gaussian density function in θ. To be specific,

$$f_G(\theta) = \frac{1}{\sqrt{2\pi\sigma^2_{T_i}}} exp\left\{-\frac{(\theta - \widehat{\theta})^2}{2\sigma^2_{T_i}}\right\} \quad (3)$$

Note that we modify the variance of the Gaussian to be task dependent σ_{T_i}. The effect of task directly on the estimation of raw gaze is largely evident in the case of distractions. In those cases peripheral vision plays a larger role (motion and color cues affect periphery more). In the distracted state it then becomes necessary to increase the variance of the gaze estimate to include a wider field of view. In the primary and secondary-task oriented states, we can model the field of view after the typical human parafoveal field of view, which is approximately 10° [15].

3.3. Visual Saliency: $f_A(a|G = g, T = T_i)$

The term $f_A(a|G = g, T = T_i)$ in Equation 2 corresponds to the likelihood of paying attention to a particular location, given that the subject is looking at that direction and engaged in a given task. This most directly evokes the notion of visual saliency, that is, the regions of a scene that are most likely to draw attention. Given a gaze estimate in some direction, there are likely going to be locations in that field of view that are more salient than others. In this section we try to model those effects.

As discussed above, the derivation of visual saliency of a scene is a well-studied problem. Recent studies have demonstrated that task and context are the most significant factors in predicting gaze patterns and attention [17]. However the relationship is not clear, especially when subjects are cognitively or visually distracted. In driving situations, distraction is an extremely consequential issue [1]. There have been many attempts to model glance behavior based on interest point detection in a "bottom-up" approach [8]; here we can make use of the context of highway driving to develop more relevant saliency maps correlating to driver distraction [4].

In the case of the goal-oriented behavior in the lane-keeping and lane-changing tasks, we presume that the most salient regions in the scene, are those where the driver should be scanning in normal situations. Namely, we set the front of the vehicle as the most salient region in lane-keeping, and the sides of the vehicle as most salient under lane-changing. Saliency in these situations could be modified to include detection of objects relevant to the task as well, however we leave that beyond the scope of this work. These regions would correspond to "top-down" processing of saliency in a scene.

In the "No-task" or "Distracted" state, the drivers are more likely to be paying attention to abnormal events in the scene. In the case of driving on highways, motions and color changes tend to be most salient cues [7]; these are the cues most likely picked up by peripheral vision. Therefore we make use of a motion-based saliency map in this case, where normal "background" scene motions are calculated using optical flow, and then "foreground" motions are subtracted out [4].

$$f_A(d,\theta|g=\widehat{\theta},T=T_3)=\frac{m(d,\theta)-\overline{m}(d,\theta)}{\sum_{d,\theta} m(d,\theta)-\overline{m}(d,\theta)} \quad (4)$$

The current frame motion is m, and $\overline{m}$ represents an average "background" motion. Examples of this sort motion-based saliency can be seen in Figure 3.

3.4. Orienting of Attention: $f_G(g|A=a,T=T_i)$

The final term in Equation 2, $f_G(g|A=a,T=T_i)$, corresponds to the distribution of gaze given the location of attention. This term could encode any style of gaze inhibitor, or those cognitive processes which prevent gaze from moving to the focus of attention.

Of primary concern is "Inhibition of Return", the phenomena by which human subjects tend not to glance at objects which they have already seen [14]. This response could be due to the subject already having paid attention to those objects; by keeping track of them the subject does not need to look at that particular area again.

The task also plays a role in linking gaze to attention. In a goal-oriented state, a subject may be more focused on the task at hand. However in a distracted state, the subject may be more likely to be paying attention to something unrelated to the visual field of view. Such "cognitive distractions" have been found to significantly alter the safety of driving situations in particular [1].

In order to account for such notions, we define a task-oriented PDF, $f_G(g|A=a,T=\{T_1,T_2\})=C$, which amounts to a uniform distribution. That would indicate that no particular area is more likely to be glanced at. The distraction-driven PDF, $f_G(g|A=a,T=T_3)$ places slightly more emphasis on the gaze angle, which corresponds to a decreased likelihood of paying attention to the area where gaze is located. Such a model could be modified or trained to account for more subtle variations in these gaze inhibitors, given more detailed or informative training data.

4. Experimental Validation

In order to validate the model we take examples from real-world, naturalistic driving data. The vehicular testbed is outfitted with a commercial head tracking unit, which involves a monocular camera pointed at the driver. This system outputs head pose information; eye gaze trackers were implemented but deemed too unreliable in many situations.

Additionally, an omni-directional camera is mounted above the vehicle directly above the driver's seat. In this manner, the camera is able to see most of the driver's field of view. Sample images from this camera can be seen in Figure 3.

For every frame in the sequence, and every task, the terms in Equation 2 are computed and stored. The component of attention given each task is shown in columns 3-5 of Figure 3. The effects of the tasks on attention are clear; for the "lane keeping" task T_1 the model presumes that the attention is mostly on the forward areas. The attention given the "lane change" task T_2 are less focused on the front and more on the side.

However in the second sample row, the "distracted" state T_3 causes the attention to shift to the vehicle driving by. This effect is caused by the wide gaze estimate along with the motion-based saliency map. Assuming the driver is distracted, they would more likely notice the vehicle in their peripheral vision and then turn to pay attention to it. The vehicle is seen in both the tasks T_2 and T_3, so it would be a bit difficult to presume the actual underlying task.

By marginalizing the task, we arrive at a distribution function of the attention, $f_A(a)$, seen in the right-most column of Figure 3. The resulting estimate of attention can be obtained in a maximum likelihood manner, or the attentional distribution can be used as a prior for some temporal attention tracking scheme. In this setup we simply use

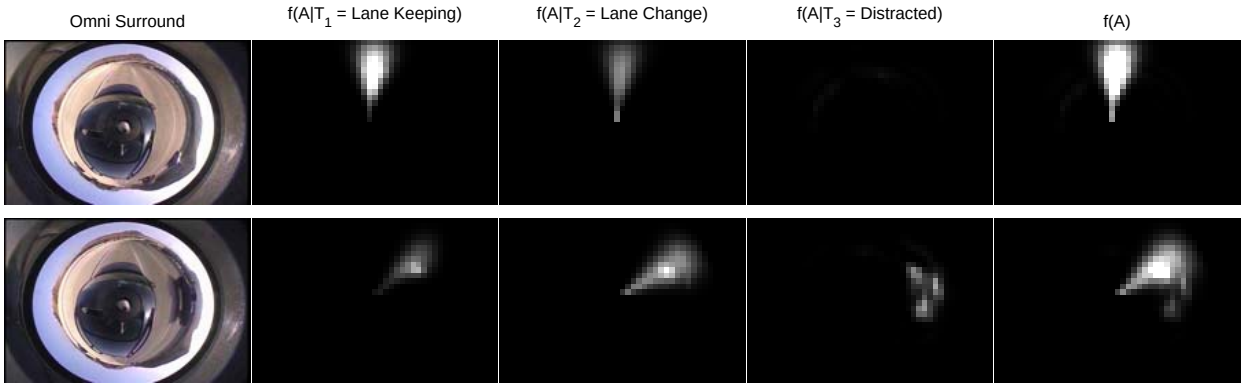

Figure 3. Sample data of the omni surround view, shown on the left. The task-dependent pdf's of attention corresponding to Equation 2 are in columns 3-5, and the last column shows the final density of the attention estimate. The top row corresponds to a normal lane-keeping state, and the bottom row shows data prior to a lane change.

$max(f_A(a))$ to obtain the final estimate.

4.1. Quantitative Evaluation

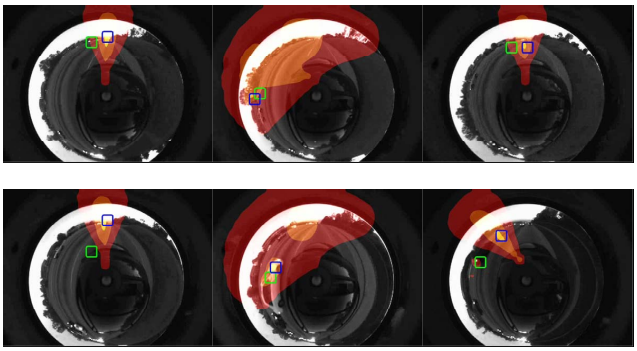

Figure 4. Results showing the estimated PDF of focus-of-attention. The blue box marks the estimated point of focus, and the green represents the labeled ground truth.

In this section we discuss the results of an experimental evaluation over several hundred frames of manually labeled data. In these cases the external environment images were shown to an annotator along with the images of the driver. Ground truth for the attentional location was determined from this manually labeled data. The following tables and graphs demonstrate the experimental accuracy of the proposed system.

The validation of the system should ideally include some notion of the ground truth of the driver's attention patterns. However it is difficult to obtain actual ground truth, without an extremely detailed understanding and measurement of the attentional processes in the driver's neurological system. Several simulation-based studies have included a questionnaire asking the driver to annotate their own data.

In this case we would like to emulate the performance of an expert assistance system; generally speaking a passenger sitting next to the driver has a good awareness of where the driver is paying attention. It has been shown that having a conversation with a passenger is much safer (than using a cell phone) because the passenger can modulate the conversation based on the driver's attentional and stress patterns [18]. An assistance system that approaches the performance of a passenger or human onlooker would be fundamentally useful. In this case we ask a human labeler to examine 461 frames of video of the driver and of the environment, acting as a passenger. Using those videos, the "passenger" then marks their interpretation of the focus of attention of the driver.

As points of comparison for the proposed Bayesian approach for estimating attention through simultaneous viewer and view observations (*BRAVVO*), we also include results of two simple attention estimators, one using only head pose (*HeadPose*) and one using only a raw saliency map (*Saliency*). The head pose-based attention estimator places the focus of attention on the central field of view of the driver's head. The saliency-based detector uses the motion-based saliency map derived above, and places the estimated attention at the point of greatest saliency.

Table 4.1 shows comparative results using several metrics. In the first metric, we correlate the labeled focus of attention with the PDF of the estimated attention. This sum demonstrates the ability of the estimation system to capture the most relevant points in the scene. The BRAVVO approach not only outperforms the baseline approaches, but it is also apparent that a simple combination of the two baseline approaches would still not reach the level of the proposed system.

We also calculate the error distance between the labeled and estimated focus of attention. Average results over the entire data set can be seen in the final column of Table 4.1. Figure 5 shows the performance improvement of BRAVVO over the baseline estimators, for a sequence of data.

It is interesting to note that in the cases when no head

Estimator	*Avg. Correlation*	*Avg. Error*
HeadPose-based	0.23	123.09
Saliency-based	0.18	183.80
BRAVVO (proposed)	**0.58**	**66.88**

Table 1. Comparison of proposed Bayesian-Attention Estimator with approaches based on Head Pose alone and Saliency alone. Avg. Correlation corresponds to the correlation between the estimated PDFs and the labeled Focus of Attention. Also shown is the error (in pixels) between the estimated and labeled Focus of Attention.

pose was detected due to the head being out of range of the head pose detection system (highlighted in red in Figure 5), the BRAVVO system still performs quite well. It is generally able to do so because it can revert to the saliency map as an estimate of what the driver should be looking at, while the head pose is out of range. Overall, the BRAVVO system clearly outperforms both baseline systems and more closely approaches the performance of an ideal agent.

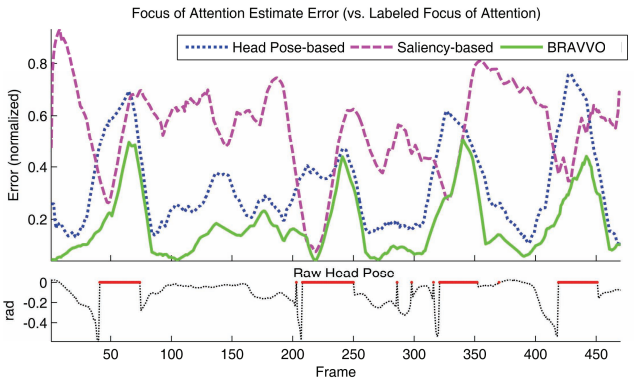

Figure 5. Improvement of Proposed BRAVVO approach compared to Baseline Head-Pose-based and Saliency-based Attention Estimators. The raw head pose is shown as a dotted line, with those areas of no head pose highlighted in red. Note that the Bayesian approach improves on both baseline estimators, with and without the presence of an accurate head pose reading.

5. Discussion and Future Directions

We have introduced a new approach to analyzing the attentive state of a human subject, given cameras focused on the subject and their environment. In particular, we are motivated by and focus on the task of analyzing the focus of attention of a human driver. We have developed a new Bayesian paradigm for estimating human attention specifically addressing the problems arising in complex, dynamic situations. The model incorporates vision-based gaze estimation and visual saliency maps, with cognitive considerations including "inhibition of return" and "center bias" that affect the relationship between attention and gaze.

The results demonstrate the potential of the model. We are able to demonstrate a 45.67% improvement in Focus-of-attention estimation over a baseline system based on head pose alone, and a 63.62% improvement over a saliency-based attention estimation system. These results show the capability of the system to more closely approximate an ideal assistive agent who is aware of the attentive state, based only on observations of the subject and surrounding environment. Further analysis and evaluation will require a careful design of experiments to measure the true attentive state of the subject. The model could be modified to use various kinds of saliency maps with other sets of cognitive factors. As seen in Figure 6, it is general enough to be potentially applicable to any human-machine interface in complex environments.

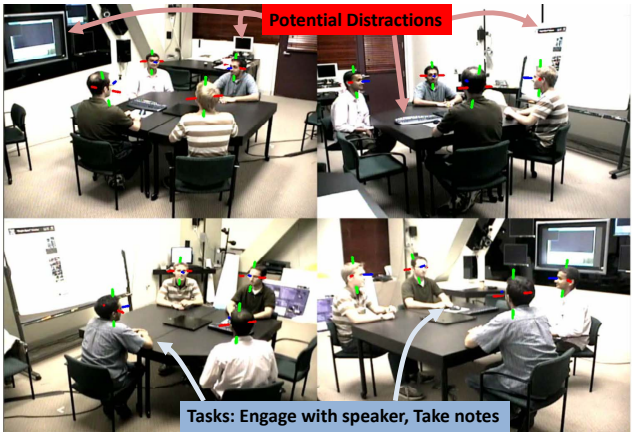

Figure 6. Meeting room environment, showing head pose estimation results from [12]. The proposed framework could generalize to estimate attention in environments such as meeting rooms where it may be important to differentiate tasks and distractions (such as people walking around, and various displays showing information in the background) while engaging and analyzing the behavior of participants in the meeting.

References

[1] L. Angell, J. Auflick, P. A. Austria, D. Kochhar, L. Tijerina, W. Biever, T. Diptiman, J. Hogsett, and S. Kiger. Driver workload metrics task 2 final report. *Report DOT HS 810635, NHTSA, U.S. Department of Transportation*, Nov 2006. 2, 3, 4

[2] S. Ba and J. M. Odobez. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics Part B*, 39(1), 2009. 2

[3] A. Doshi and M. M. Trivedi. Head and gaze dynamics in visual attention and context learning. *CVPR Workshop on Visual and Contextual Learning (VCL)*, 2009. 3

[4] A. Doshi and M. M. Trivedi. Investigating the relationships between gaze patterns, dynamic vehicle surround analysis, and driver intentions. *IEEE Intelligent Vehicles Symposium*, June 2009. 4

[5] H. V. Helmholtz and J. P. C. Southall. Helmholtz's treatise on psychological optics, 3. *The Optical Society of America, Rochester, NY*, 1924. 1

[6] J. E. Hoffman. Visual attention and eye movements. In H. Pashler, editor, *Attention*, chapter 3. Psychology Press, 1998. 1

[7] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, 2005. 4

[8] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000. 1, 2, 4

[9] M. F. Land. Predictable eye-head coordination during driving. *Nature*, 359:318–320, September 1992. 2, 3

[10] Y. C. Lee and J. D. L. amd L. N. Boyle. Visual attention in driving: the effects of cognitive load and visual disruption. *Human Factors*, 49(4):721–733, 2007. 2

[11] J. C. McCall, D. Wipf, M. M. Trivedi, and B. Rao. Lane change intent analysis using robust operators and sparse bayesian learning. *IEEE Transactions on Intelligent Transportation Systems*, Sept. 2007. 3

[12] E. Murphy-Chutorian and M. M. Trivedi. 3D tracking and dynamic analysis of human head movements and attentional targets. *IEEE/ACM International Conference on Distributed Smart Cameras*, June 2008. 2, 6

[13] R. J. Peters and L. Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 2

[14] M. I. Posner and Y. Cohen. Components of visual orienting. *Chapter in Attention and Performance X, Bouma H. and Bouwhuis D., eds*, pages 531–556, 1984. 2, 4

[15] M. Rizzo and I. L. Kellison. Eyes, Brains, and Autos. *Arch Ophthalmol*, 122:641–645, Apr. 2004. 3

[16] A. L. Rothenstein and J. K. Tsotsos. Attention links sensing to recognition. *Image and Vision Computing*, 26:114–126, 2008. 1, 2

[17] C. A. Rothkopf, D. Ballard, and M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14):1–20, 2007. 2, 3, 4

[18] D. L. Strayer and F. A. Davis. Cell-phone-induced driver distraction. *Current Directions in Psychological Science*, 16(3):128–131, 2007. 5

[19] S. Yantis. Control of visual attention. *in Attention, ed. H. Pashler. 13-74 Psychology Press, Hove, UK.*, 1998. 2

[20] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):1–20, 2008. 3