Inferring Shared Attention in Social Scene Videos

Lifeng Fan*, Yixin Chen*, Ping Wei, Wenguan Wang, Song-Chun Zhu

University of California, Los Angeles

1. Introduction

Shared attention is the attention focus shared by two or more people on one object or human. Human communication is only possible when the people involved in such communications have built a common conceptual ground consisting of shared attention, shared experience, common cultural knowledge, *etc*. [7]. Shared attention is a crucial first step towards social interaction, as well as the primary basis of social intelligence and a precursor of theory of mind [3]. Humans can easily recognize and form shared attention in a social group. Patients with autism feel it difficult to interact with people around due to the lack of ability to build shared attention with others [1].

The study of shared attention is important because it helps a computer vision system to better understand and interpret human activities in images or videos. Robotics equipped with the ability to detect and understand human shared attention can also be more intelligent when interacting with humans. However, works on shared attention are quite limited in the computer vision community at present.

Given a third-person social scene video clip, we want to detect which frames contain shared attention and where is the shared attention in those frames. We collect a new dataset VideoCoAtt[1], which covers diverse social scenes with full annotations. We also build a deep spatial-temporal neural network with four modules: gaze estimation module, region proposal module, spatial detection module and temporal optimization module. The proposed model explicitly leverages human gaze direction, target region candidates, and temporal inter-frame constraints for identifying shared attention.

2. VideoCoAtt Dataset

In this section we describe our proposed VideoCoAtt dataset, which is specifically designed for studying shared attention in social scenes. Some example frames with annotations are presented in Fig. 1.

[1]This dataset is available at: http://www.stat.ucla.edu/~lifengfan/shared_attention.

Dataset Collection. The following principles drive the collection of our dataset: 1) *Natural social interaction.* TV show is a good choice because social interactions in TV shows appear to be relatively more natural. The videos are sourced from 20 different TV shows on Youtube. 2) *Large scale and high quality.* We collect 380 RGB video sequences, each of which lasts for various time, from around 20s to more than 1 minute with a frame rate of 25 fps. In total, there are 492,100 frames at the spatial resolution of 320×480. 3) *Diversity and generality.* The videos in the VideoCoAtt dataset cover different countries and cultures, such as American, Chinese, Indian, European, *etc*. The appearances of actors/actresses, the costume and props vary a lot. There are also diverse scenario settings in VideoCoAtt, including living room, kitchen, restaurant, Cafe, office, outdoor, *etc*. Moreover, the number of shared attentions per frame and the number of involved people per shared attention can vary in different frames and videos. This generality in VideoCoAtt dataset is beneficial for the trained model to deal with multiple cases as in real life.

Dataset Annotation. We manually annotate all the video frames using the online tool Vatic [8]. For each frame, we mark whether there is shared attention in the scene. If there is on-going shared attention in the scene, we mark all the shared attentions with bounding boxes. Only those shared attentions within the view of the scene will be annotated; those out of view or occluded will not be counted as shared attention. Furthermore, for each shared attention, we annotated all the heads that are currently engaged in the certain shared attention using bounding boxes and attributes related to the shared attention numbering.

Dataset Splitting. We split our VideoCoAtt dataset into three parts for training, validation and testing respectively. There are 181 videos (250,030 frames) in the training set, 90 videos (128,260 frames) in the validation set and 109 videos (113,810 frames) in the testing set. To avoid overfitting caused by similarities in human appearances and scenario settings, we split our videos by different sources, which we believe is necessary and will require a strong generalization ability of our shared attention model.

Figure 1. **Example frames from VideoCoAtt Dataset,** where the shared attentions are annotated by red rectangles and red points. Different groups of people involved in different shared attentions are annotated by rectangles in different colors. Best viewed in color.

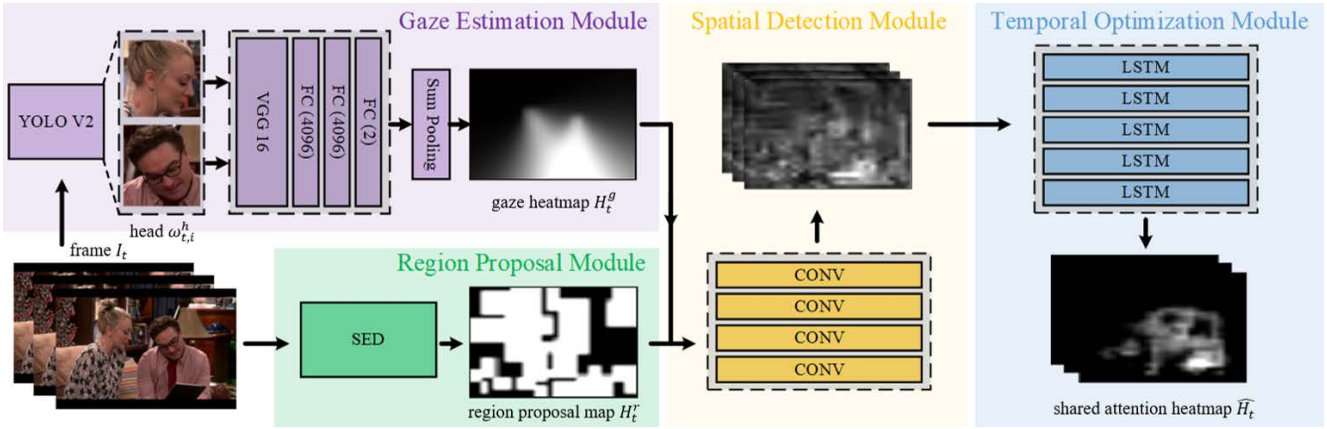

Figure 2. Illustration of our model architecture consisting of four modules.

3. Our Model

Our shared attention detection model comprises of four modules: 1) **Gaze Estimation Module.** For an input frame I_t, we use retrained YOLO V2 to detect all the head location $q_{t,i}$. The corresponding closeup image patch for head location $q_{t,i}$ is cropped out from I_t and denoted as $w_{t,i}^h$. A modified VGG 16 is then applied to regress a gaze direction $d_{t,i} \in [-1,1]^2$ for $w_{t,i}^h$. We use a Gaussian distribution to model the variation of a gaze ray with respect to the predicted primary gaze direction $d_{t,i}$ and generate a gaze heatmap $H_{t,i}^g$ for each detected head $q_{t,i}$. Then a final gaze heatmap H_t^g for frame I_t is generated via sum-pooling all the prepared $H_{t,i}^g$. 2) **Region Proposal Module.** To exploit context information, we use a region proposal module to generate a binary region proposal map H_t^r for input image I_t, which is implemented by Structured Edge Detector (SED) [9]. 3) **Spatial Detection Module.** Shared attention detection is firstly conducted in a frame-by-frame style. We apply a spatial detection module consisting of several convolutional layers to combine the gaze heatmap H_t^g and region proposal map H_t^r for intra-frame shared attention detection and output the intermediate shared attention heatmap $\tilde{H}_t$. 4) **Temporal Optimization Module.** To further exploit the temporal inter-frame constraints in videos, we add a temporal optimization module $LSTM(\cdot)$ that consists of several convolutional Long Short-Term Memory (convLSTM) network [6] layers to get the optimized eventual shared attention heatmap $\hat{H}_t$. An illustration of our whole model architecture is presented in Fig. 2.

We apply the Mean Squared Error (MSE) between the predicted shared attention heatmap $\hat{H}_t$ and the ground truth shared attention binary map H_t as the loss function. The inference is possible given the predicted shared attention heatmap $\hat{H}_t$, based on which we can compute the cumulative score for each region proposal bounding box $b_{t,i}$. We only keep those proposal bounding boxes with a score higher than a threshold. Then we conduct a Non-Maximum Suppression (NMS) [2] and treat the remaining bounding boxes as our final shared attention prediction for frame I_t.

Model	Prediction Acc.	L^2 Dist.
Raw Img.	52.3 %	188
Only Gaze	64.0 %	108
Only RP	58.0 %	110
Gaze+RP	68.5 %	74
Gaze+RP+Img.	54.0 %	72
Fixed Bias	52.4 %	122
Random	50.8 %	286
Gaze Follow [5]	58.7 %	102
Gaze+Saliency[4]	59.4 %	83
Gaze+Saliency[4]+LSTM	66.2 %	71
Ours (Gaze+RP+LSTM)	**71.4 %**	**62**

Table 1. Quantitative evaluation results with Prediction Accuracy and L_2 Distance.

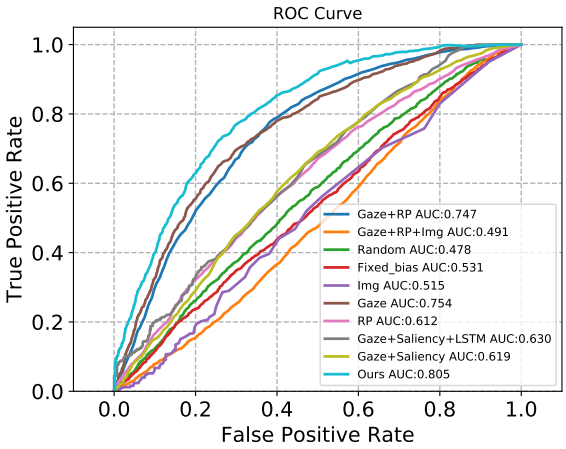

Figure 3. Quantitative evaluation results with ROC Curve, computed over the test set of the VideoCoAtt dataset.

4. Experiments

4.1. Experimental Setup

Evaluation Metrics. The percentage of frames with right shared attention existence prediction over all the video frames is applied as a metric ***Prediction Accuracy***. We also use the region proposal bounding boxes and shared attention heatmap to generate a ***ROC Curve***, reflecting the precision and recall when predicting shared attention bounding boxes under different score thresholds. ***AUC*** refers to the area under the ROC curve (higher is better). Then given a certain score threshold, the L^2 ***Distance*** (measured in pixel) is the Euclidean distance between the predicted shared attention bbx and the annotated ground truth.

Baseline Methods. ***Random***: A Gaussian heatmap with random mean and variance. ***Fixed Bias***: A Gaussian heatmap with mean and variance learned from our dataset. ***Gaze Follow***: We apply the gaze following model in [5] to detect all the people's gaze fixations and gaze concurrences in a frame as a baseline. ***Gaze+Saliency*** and ***Gaze+Saliency +LSTM***: We replace our region proposal module with a top-performance saliency model [4], and consider two baselines with and without the temporal optimization module respectively.

Ablation Study. ***Raw Img.***: Only raw image as input to the spatial detection module. ***Only Gaze***: Gaze estimation and spatial detection module. ***Only RP***: Region proposal and spatial detection module. ***Gaze+RP***: Gaze estimation, region proposal and spatial detection module. ***Gaze+RP+Img.***: Gaze, region proposal and raw image feature as input to spatial detection module.

4.2. Results and Analysis

Quantitative results. Table 1 shows the comparison of our model with baseline methods and several ablation models by two evaluation metrics *Prediction Accuracy* and L^2 *Distance*. Our model achieves the best performance in both the shared attention interval detection task (Prediction Acc.: 71.4%) and the shared attention location prediction task (L^2 Dist.: 62).

Among all the baseline models, the second best model is *Gaze+Saliency+LSTM* with a Prediction Acc. of 66.2% and a L^2 Dist. of 71. The replacement of region proposal module with a saliency model impairs our model performance because the shared attention of people in a social interaction may not be the most visually salient object in the scene, but more influenced by the on-going interaction. The performance of the Gaze Follow baseline in detecting shared attention is mediocre, which is mainly because that shared attention of a social group is goal-driven and object-related, not just the concurrence of human gazes.

Among all the ablation models, *Gaze+RP* shows a overall best performance (Prediction Acc.: 68.5% and L^2 Dist.: 74), but is still inferior to our full model with all the four modules. And overall *Only Gaze* performs better than *Only RP*, indicating the gaze estimation module plays a more important role than the region proposal module in shared attention detection, which is consistent with our intuitions. The simplest model without any module design *Raw Img.* performs worst. The ablation study shows that each of the four modules proposed by our model (§ 3) is important and necessary for shared attention detection in videos.

Fig. 3 shows the ROC Curve and AUC comparison results among our full model, baseline models and ablation models. Our model has the best precision and recall performance and the largest AUC value than all the other models.

Qualitative results. Fig. 4 exhibits an internal visualization of shared attention detection results by our full model on some example frames. The *Gaze Heatmap* roughly features the attention of each individual in the social scene and is not enough to accurately feature shared attention. The *Region Proposal Map* gives some potential shared attention proposals and provides the important spatial constraints. *Single-frame Detection* combines the *Gaze Heatmap* and

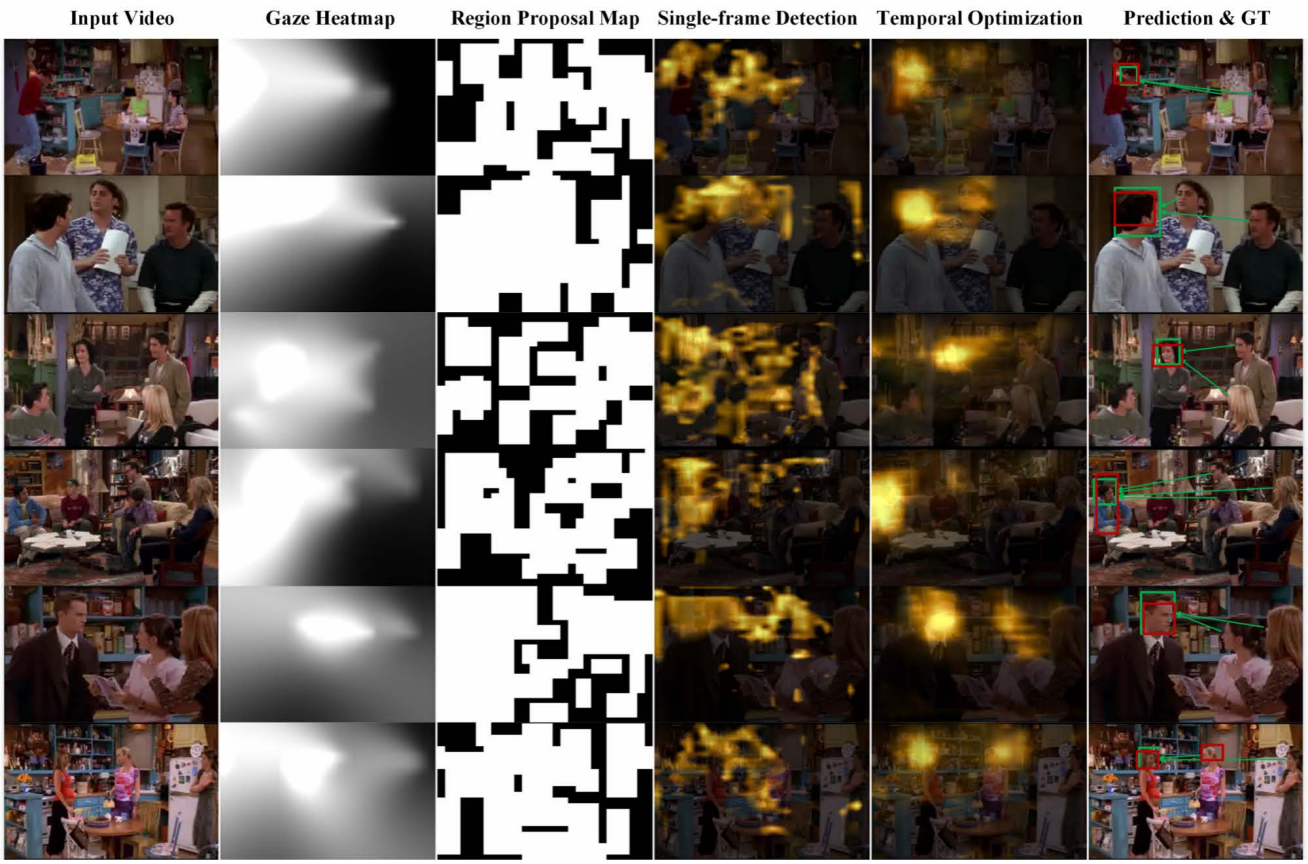

Figure 4. Shared attention detection results on example frames.

the *Region Proposal Map* to generate a preliminary shared attention heatmap, which still has too much noises. After the *Temporal Optimization* by convLSTM, the shared attention heatmap is much clearer and can provide more accurate shared attention distribution information. The final column in Fig. 4 compares our eventual shared attention prediction results (depicted in red rectangles) with the ground truth shared attention annotations (depicted in green rectangles). As shown, there are good predictions that can exactly locate the shared attention in the social scenes, like the prediction in the first example. However, there are also some false alarms existing. For example, The scene in the last row actually has only one shared attention, but our model gives two predictions located near the two human faces. This is an interesting failure example since whether the third person on the right side is looking at the person on the left side or the person in the middle is somehow ambiguous for our model to distinguish. That's why the shared attention heatmap gets two peaks for this example. But similar situation in the fifth scene is successfully solved by our model.

Acknowledgement. This work was supported by ONR MURI project N00014-16-1-2007, DARPA XAI Award N66001-17-2-4029, and NSF IIS 1423305.

References

[1] S. Baron-Cohen. *Mindblindness: An Essay on Autism and Theory of Mind.* MIT Press, 1995. 1

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE TPAMI*, 32(9):1627–1645, 2010. 2

[3] C. Moore and P. J. Dunham. *Joint Attention: Its Origins and Role in Development.* Lawrence Erlbaum Associates, 1995. 1

[4] J. Pan, E. Sayrol, X. Giro-I-Nieto, K. McGuinness, and N. E. OConnor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, 2016. 3

[5] A. Recasens, A. Khosla, C. Vondrick, and A. Torralba. Where are they looking? In *NIPS*, 2015. 3

[6] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In *NIPS*. 2015. 2

[7] M. Tomasello. *Origins of Human Communication*. The MIT Press, 2008. 1

[8] C. Vondrick, D. Patterson, and D. Ramanan. Efficiently scaling up crowdsourced video annotation. *IJCV*, 101(1):184–204, 2013. 1

[9] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *ECCV*, 2014. 2