3rd International workshop on Gaze Estimation and Prediction in the Wild, at CVPR 2021

Visual Focus of Attention Estimation in 3D Scene with an Arbitrary Number of Targets

Rémy Siegfried and Jean-Marc Odobez

Idiap Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne, Switzerland
remy.siegfried@idiap.ch, odobez@idiap.ch

Abstract

Visual Focus of Attention (VFOA) estimation in conversation is challenging as it relies on difficult to estimate information (gaze) combined with scene features like target positions and other contextual information (speaking status) allowing to disambiguate situations. Previous VFOA models fusing all these features are usually trained for a specific setup and using a fixed number of interacting people, and should be retrained to be applied to another one, which limits their usability. To address these limitations, we propose a novel deep learning method that encodes all input features as a fixed number of 2D maps, which makes the input more naturally processed by a convolutional neural network, provides scene normalization, and allows to consider an arbitrary number of targets. Experiments performed on two publicly available datasets demonstrate that the proposed method can be trained in a cross-dataset fashion without loss in VFOA accuracy compared to intra-dataset training.

1. Introduction

Estimating the visual focus of attention (VFOA) of people from videos is an important problem in interaction analysis, with application in many fields like multiparty conversation [10], human-computer/-robot interaction [1], and psychological studies [15] among others. Formally, VFOA can be defined as the target that a person is looking at and is in that sense a discretization of the continuous gaze direction.

VFOA estimation is a challenging task as it relies on many features like the head pose and gaze of the person, and also scene information to know the positions of the person and VFOA targets or other potential contextual information (e.g. who is speaking), which can be difficult to estimate accurately depending on the setup. In particular, despite recent progress, the accuracy of gaze estimation is often limited when recording naturally acting people with remote sensors due to low image resolution and high variability in appearance (pose, eye), in particular for large head poses.

When handling 3D scenes, one simple way to estimate the VFOA consists of comparing the gaze direction and target positions using an angular distance measure. This method is always applicable but neglects the use of scene cues like conversation regime [9] (which encodes the prior that people tend to look at their interlocutor or the speaker) or task context [20] that were shown to improve VFOA estimation accuracy and counterbalance ambiguities due to close VFOA targets and noisy gaze estimates. Principled methods have been proposed to integrate such cues in the VFOA inference, like bayesian [3, 13], random forest [5] and deep learning [17] schemes, but they usually rely on predefined target sets leading to fixed input/output sizes and inference structures. Thus, these models are usually trained for a specific setup and should be retrained to be applied to another one, which limits their usability as VFOA annotations can be difficult or expensive to gather. Having a model that can generalize to new setups and situations (geometry, number of people, and salient objects) would be useful.

To work towards this goal, we propose to reformulate the problem. The main idea is to reformat the input features into several 2D maps associated with the subject's field of view, allowing to encode all inputs and contextual cues (4 maps for head pose, gaze direction, directions of speaking sources, and person speaking status) within a single referential, as well as providing as input a visual saliency 2D map encoding an arbitrary number of candidate VFOA targets. This leads to a fixed number of maps that can be stacked as a tensor and processed as an image to produce a 2D map of

VFOA direction probabilities, whose maximum is used in the final VFOA classification (details in Sec. 3). This has four advantages:

- it normalizes the inputs, removing spatial and feature dependencies to the camera view points and more generally the setup;
- it allows to consider an arbitrary number of targets and to encode all contextual cues in the same referential;
- it makes the input more suited to be processed by convolutional neural networks (CNN), as images naturally encode proximity in space and channels;
- it makes data augmentation easier (an important step for generilizing to other situations), as targets can easily be be added and removed or context be modified during training.

Our contribution is thus a novel method to estimate the VFOA of a subject given an arbitrary number of targets and contextual cues, which combines the above advantages and allows the application of a learned model to different setups. We evaluated this method and the impact of the different input features on two publicly available datasets, namely UBImpressed [15] and KTH-Idiap Group-Interviewing [16], including convincing cross-datasets/setup experiments.

2. Related works

Back when accurate gaze trackers were not available, VFOA was inferred from head pose using behavioral models [22]. Inference mechanisms like GMM, HMM, or Dynamical Bayesian Networks [18, 3] were used to estimate the VFOA directly from the head pose, potentially taking into account context information [9, 20] or modeling the joint VFOA of all participants [4, 13]. Nevertheless, with the recent improvement of gaze estimation, even simple frame-based geometrical models were shown efficient to estimate VFOA [24]. However, remote gaze estimation is often noisy and unreliable in extreme head pose cases, which limits the accuracy of such methods. To address this problem, as with earlier methods, the gaze direction is usually combined with other cues like the head pose [2] or the participants' speaking status [21, 14]. On the methodological side, recent works rely more on deep learning, as it provides more efficient ways to fuse different cues and learn more complex inter-modality and cue relations. In this context, temporal neural networks (CNN, RNN) were shown more accurate than their Bayesian counterparts [17].

However, although VFOA models were improved using new methods and additional context features, with the aim of modeling conversations, little work was done to improve their flexibility. Indeed, conversation models have the advantage of representing all participants' behaviours together to take into account dependencies between them and they were shown to be effective at learning models for VFOA estimation in setups for which annotations are available. However, such models are often trained on a single specific setup, in which a defined number of static visual targets are usually assumed and 3D scene representation is sometimes absent, leading the model to learn feature clustering rather than geometrical reasoning, so it can not generalize to unseen setups with a different number of people or a different geometry and can not handle people that join or leave the conversation. [12] addresses this issue by learning two-person VFOA predictors (do they look at each other or not), but their work addresses 2D VFOA estimation in images which is not multimodal and different from VFOA estimation in 3D scenes, as it relies on image processing rather than on geometrical reasoning. Here we propose a new input data format that is more naturally processed by CNNs and allows the same trained model to handle an arbitrary number of targets as well as different setups. In this regard, unlike the above methods, the proposed method estimates the VFOA of each person individually (rather than jointly) using 2D maps while still encoding the conversation features from all subject features as well as scene information.

Representing the gaze direction as an image was already studied in the context of gaze refinement in a screen-based setting [19], where it was used to align the point of gaze with the screen image content. In this work, we extend this idea by representing all subject's and scene's features as images (i.e. 2D maps) and by considering the whole 3D field of view of the subject. In contrast to [19], we do not have access to what the subject sees or the gaze ground truth, so we must work in a virtual field of view that we fill with the scene information. Also, we must handle aversion cases without knowing where the subject is effectively looking.

3. Method

The proposed method is presented in Fig. 1. It can be divided into three main parts:

1. the extraction of the required features (head pose, gaze direction, target directions, and speaking status) from the input video and scene information;
2. the translation of these features into a fixed number of 2D maps;
3. the estimation of the VFOA from the 2D maps.

They are described in more detail below.

3.1. Features extraction

We consider the case where the scene is monitored, i.e. that the 3D positions and the speaking status (binary) of

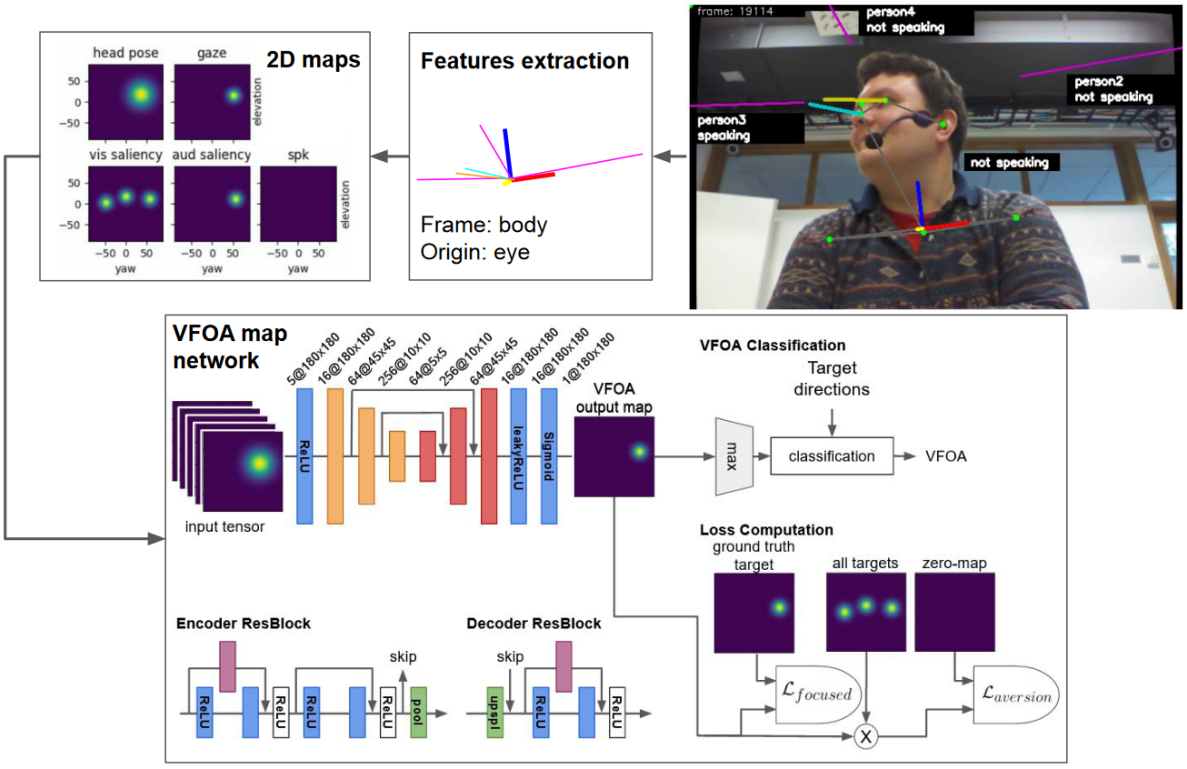

Figure 1. Proposed workflow. First, head pose (cyan) and gaze direction (yellow) are estimated from the RGB-D video while target directions (purple) and people speaking status are recovered from scene monitoring. Second, all these features are expressed in the body frame (red-blue-yellow) and represented as yaw and elevation angles which are then further encoded into five 2D maps. The resulting input tensor is processed by a CNN to generate a VFOA probability map whose maximum is used to derive the actual VFOA. Network layer types are represented by colors: 3x3 conv. layers (blue, with mentioned activation), encoder residual blocks (orange), decoder residual blocks (red), 1x1 conv. (purple), activation alone (white), and up-/downsampling (green). Elements for computing the loss are indicated at the bottom right.

each person involved in the interaction are available. Our goal is to express all features (head pose, gaze, and target directions) as yaw and elevation angles in a frame associated with the body orientation. In this way, the representation can potentially exploit coordination patterns between the body, the head pose, and the gaze, and allows to normalize the data independently of the camera position. Target directions are easily computed as the eye-to-target (most forward of both eyes) vector and then translated to angles, while head pose and gaze direction are extracted from RGB-D video capturing the subject, as summarized below.

Body frame estimation. The 3D positions of the subject's joints are extracted by combining the body 2D keypoints from OpenPose [6] and the depth provided by RGB-D cameras. To catch the orientation of the subject, the body frame is built using the vector going from the right to the left shoulder and the vertical axis of the camera frame. The latter axis was selected since available videos only provide upper body views of the people, so the hip keypoints are not available and the estimation of the vertical axis of the body is then difficult. We consider it a minor drawback, as most of the body rotations are done in the yaw direction in our conversation scenarios.

Head pose. It is estimated from the subject's video using the Headfusion method [23], which processes RGBD videos and provides both the 3D position (define as the nose's tip) and the orientation of the subject's head. This method relies on the online fitting and tracking of both a 3D Morphable Face Model and a 3D raw representation of the head. The use of depth information and head reconstruction makes the method much more robust to large head poses variations, compared to 2D landmarks based methods. Finally, the head pose angles are expressed in the body frame, so the head yaw angle is 0 when the head is in a neutral position, even if the subject is not facing the camera.

Gaze estimation. Using the head pose, the 3D textured mesh obtained from RGB and depth image is rotated to get a frontal image of the face [8]. Then, a facial landmark detector [11] is applied to this frontal image to locate the position of the eye corners and crop the eye images (36x60 pixels). This method normalizes the size and appearance of the eye

images. Then, to estimate the gaze direction, we used the GazeNet [25] architecture trained on the Eyediap dataset [7] (floating target, mobile pose), as this dataset provides depth and thus allows the same eye image normalization method as above. We obtained state-of-the-art results of 6.3° mean angular error on the Eyediap test set.

3.2. 2D feature maps

The proposed method takes five types of features as input: head pose angles, gaze direction angles, speaking status of the subject, the directions of potential VFOA targets (i.e. visual saliency), and the directions associated with speaking targets (i.e. audio saliency). To bring all these features in the same space and allow an arbitrary number of targets to be represented, we set in place two main elements. First, as explained earlier, all directions, whether from the scene (target directions) or from the human subject (i.e. head pose, gaze) have been expressed in the same reference frame (the subject's body frame). Secondly, each input feature type is encoded as a 2D map with a resolution of 180x180 pixels, in which each pixel represents an angle of 1 degree in both yaw and elevation axes. As a result, the 2D maps represent a unified view of the gazing activity and scene information in front of the person.

To generate these maps (see one example of such maps in the top left of Fig. 1), we proceed as follows. If $\{p_i, i = 1, \ldots, N\}$ denotes the set of directions to be encoded in the map D, we simply place 2D gaussians of covariance Σ^m at each direction p_i to provide information regarding this direction while taking into account the estimation noise or the size of targets. More formally:

$$D(p) \propto \max_{i=1,\ldots,N} \mathcal{N}(p - p_i; \Sigma^m) \tag{1}$$

Using this process, the head pose map is created by using as p_i the (single) head pose direction, the gaze map is built using the gaze direction, the video saliency map is produced using as p_i the set potential VFOA target directions (people in the conversation in our case), and the audio saliency map using the directions of people who are talking. The map associated with the speaking status is the exception. As it is not associated with any direction, we chose to fill it with its value, i.e. it is full of ones when the subject speaks and full of zeros otherwise.

Finally, all the five above maps are gathered into a single tensor, so that it can be processed as a 5-channel input by a convolutional neural network.

3.3. VFOA network and classification

Architecture. It is similar to the one in [19] and is a kind of hourglass network with one initial 3x3 convolutional layer, 3 down- and upsampling layers built from residual blocks (two blocks per encoder, one per decoder, as shown in at the bottom left of Figure 1), and two final 3x3 convolutional layers ending up with sigmoid activation. Down- and upsampling are done using maxpool and upsample layers respectively.

VFOA classification. The predicted VFOA map is transformed into yaw and elevation angles by taking the angle coordinates of the map's maximal value. Then, VFOA classification is performed using the angular distance (i.e. cosine distance) to each target: it is an aversion if the minimal angular distance to all targets is above a given threshold τ_{vfoa} and the nearest target otherwise.

Training. The VFOA map network is trained frame by frame using the binary cross-entropy (BCE) loss distinguishing two cases, as shown in the bottom right of Fig. 1. When the subject is looking at another person (focused), we want the output VFOA map to fit the target position, and the loss is the BCE between the output map and a target map featuring only the ground truth VFOA target. In case of aversion, we want the output VFOA map values to be low where there are targets, so the loss is the BCE between a zero map and the output map masked by the visual saliency map to only keep VFOA outputs close from targets (and thus remove outputs far from targets which can be considered as valid outputs). So, we have:

$$\mathcal{L}_{VFOAmap} = f \cdot \mathcal{L}_{focused} + (1 - f) \cdot \mathcal{L}_{aversion}, \tag{2}$$

$$\mathcal{L}_{focused} = \textbf{BCE}(M_{out}, M_{tar}), \tag{3}$$

$$\mathcal{L}_{aversion} = \textbf{BCE}(\frac{M_{vsal}}{max(M_{vsal})} \cdot M_{out}, 0_{map}), \tag{4}$$

where M_{out}, M_{tar}, M_{vsal}, and f are respectively the output map, the target map, the visual saliency map (consisting of all potential targets), and a binary indicator equal to 0 if the ground truth VFOA is "aversion" and 1 otherwise.

Data augmentation. To increase the generalization abilities of the trained model, we use several data augmentation strategies:

- target removal: a random number of targets (between 0 and the total number of targets minus 1) are removed from visual and audio saliency maps. If a removed target corresponds to the VFOA ground truth, the label is turned to "aversion";
- target addition: a random number of fake targets (between 0 and 2) are added to the visual saliency map. Their locations are sampled using the mean of the real target positions and a variance scaled by 1.5 in the yaw direction. Each fake target can also appear on the audio saliency map, as if it was speaking, with a probability of 0.5;
- global noise: random white noise ($\sigma = 5°$) is added to all angles (head pose, gaze, and target positions).

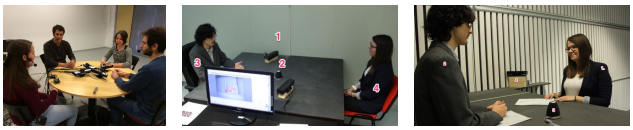

Figure 2. Picture of the KTH-Idiap (left), UBImpressed 'Interviews' (center) and UBImpressed 'Desk' (right) setups.

- feature noise: random white noise ($\sigma = 2°$) is added to each angle separately (head pose, gaze, and target positions).

When data augmentation is used, these four strategies are applied to each samples, meaning that the number of training samples does not increase.

4. Experiments

4.1. Datasets

In this work, we rely on two conversation datasets, which present different setups (see Fig. 2) in terms of scenario, length, and number of participants, but provide the same kind of data (RGB-D recording and speaking turns of each person).

UBImpressed dataset [15]. It consists of short dyadic interactions (five to ten minutes) in which a participant interacts with an actor in two different scenarios: a job interview, in which the two persons are sitting in front of each other in a formal setup, and a reception desk, where they are standing and moving freely. Data were acquired with Kinect2 sensors (RGB-D, HD color images, 30 fps) placed on the table, recording each participant individually with a sideway point of view. Also, a microphone array recorded the conversation and indicated who is talking in each video frame. VFOA was annotated on the first minute plus 5 additional segments of 10 seconds every minute in 4 'Interviews' and 4 'Desk' sessions (so 16 videos in total). Removing samples where the subject is blinking or when the annotation is uncertain, sums up to around 36K annotated samples, with 39% of aversions.

KTH-Idiap dataset [16]. It consists of one-hour four-party meetings in which three students present their projects to an interviewer who leads the discussion. Compared to UBImpressed, this setup presents a more relaxed and more dynamic type of social interaction, with alternation of monologue, dialogue and discussions.

Data were acquired with Kinect1 sensors (RGB-D, VGA color images, 30 fps) placed on the table at around 0.8 meters in front of each participant, and lapel microphones indicating who is talking in each video frame. VFOA was annotated on the first minute of interaction and 9 additional segments of 30 seconds spread on the entire video in all 5 sessions (so 20 videos in total). In this dataset, annotated samples sum up to around 115K, with 17% of aversions.

Scene monitoring. As specified in Section 3.1, both datasets provide the speaking status of each participant at each video frame. They also provide the recording of each participant and the relative position of their corresponding cameras, so that the target (i.e. people) positions can be extracted by processing each target person video using the Headfusion method (see Section 3.1) and by projecting their 3D positions in the subject's camera coordinate system, before expressing them in the subject's body frame.

4.2. Baseline model

We looked for a baseline that is comparable to our method in terms of input features and usability, i.e. a method allowing to predict the VFOA for an arbitrary number of targets without retraining. However, to the best of our knowledge, previous state-of-the-art methods focused on predicting VFOA in scenarios involving a fixed number of people in the conversation and within a fixed setting [4, 17, 5]. No cross-dataset experiments were performed. As discussed in Section 2, applying these methods to a new setting with a different geometry or number of participants is not trivial without retraining the model from scratch, since these models do not introduce explicit spatial relationships, and they do not handle moving people.

Thus, we propose to use as baseline a strong multimodal statistical binary classifier predicting the probability that a subject looks to a target given the target direction, the subject's gaze and head pose, and the speaking status of all persons in the scene. Doing so allows this classifier to be applied to any number of potentially moving targets without retraining or fine-tuning and uses the same features as the proposed method.

More formally, let us denote F_i the subject's focus status toward the target i (1 when being *focused* on target, and 0 otherwise) g his/her gaze direction, h his/her head pose, t_i the direction of target i, and S the speaking status of the scene, defined as the combination of three speaking status $S = (S_{subject}, S_{target}, S_{other})$ and can thus take 8 values. For example, it can take values such has $(0, 0, 0)$ (nobody speaks) or $(0, 1, 1)$ i.e. the subject does not speak, but the target and at least another person speak. Note that g, h and t_i are all expressed as 2D angles. With these notation, we define the probability of the subject' focus status as follows:

$$p(F_i|Z_i, S) \propto p(Z_i|F_i)p(F_i|S), \tag{5}$$

$$\text{with } Z_i = [g - t_i, h - t_i]^T \tag{6}$$

where we made the assumption that the gaze and head pose errors $g - t_i$ and $h - t_i$ do not depend on the speaking status. We then further define the likelihood $p(Z_i|F_i)$ as a multivariate Gaussian for each possible value of F_i. It can

formally be written as:

$$p(Z_i|F_i) = F_i\,\mathcal{N}(0, \Sigma^{foc}) + (1 - F_i)\,\mathcal{N}(0, \Sigma^{Unf}). \quad (7)$$

Regarding the prior $p(F_i|S)$, it is defined as a categorical distribution over the eight speaking status.

Finally, at inference, to make a decision we first find the target $\hat{i}$ for which the likelihood $p(F_{\hat{i}}|Z_{\hat{i}}, S)$ of looking at this target is maximal. Then, if

$$p(F_{\hat{i}} = foc|Z_{\hat{i}}, S) > p(F_{\hat{i}} = Unfoc\,|\,Z_{\hat{i}}, S),$$

$\hat{i}$ is set as the subject's VFOA, otherwise it is defined as being *aversion*.

4.3. Experimental Protocol

Performance measure. To compare methods, we report the mean and standard deviation of VFOA classification accuracy of the subjects, where the accuracy per subjects is computed as:

$$vfoaAcc = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}_{\hat{f}_t = f_t}, \quad (8)$$

where $\hat{f}_t$ is the estimated VFOA and f_t the ground truth. Moreover, we report the mean of VFOA classification macro F1-score of the subjects, to ensure that the model does not exploit classes' unbalance to reach a good accuracy.

Experimental protocol. Regarding the protocol, our method and the baseline are first evaluated on both datasets separately with a leave-one-out protocol, reporting the average of the mean VFOA accuracy computed on each subject. For *cross-datasets* experiments, a single model is trained on one of the datasets and evaluated on the second dataset without any adaptation, and we also compute the mean VFOA accuracy for each subject in the test dataset and report their average. Finally, our method is evaluated by training and testing a model on both datasets together (*all-datasets* experiments), for which we use a 4-fold cross-validation protocol, with 1 KTH-Idiap, 1 UBImpressed 'Interviews', and 1 UBImpressed 'Desk' sessions per fold. We compute the mean VFOA accuracy for each subject and report their average by dataset to allow comparison with other experiments.

Parameters. We fixed the VFOA classification threshold τ_{vfoa} (see Section 3.3) to 10°, which corresponds to a tolerance of 35cm at 2 meters. In addition, in our experiments, to produce the feature maps, we used an isotropic Gaussian kernel Σ^m with a standard deviation of 10° for all maps, except for the head pose map where we used 20° which better encompasses the range from the head pose where the gaze can be.

Table 1. VFOA classification accuracy mean, standard deviation, and macro F1-score across subjects. (Abbreviations: 'h' stands for head pose, 'g' for gaze direction, 'vsal' for visual saliency, 'asal' for audio saliency, and 'spk' for subject's speaking status).

Method	UBImpressed		KTH-Idiap	
	vfoaAcc	$F1$	*vfoaAcc*	$F1$
a) Overall results				
baseline	0.84 ± 0.13	0.80	0.80 ± 0.11	0.74
VFOAmap Net	0.85 ± 0.13	0.82	0.81 ± 0.15	0.75
VFOAmap Net + dataAug	0.82 ± 0.12	0.78	0.82 ± 0.15	0.74
b) Input ablation study				
headGaze (h-g)	0.80 ± 0.15	0.78	0.60 ± 0.17	0.56
onlyScene (vsal-asal-spk)	0.60 ± 0.14	0.37	0.63 ± 0.15	0.51
noGaze (h-vsal-asal-spk)	0.67 ± 0.12	0.55	0.73 ± 0.14	0.59
noHead (g-vsal-asal-spk)	0.83 ± 0.12	0.79	0.82 ± 0.15	0.74
noAudio (h-g-vsal)	0.88 ± 0.10	0.80	0.78 ± 0.14	0.70
c) Cross-datasets evaluation				
baseline	0.71 ± 0.11	0.58	0.62 ± 0.15	0.56
VFOAmap Net	0.74 ± 0.14	0.65	0.70 ± 0.15	0.61
VFOAmap Net + dataAug	0.85 ± 0.12	0.82	0.79 ± 0.13	0.71
c) All-datasets evaluation				
baseline	0.80 ± 0.10	0.77	0.82 ± 0.12	0.75
VFOAmap Net	0.85 ± 0.09	0.85	0.85 ± 0.17	0.75
VFOAmap Net + dataAug	0.87 ± 0.10	0.83	0.85 ± 0.17	0.77

4.4. Results

Intra-datasets evaluation. Intra-dataset results are reported in Tab. 1a. Given the difficulty of the task, we can see that the multimodal baseline already produces very good results on both datasets. Looking at the standard deviation, we can also notice an important differences between subjects, which remains in all experiments. In intra-dataset experiments, the proposed method achieves marginally better than the baseline. Also, the F1-score, which puts more weight on incorrectly classified cases compared to accuracy, shows a similar trend, showing that performances are not only due to more *target* and fewer *aversion* predictions but to good recognition of all classes. The data augmentation does not help here, which is probably due to the amount of available data compared to the relatively low target positions variance in the datasets, and the actual potential overfitting when conducting such intra-dataset experiments.

Input ablation study. In the proposed approach, the input consists of five 2D maps and we are interested in testing the contribution of the different features to the overall performance. To do so, we removed some maps to see how it affects performance. We tested five combinations of inputs, and results are given in Tab. 1b.

Results confirm that VFOA estimation benefits both from subject features and scene information, as experiments with only head and gaze (*headGaze*) or scene cues (*onlyScene*, an experiment which allows to check the impact of only prior on results), do not reach the performance of the proposed approach. In addition, while adding head pose improves the performances of *onlyScene* (*noGaze* experiments), it is not as strong as adding gaze alone (see *noHead*) which almost reaches the performance of the proposed approach (*VFOAmap Net*), indicating that in our data, the

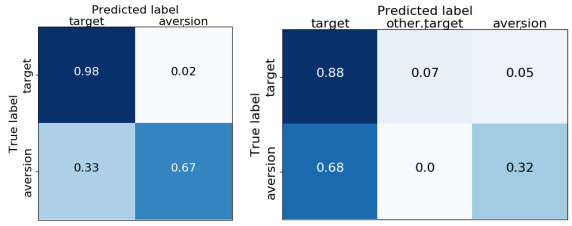

Figure 3. Confusion matrices for UBImpressed (left) and KTH-Idiap (right) datasets after cross-dataset training with data augmentation.

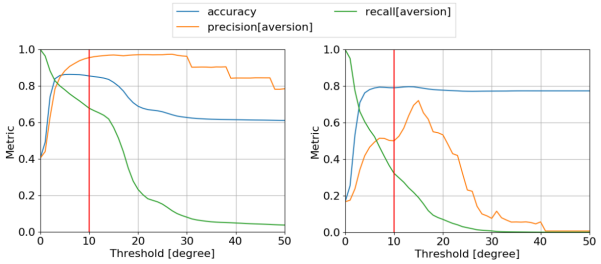

Figure 4. Accuracy, as well as the precision and recall curves of the *aversion* class against the VFOA classification threshold for UBImpressed (left) and KTH-Idiap (right) datasets, after cross-dataset training with data augmentation. The red vertical line indicates the default value of $\tau_{vfoa} = 10^\circ$ used in our experiments.

head pose does not contribute much when the gaze is available. Finally, audio information (subject's speaking status and audio saliency) seems to be more relevant in the multi-target case of the KTH-Idiap dataset (compare *noAudio* to *VFOAmap Net*), which is expected as in such a case, the tendency of looking at the speaker can help to solve ambiguities. These results show that the proposed method mainly exploits gaze and visual saliency when they are available, but that all inputs contribute to robustness (even if they are redundant, e.g. head pose).

Cross-datasets evaluation. Tab. 1c reports the resulting VFOA accuracy when the model is trained on the other dataset (i.e. trained on KTH-Idiap and evaluated on UBImpressed and vice-versa). These clearly demonstrate the generalization capabilities of the proposed method. In particular, while the results achieved only with the available training data are below the intra-dataset results by 10%, using data augmentation (*dataAug*) allows to close the gap and to achieve results as good as if the method was trained on the dataset itself. For comparison, the baseline's accuracy decreases of respectively 13% and 18%, which shows that generalizing from a dataset to another is not trivial.

All-datasets evaluation. In our case, training on both UBImpressed and KTH-Idiap together (see Tab. 1d) slightly improves the performances compared to cross-datasets experiments ($+2\%$ and $+6\%$ with data augmentation). Also, data augmentation only marginally improves the results, probably for the same reasons as in the intra-dataset case. These results, which are the best among our experiments, show the advantage of the proposed method that can successfully train a single model using several datasets with different setups and target numbers.

Confusion matrices. Figure 3 shows the confusion matrices for the models trained in a cross-dataset fashion with data augmentation. We did not report the confusion matrices for the intra-dataset experiments as they are very similar to these. In the KTh-Idiap case, computing a confusion matrix is difficult as when the VFOA ground truth is a target, the network can output *aversion* (false negative), the correct target (true positive), or another target. We fixed the latter case by adding an *other target* column in the confusion matrix.

Looking at the resulting matrices, the proposed method has more difficulties to detect aversions, as it achieves a better *target* recall (0.98 for UBImpressed and 0.88 for KTH-Idiap) compared to *aversion* recall (0.67 and 0.32 recall respectively).

Looking at the KTH-Idiap dataset, most of the errors come from the model predicting a *target* instead of *aversion*. This and the very low *aversion* recall can be explained by the class imbalance, as only 17% of VFOA are aversions. Nevertheless, class balancing strategies might not be desired, as this imbalance is a characteristic of multi-party meetings and from an application viewpoint, there is no obvious reason to favor recall over accuracy. The low aversion recall might also be explained by the defined loss, which does not set a precise prediction target in aversion cases. It should be noted that we reported only the confusion matrices for a VFOA threshold τ_{vfoa} of 10°, without searching to maximize the recall.

Accuracy versus VFOA classification threshold. In the above experiments, we set the VFOA classification threshold τ_{vfoa} to an arbitrary value of 10°. Figure 4 shows the impact of this parameter on the accuracy, precision, and recall of the *aversion* class. All these metrics were computed for each subject, and we report their average for each value of τ_{vfoa}.

Overall, the accuracy is maximal in a region between 5° and 15°, which is probably due to the choice of the Gaussian kernel's standard deviation. Also, one can see that we could increase *aversion* recall without losing accuracy by choosing a smaller threshold.

Looking at the KTH-Idiap case, the accuracy peak is smaller and increasing the threshold makes the accuracy saturate toward a value of 0.77, which is near to that *target* class ratio in the dataset. This may suggest that the network's good score is particularly due to its ability to chose between the different targets. Still, when the threshold is around 5°, both accuracy and *aversion* recall are above 0.60,

showing that the network is somehow able to distinguish aversion from target.

5. Conclusion

In this work, we propose a deep learning based method that estimates VFOA estimation from visual and audio features encoded as 2D maps, which provides setup normalization and allows to consider an arbitrary number of targets. Especially, the proposed method was shown successful in cross-datasets experiments, which is a promising step to estimate VFOA in new setups without needing to retrain or fine-tune the model.

One limitation of the proposed method is the need for the 3D position and speaking status of all participants in the scene. It is usually considered as available in research targeting HHI [4, 14, 17, 5] or HRI [20] applications where one of the main goals is to monitor the conversation and interactions between participants, including the speaking status. However, in other applications, e.g. TV shows or internet videos, how to extract this information reliably and the impact on performance will require further investigation Future work will also consist of testing our method on other datasets with even more intra-setup variance in terms of target position and number. Indeed, in both presented datasets, the number of targets does not change during the interaction, even if cross-dataset results are promising. Also, the proposed network could be enhanced with temporal information, using recurrent layers for example, or by adding other maps like encoding the gaze of the targets. Finally, it would be interesting to see if this approach could be applied to different tasks, like gaze refinement or gaze synthesis.

Acknowledgements

This research has been supported by the European Union Horizon 2020 research and innovation programme (grant agreement no. 688147, MuMMER project) and by the Idiap research institute's "Valais-Wallis Ambition" initiative. Also, it uses the KTH-Idiap database made available by KTH, Sweden.

References

[1] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 25–32. IEEE, 2014. 1

[2] Stylianos Asteriadis, Kostas Karpouzis, and Stefanos Kollias. Visual focus of attention in non-calibrated environments using gaze estimation. *International Journal of Computer Vision*, 107(3):293–316, 2014. 2

[3] Sileye Ba and Jean-Marc Odobez. Recognizing visual focus of attention from head pose in natural meetings. *Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1), 2008. 1, 2

[4] Sileye Ba and Jean-Marc Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *Transactions on Pattern Analysis and Machine Intelligence*, 33(1), 2011. 2, 5, 8

[5] Chongyang Bai, Srijan Kumar, Jure Leskovec, Miriam Metzger, Jay Nunamaker, and V. S. Subrahmanian. Predicting the visual focus of attention in multi-person discussion videos. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 4504–4510. International Joint Conferences on Artificial Intelligence Organization, 2019. 1, 5, 8

[6] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 3

[7] Kenneth Funes, Florent Monay, and Jean-Marc Odobez. Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *Proc. of the ACM Symposium on Eye Tracking Research and Applications*, 2014. 4

[8] Kenneth Funes and Jean-Marc Odobez. Gaze estimation in the 3d space using rgb-d sensors. *International Journal of Computer Vision*, 118:194–216, 2016. 4

[9] Sebastian Gorga and Kazuhiro Otsuka. Conversation scene analysis based on dynamic Bayesian network and image-based gaze detection. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction on - ICMI-MLMI '10*. ACM Press, 2010. 1, 2

[10] Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Transactions on Interactive Intelligent Systems (TIIS)*, 6(1):1–31, 2016. 1

[11] Davis E. King. Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758, 2009. 4

[12] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019. 2

[13] Benoit Masse, Sileye Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2711–2724, 2018. 1, 2

[14] Philipp Müller, Michael Xuelin Huang, Xucong Zhang, and Andreas Bulling. Robust eye contact detection in natural multi-person interactions using gaze and speaking behaviour. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, pages 1–10, 2018. 2, 8

[15] Skanda Muralidhar, Remy Siegfried, Jean-Marc Odobez, and Daniel Gatica-Perez. Facing employers and customers: What do gaze and expressions tell about soft skills? In *Proceedings of the International Conference on Mobile and Ubiquitous Multimedia*, pages 121–126. ACM, 2018. 1, 2, 5

[16] Catharine Oertel, Kenneth A Funes, Samira Sheikhi, Jean-Marc Odobez, and Joakim Gustafson. Who will get the

grant?: A multimodal corpus for the analysis of conversational behaviours in group interviews. In *Proceedings of the workshop on Understanding and Modeling Multiparty, Multimodal Interactions*, pages 27–32, 2014. 2, 5

[17] Kazuhiro Otsuka, Keisuke Kasuga, and Martina Kohler. Estimating visual focus of attention in multiparty meetings using deep convolutional neural networks. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, pages 191–199, 2018. 1, 2, 5, 8

[18] Kazuhiro Otsuka, Junji Yamato, Yoshinao Takemae, and Hiroshi Murase. Conversation scene analysis with dynamic bayesian network basedon visual head tracking. In *2006 IEEE International Conference on Multimedia and Expo*, pages 949–952. IEEE, 2006. 2

[19] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *European Conference on Computer Vision*, pages 747–763. Springer, 2020. 2, 4

[20] Samira Sheikhi and Jean-Marc Odobez. Combining dynamic head pose-gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognition Letters*, 66:81–90, 2015. 1, 2, 8

[21] Remy Siegfried, Yu Yu, and Jean-Marc Odobez. Towards the use of social interaction conventions as prior for gaze model adaptation. In *Proceedings of the International Conference on Multimodal Interaction*, pages 154–162. ACM, 2017. 2

[22] Rainer Stiefelhagen, Michael Finke, Jie Yang, and Alex Waibel. From gaze to focus of attention. In *International Conference on Advances in Visual Information Systems*, pages 765–772. Springer, 1999. 2

[23] Yu Yu, Kenneth Funes, and Jean-Marc Odobez. Headfusion: 360 degree head pose tracking combining 3d morphable model and 3d reconstruction. *Transactions on Pattern Analysis and Machine Intelligence*, 40(11), 2018. 3

[24] Zeynep Yücel, Albert Ali Salah, Çetin Meriçli, Tekin Meriçli, Roberto Valenti, and Theo Gevers. Joint attention by gaze interpolation and saliency. *IEEE Transactions on cybernetics*, 43(3):829–842, 2013. 2

[25] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2017. 4