Predicting the Visual Focus of Attention in Multi-Person Discussion Videos

Chongyang Bai[1*], **Srijan Kumar**[2,3], **Jure Leskovec**[2], **Miriam Metzger**[4], **Jay F. Nunamaker**[5] and **V.S. Subrahmanian**[1]
[1]Dartmouth College
[2]Stanford University
[3]Georgia Institute of Technology
[4]University of California Santa Barbara
[5]University of Arizona
cy@cs.dartmouth.edu, {srijan,jure}@cs.stanford.edu,
metzger@ucsb.edu, jnunamaker@cmi.arizona.edu, vs@dartmouth.edu

Abstract

Visual focus of attention in multi-person discussions is a crucial nonverbal indicator in tasks such as inter-personal relation inference, speech transcription, and deception detection. However, predicting the focus of attention remains a challenge because the focus changes rapidly, the discussions are highly dynamic, and the people's behaviors are inter-dependent. Here we propose *ICAF* (Iterative Collective Attention Focus), a collective classification model to jointly learn the visual focus of attention of all people. Every person is modeled using a separate classifier. *ICAF* models the people collectively—the predictions of all other people's classifiers are used as inputs to each person's classifier. This explicitly incorporates interdependencies between all people's behaviors. We evaluate *ICAF* on a novel dataset of 5 videos (35 people, 109 minutes, 7604 labels in all) of the popular Resistance game and a widely-studied meeting dataset with supervised prediction. *ICAF* outperforms the strongest baseline by 1%–5% accuracy in predicting the people's visual focus of attention. Further, we propose a lightly supervised technique to train models in the absence of training labels. We show that light-supervised *ICAF* performs similar to the supervised *ICAF*, thus showing its effectiveness and generality to previously unseen videos.

1 Introduction

Given a group G of people, a person $P \in G$, and a short video clip v (1/3rd sec), the Visual Focus of Attention (VFOA) problem is to automatically predict who person P is looking at among all people in G in the video clip v. Solving the VFOA problem can provide profound insights into a number of factors, e.g., who is the dominant person in the group [Hall *et al.*, 2005], who supports/opposes who in the group, who trusts/distrusts who in the group [Knapp *et al.*, 2013].

Figure 1(a) illustrates some of the challenges involved. First, even within a very short 1 second clip, a person may look at many people. The four frames shown in Figure 1(a) show the pictured subject looking at three people. Second, multi-person discussions are highly dynamic because many people may speak at the same time and the speakers change rapidly (Figure 1) — and as people often look at a speaker, solving VFOA requires the ability to rapidly estimate the VFOA. Third, non-verbal behaviors (e.g. eye rolling, head shaking) of people may influence another person's VFOA. Returning to Figure 1(a), one would expect people to look at the lady shown when she is speaking — however, their gaze may turn elsewhere if some unseen person makes a gesture. Alternatively, predicting the VFOA of person P might depend on predicting the VFOA of person P1 as both of them might be looking at the same person P2 who is speaking or gesturing. *In short, solving VFOA requires reasoning at the sub-second level and making rapid changes that take into account not only video of the person P whose gaze we are trying to predict, but also that of others.*

We address these challenges via a novel algorithm called *ICAF* (stands for Iterative Collective Attention Focus) which: (i) reasons at the 1/3 second level that prior research has established as the normal duration humans need to visually focus their attention [Rayner, 2009], (ii) incorporates collective classification [Sen *et al.*, 2008; Kong *et al.*, 2012] intuitions to capture the fact that where person P is looking might depend on where others are looking, and simultaneously assign VFOAs to all people rather than doing so independently, and (iii) *ICAF* iteratively builds a multi-layer network that captures the evolution of the collective classification. This captures the idea that predictions of who P is looking at depends on predictions of who others in the group are looking at. (iv) *ICAF* specifically captures the temporal dependency of VFOA, e.g. the conditional probability that P is looking at Q, given that she was looking at Q in the previous 1/3 sec. *To the best of our knowledge, no prior work on gaze estimation has considered using where others are currently looking and using this to arrive at a joint prediction as we do.*

We introduce a novel dataset (109 mins of video from 5 episodes of the Resistance game in 3 different countries with

*Contact Author

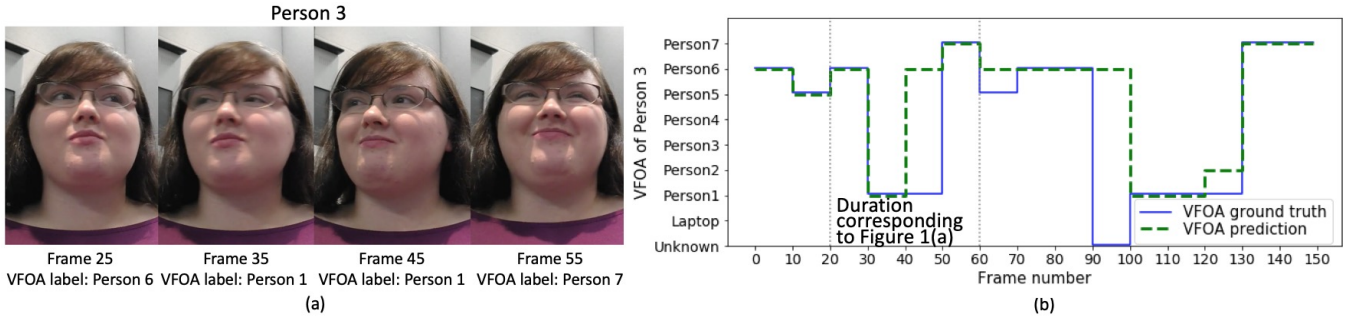

Figure 1: (a) An example of a person's (Person 3) Visual Focus of Attention (VFOA) in 4 frames out of a contiguous 4/3 second (40 frames) during a discussion. person 3's VFOA changes rapidly within this short time period, from looking at persons 6, 1, 1, 7, in frames 25, 35, 45, and 55, respectively. Note that even though the head pose in frames 25 and 55 are similar, the VFOA is different (6 vs 7) (b) Person 3's ground truth VFOA and predicted VFOA made by the proposed method, *ICAF*, of a 5-second discussion clip in which frames 20–60 correspond to Figure 1 (a). We observe that *ICAF* is able to efficiently predict the rapid change in VFOA.

35 people). The data was annotated with ground truth VFOA at the 1/3 second level (a huge task by itself leading to over 19,000 annotated 1/3 second clips). Resistance is an immensely popular, dynamic, animated (and sometimes very noisy) party game involving 5-8 people per game.

We experimentally show that *ICAF* outperforms several strong baselines in predicting people's next VFOA by over 1.3%, i.e. given a training video up to second t, we predict where each person looks at second $t + 1/3$. Moreover, *ICAF* outperforms the best baseline between 1%–5% when predicting next k VFOAs. For example Figure 1(b) shows that even though Person 3 rapidly changes her VFOA during a 5 second multi-person discussion, *ICAF* predicts her VFOA correctly in 11 out of 14 points (78.6% accuracy). Finally, we experimentally show that both temporal dependency and collective classification boost *ICAF*'s performance.

Since getting ground truth labels is a tedious task, we create a lightly supervised version of *ICAF* that uses the speaker label to make predictions. We experimentally show that lightly supervised *ICAF* has similar performance to *ICAF*, showing the potential of using *ICAF* for previously unseen videos.

The demo, code, and predicted VFOA networks are available at: https://cs.dartmouth.edu/dsail/demos/icaf.

2 Related Work

As tracking eye gaze in video is difficult (video resolution, eye visibility, etc.), many estimate head pose as VFOA [Stiefelhagen *et al.*, 1999; Voit and Stiefelhagen, 2008; Zhang *et al.*, 2008; Stiefelhagen and Zhu, 2002]. In real cases, head pose and VFOA may differ. Figure 1(a) shows an example in our dataset—while the person's head pose is similar in frames 25 and 55, her VFOA is different. [Asteriadis *et al.*, 2014] fused head pose and eye gaze to reduce prediction error. Our *ICAF* additionally adds speaking probabilities as features.

[Ba and Odobez, 2009] used head pose to model VFOA by GMM and HMM with person-based Maximum A Posterior parameters. [Sheikhi and Odobez, 2012] added temporal gaze change in HMM. Their methods predict VFOA individually. Instead, our collective classification model enables joint predictions of all people based on head pose and eye gaze.

In group settings, people's VFOA are influenced by each other. [Stiefelhagen *et al.*, 2002] introduced speaking priors to capture VFOA. [Ba and Odobez, 2008] further used meeting context (e.g slides updating) prior. [Ba and Odobez, 2011] additionally created a Dynamic Bayesian Network capturing the shared VFOA, but the sharing prior is constant and same for all people. In contrast, our *ICAF* adds inter-person dependency, enabling the classifiers to learn the weights for other inputs, allowing changes over time as behaviors shift during a video. [Massé *et al.*, 2017] proposed a temporal graphical model to jointly track people's gaze and VFOA. Unlike us, they assumed conditional independence of people's VFOAs given their observed head poses. [Duffner and Garcia, 2013; Duffner and Garcia, 2016] clustered VFOA via Histogram of Gradient features. Unlike them, we use a speak prior for light supervision and show its efficacy by comparing with fully supervised results.

Collective classification. Collective classification methods are widely used in graph mining tasks such as node labeling [Sen *et al.*, 2008; Kong *et al.*, 2012], link prediction [Taskar *et al.*, 2004] and a combination of both [Bilgic *et al.*, 2007]. These methods are able to correlate node/edge attributes to train a mutually dependent classifier ensemble. However, none of these models directly predicting VFOA from videos. To the best of our knowledge, *ICAF* is the first method to use collective classification to predict the VFOA of all people simultaneously in a multi-person video.

3 Dataset and Problem Setup

We collected a dataset involving the Resistance game[1] containing five games from five different locations—three from U.S.A., one from Israel, and one from Singapore. In each game, up to eight people are seated in an octagon layout (Figure 2). It has a total of 35 people whose goal is to identify deceptive people for additional financial reward. Each person has a tablet in front of them which records their activity. At the start of every game, all people introduce themselves, followed by several rounds of discussion

[1]https://en.wikipedia.org/wiki/The_Resistance_(game)

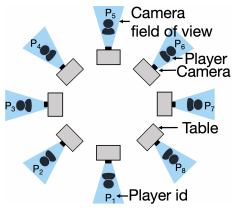

Figure 2: Data collection setup

Video id	Number of seconds	10-frame segments	Number of labels
1	1062	3186	1086
2	896	2688	1541
3	1435	4305	1516
4	1984	5952	2060
5	1134	3402	1401
Total	6511	19533	7604

Table 1: Resistance dataset

where 2-3 people are deceptive and do not want to be identified by the other people whose goal is to unmask them. The people may not leave their seats. The discussions are emergent as there is no pre-determined presenter or leader.

We generated ground-truth labels for people's VFOA for every 10 frames (1/3 seconds in 30 frames per second videos), the time taken to register one's attention [Rayner, 2009]. Figure1(a) is an example. An expert manually assigned one label for every 10 frame segment of each person. For each person, there are eight possible points of focus—one of the other 7 people and the tablet. A label is assigned if the person looks at the object (person or tablet) for the majority of the 10 frames, otherwise, an 'unknown' label is assigned. This results in a total of 7604 valid labeled segments. The 'unknown'-labeled segments are not used for training or testing.

We extract 3 clips from each game—the entire introduction round (where at most one person is speaking at a time), and two 5-second discussions (where multiple people are simultaneously speaking). This gives 6511 seconds of data in total for the 5 games. Table 1 shows the data distribution by game.

AMI corpus. We also used the widely-studied AMI meeting corpus [McCowan *et al.*, 2005], which is highly structured. In this dataset, we used closeup videos of 12 meetings with available VFOA annotation. Each meeting has 4 people and lasts 25 minutes on average. The VOFA targets are 4 people, table, whiteboard and slide screen.

3.1 Feature Extraction

We extract two sets of features from the clips: face-based features and speaking probability features. As with face-based features, we extract the person's head pose angles and eye gaze vectors using OpenFace [Baltrusaitis *et al.*, 2018] since the tablet cameras can capture close-up video of each person.

Speaking prediction. We use visual information to predict if a person is speaking at an instance. First, we get 2-dimensional lip contour points $X^{(t)} = \{(x_i^{(t)}, y_i^{(t)}), i = 1, \ldots, n\}$ at frame t from OpenFace and normalize $X^{(t)}$ by its bounding box to avoid the influence of head movement. Second, we compute the gradient of point positions over time to capture mouth movement, which is $\vec{g}_i^{(t)} = (x_i^{(t)} - x_i^{(t-1)}, y_i^{(t)} - y_i^{(t-1)}), i = 1, \ldots, n$, and aggregate them as a frame feature vector $\vec{g}^{(t)}$. Third, we get feature $G^{(t)}$ by concatenating $(\vec{g}^{(t-s+1)}, \vec{g}^{(t-s+2)}, \ldots, \vec{g}^{(t)}, \ldots, \vec{g}^{(t+s)})$ around time t, in a window of size $2s$. This forms a sliding window over time. We use $G^{(t)}$ as a feature, and the introduction part of a game from this dataset to train a general speaking detection model **SP**. Finally, the speaking probability of a person at time t is given by $s = \mathbf{SP}(G^{(t)})$.

We do not create a new model for head pose angles or eye gaze vector extraction. Instead, we use these as inputs to our model to improve the predictions by using them collectively, instead of independently. *ICAF* takes the head-based features and speaking probability features as inputs.

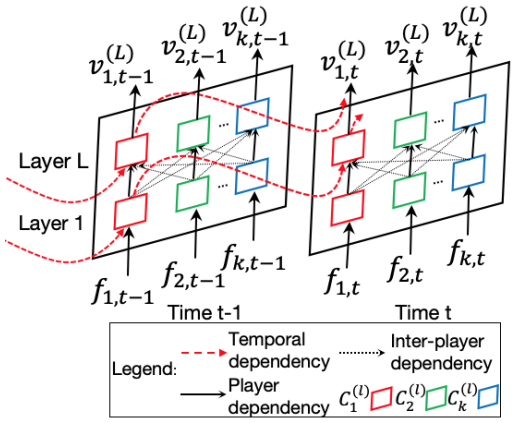

Figure 3: Architecture of the iterative collective classification model, *ICAF*. Each classifier C_i takes three inputs: output of its previous layer (person dependency), previous time (temporal dependency), and other people's output (inter-person dependency). Figure is best viewed in color.

4 *ICAF*: Iterative Collective Classification

Here we describe *ICAF*, the collective classification methods that incorporates inter-person dependencies and temporal consistency to jointly predict the VFOA of all people.

Let $\boldsymbol{f}_{i,t}$ denote the raw input feature vector of person $P_i \in \{P_1, \ldots P_k\}$ at time t. The raw input features for P_i include the head pose angles vector, the eye gaze vector and speaking probabilities vector $\vec{s} = (s_1, \ldots, s_{i-1}, 0, s_{i+1}, \ldots, s_k)$. Note that we don't use P_i's speaking probability s_i in $\vec{s}$, as P_i's speaking activity doesn't directly influence her VFOA. Let C_i denote the VFOA prediction model for P_i. *ICAF* builds separate models C_i for each person P_i. C_i outputs a vector $\boldsymbol{v}_{i,t}$, the probability distribution of person P_i's visual focus of attention at time t. This output vector specifies the probability that P_i's VFOA is person P_j (or the tablet) for each j. The ground truth label for person P_i at time t is denoted by $y_{i,t}$.

Figure 3 illustrates *ICAF* for k people and an L-layer network. Each person P_i has one classifier $C_i^{(l)}$ for each layer l. Raw features $\boldsymbol{f}_{i,t}$ are used as input for P_i at time t. The model has multiple layers $1, \ldots, L$ to add inter-person dependencies by using the output of other people's classifiers as input (shown in dotted lines). Each classifier also takes the previous timestep's output as input (shown in dashed lines only for C_1 for simplicity). The final output vectors are $\boldsymbol{v}_{i,t}^{(L)}$.

ICAF has three major inputs for each classifier $C_i^{(l)}$ at every time t and layer l as follows: (i) raw features $\boldsymbol{f}_{i,t}$ associated with P_i, (ii) inter-person dependencies $\boldsymbol{v}_{j,t}^{(l-1)} (j = 1, \ldots, k, j \neq i)$ incorporating the influence of the behavior of other people, and (iii) temporal consistency $\boldsymbol{v}_{i,t-1}^{(l-1)}$ en-

Algorithm 1: *ICAF* MODEL

Input : Raw features $\boldsymbol{f}_{i,t}\ \forall i \in [1,\dots k], t \in [1,\dots T]$, Number of layers L.

Output: Predictions $\boldsymbol{v}_{i,t}^{(L)}$ of all people i at all times t

1 $\boldsymbol{v}_{i,0}^{(l)} = (\frac{1}{k+1}, \frac{1}{k+1}, .., \frac{1}{k+1})$
2 $\boldsymbol{v}_{i,t}^{(0)} = C_i^{(0)}(\boldsymbol{f}_{i,t})$
3 **for** $t \in [1,\dots T]$ **do**
  /* Operate on every time step t */
4 **for** $l \in [1,\dots L]$ **do**
    /* Process every layer l */
5 **for** $i \in [1,\dots k]$ **do**
      /* Update person P_i */
6 $S(V) = \sum_{j\in\{1,\dots k\}-\{i\}} \boldsymbol{v}_{j,t}^{(l-1)}$
7 $\boldsymbol{v}_{i,t}^{(l)} = C_i^{(l)}(\boldsymbol{f}_{i,t}, \boldsymbol{v}_{i,t}^{(l-1)}, \boldsymbol{v}_{i,t-1}^{(l-1)}, S(V))$
8 **end**
    /* Make prediction and save $C_i^{(l)}$ */
9 **end**
10 **end**
11 **return** $\boldsymbol{v}_{i,t}^{(L)}\ \forall i \in [1,\dots k], t \in [1,\dots T]$

abling the model to make temporally consistent predictions. Together, this results in a collective classification model that makes predictions for all people. The overall algorithm of *ICAF* is shown in Algorithm 1.

4.1 Inter-person Dependencies

In a multi-person discussion, the behavior of one person can influence the VFOA of others. Moreover, the behavior of people is highly correlated—when a person is speaking, other people are likely looking at him [Ba and Odobez, 2011]. This mutual influence can be used to make accurate predictions.

We incorporate the person-to-person influence by adding explicit connections between their classifiers (lines 4–8 in Algorithm 1). In particular, for every person P_i's model C_i, we use the predictions of all other people's models $C_j, \forall j \in \{1,\dots,k\} - \{i\}$ as input. The resulting model is mutually-recursive. To solve this recursion, we unfold the model for multiple layers so that the output of layer l is fed as input to layer $l+1$. This is shown as layers $1,\dots L$ in Figure 3.

Thus, the input to person P_i's model $C_i^{(l)}$ at layer l is its output from layer $l-1$ and an aggregation of the set V of outputs from other people's models from layer $l-1$. The aggregation is a summation represented as $S(V)$, which is used as an input to the model (lines 6–7 in Algorithm 1).

To initialize for layer 1, let $\boldsymbol{v}_{i,t}^{(0)} = C_i^{(0)}(\boldsymbol{f}_{i,t})$, where $C_i^{(0)}$ is the classifier trained by only raw features of P_i, separately.

4.2 Temporal Consistency

The VFOA of a person at time t is linked to her VFOA at time $t-1$. The temporal consistency component of *ICAF* explicitly incorporates this dependency by using the output of the predictions made during the last timestep for the person as an input. Specifically, the output $\boldsymbol{v}_{i,t-1}^{(l-1)}$ is an input to $C_i^{(l)}$. This is shown using the dashed lines in Figure 3 and in line 7 in Algorithm 1. For each layer l, we initialize $\boldsymbol{v}_{i,0}^{(l)}$ as a uniform probability distribution for VFOA targets.

$$\boldsymbol{v}_{i,t}^{(l)} = C_i^{(l)}(\underbrace{\boldsymbol{f}_{i,t}}_{\text{Raw input}}, \underbrace{\boldsymbol{v}_{i,t}^{(l-1)}}_{\text{Person input}}, \underbrace{\boldsymbol{v}_{i,t-1}^{(l-1)}}_{\text{Temporal input}}, \underbrace{\sum_{j\in\{1,\dots k\}-\{i\}} \boldsymbol{v}_{j,t}^{(l-1)}}_{\text{Inter-person input}})$$

Figure 4: Final formulation of *ICAF* to output $\boldsymbol{v}_{i,t}^{(l)}$ of person i at time t on layer l.

The final formulation with all the components is shown in Figure 4. Overall, *ICAF* uses the real time inputs along with temporal and inter-person dependencies to jointly predict the visual focus of attention of all people.

5 Experiments

We conduct several experiments on Resistance and AMI datasets to show:

- *ICAF* outperforms all strong baselines by 1.3% in predicting VFOA in the next time step (i.e., 10 frames) with $p = 0.046$ by two-sample t-test.
- *ICAF* significantly outperforms the highest baseline by up to 5% when making predictions upto k time steps in the future ($p < 0.05$).
- Collective classification and temporal dependencies boost the performance of *ICAF* significantly.

Baselines. We compare with three sets of baselines that use head pose vector (H), eye gaze vector (E), and speaking probability vector(S) for predictions. The first set of baselines are [Ba and Odobez, 2009; Ba and Odobez, 2011; Massé *et al.*, 2017], with comparable numbers of VFOA targets in similar settings. Specifically, GMM(H), GMM(H,E) use Gaussian Mixture Model with parameters from each individual [Ba and Odobez, 2009]. HMM(H), HMM(H,E) uses Hidden Markov Model [Ba and Odobez, 2009]. DBN(H,S), DBN(H,E,S) uses Dynamic Bayesian Network (DBN) incorporating conversational dynamics and a shared constant focus prior [Ba and Odobez, 2011]. Note that the screen activity feature is removed to adapt to our dataset. G-DBN uses DBN to track VFOAs and eye gaze simultaneously with people's global head poses as inputs [Massé *et al.*, 2017]. In our dataset, people sit uniformly in a circle, so we convert their local head poses to global ones given poses of their cameras. Further, we created two more sets of baselines using three sets of features H, (H,E) and (H,E,S). The second set of baselines trains one general classifier GC for all people by including the person index as input feature vector [Ba and Odobez, 2011]. The last set of baselines trains a person-specific classifier PC for each person [Asteriadis *et al.*, 2014]. As in the case of GC, we create three baselines PC(H), PC(H,E), and PC(H,E,S).

Experimental setting. To get speaking probability features, we set the sliding window size as 30 frames (1 sec) and train a Random Forest speaking detection model **SP**. The training data uses people's introductions as speaking samples,

GMM(H,E)	0.716
HMM(H,E)	0.770
DBN(H,E,S)	0.800
G-DBN	0.782
GC(H,E,S)	0.756
PC(H,E,S)	0.818
ICAF	**0.831**

Table 2: Experiment 1: Next VFOA Prediction: Table reports accuracy of *ICAF* and baselines using all features. Note that the best results of GC, PC, and *ICAF* are achieved by RF. *All improvements of ICAF are statistically significant ($p < 0.05$).*

and other people's introductions as non-speaking samples. The introductions were not drawn from our 5 video samples. We evaluate *ICAF* and baselines by respecting the temporal order of data. Instead of doing a k-fold cross-validation, we train the model for the first T data points and test on the $T + 1^{th}$ data point (each data point consists of 10 frames). T is varied from 96.3% to 99.9%, and the results are averaged. Recall that the data for each game is divided into three parts: an introduction round and two discussion rounds. The introduction round clips are only used for training, and the temporal evaluation is done with the two discussion rounds. Both training and testing are at the frame level. Frame VFOA probabilities are further averaged over 10 frames as probabilities at each 10-frame segments. Given the generality of our model, we experiment with 4 classifiers: Random Forest (RF), Logistic Regression (LR), Linear SVM (LINSVM) and Gaussian Naive Bayes (NB). In all cases, *ICAF* has 3 layers. All models are compared using the accuracy metric.

Experiment 1: Next VFOA prediction. We compare *ICAF* with all baselines using all features. All models are trained on the first T data points and then used to predict the $T + 1^{th}$ data point. *Note that this means that we are predicting the visual focus of attention for each person 1/3 second into the future.* The features given to *ICAF* for every frame are the head pose vector (H), eye gaze vectors (E), and speaking probability vectors (S). Table 2 shows the results. For fairness, we add eye gaze features (E) to baselines GMM, HMM and DBN. (i) Person-specific baseline models perform better than the corresponding general-classifier baselines using the same set of features. Specifically, PC(H,E,S) performs at least 6.2% better than GC(H,E,S). (ii) More importantly, *ICAF* performs between 1.3%–11.2% better than all baseline models. (iii) Indeed, it is 3% higher than state-of-the-art method DBN(H,E,S).

Experiment 2: Longer-future predictions. We next evaluate the robustness of *ICAF* by predicting the $T + k^{th}$ data point while training only till the T^{th} data point. We vary k from 1 to 10, meaning that we predict who a person will look at between 0.3 and 3.3 seconds into the future. Figure 5 shows the result. *ICAF* outperforms the best baseline by up to 5%. In fact, it is better than DBN(H,E,S) by 1.5%–5.7%. Moreover, *ICAF* is relatively stable as k increases, while some baselines drop rapidly. Specifically, *ICAF*'s prediction accuracy varies only 7.5% over k, so it gives robust estimation of VFOA in the longer-term future.

Experiment 3: Contribution of collective classification. Figure 6 compares the results of *ICAF* with and without the

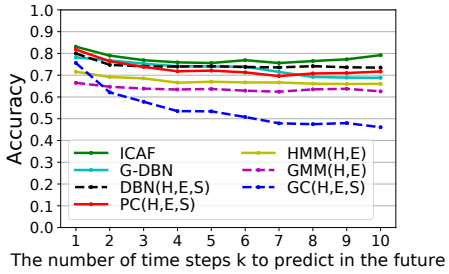

Figure 5: Experiment 2: Longer-Future Prediction: Accuracy of predicting k steps to the future. *ICAF* is the highest over all time steps, and outperforms the best baseline by up to 5% ($p < 0.05$).

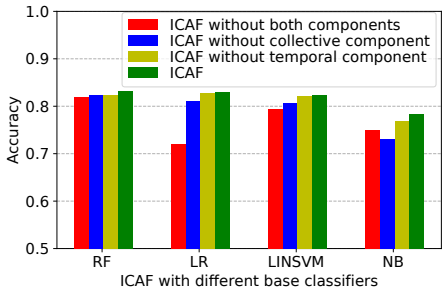

Figure 6: Experiment 3: Contribution of collective classification: The performance drops when either the collective or the temporal components is removed and drastically when both are removed.

temporal and collective classification components. Note that *ICAF* without both components is equivalent to the baseline PC(H,E,S). We observe that each of them boost the performance of *ICAF* from 0.2% to 5.3% w.r.t. all base classifiers. The combination of both components is important in *ICAF*: the performance of PC(H,E,S) is lower than *ICAF* without either of the components. Additionally, adding collective classification improves performance more than the temporal component alone. Therefore, both temporal and collective classification components of *ICAF* are essential, and the collective component results is more critical for good predictions.

Experiment 4: Comparison with different features. We next explore the effects of different features on *ICAF* and baselines. Note that RF is used as the (base) classifier to obtain best results for GC, PC, and *ICAF* . Table 3 shows the results for next VFOA prediction. First, for all models, eye gaze features E boost the predictions. It especially boosts [Ba and Odobez, 2011; Ba and Odobez, 2009] by at least 13.5%. Second, speaking features S boost all models except for GC. These demonstrate that both E and S contribute to prediction of VFOA. Third, using features including E or S, *ICAF* outperforms all baselines.

Experiment 5: Comparison between different base classifiers. Here we explore performance of *ICAF* with different kinds of base classifiers: RF, LR, NB and LINSVM. In Figure 7 we compare *ICAF* with GC and PC in the cases of both next VFOA prediction ($k = 1$) and longer-future VFOA prediction ($k > 1$). We only show 2 out of 4 plots due to space limit, but the results are similar. The colored texts show the

Model	H	H,E	H,S	H,E,S
GMM	0.525	0.716	-	-
HMM	0.623	0.770	-	-
DBN	-	-	0.665	0.800
GC	**0.719**	0.799	0.731	0.756
PC	0.716	0.805	0.771	0.818
ICAF	0.718	**0.811**	**0.784**	**0.831**

Table 3: Experiment 4: Comparison between different features: Both E and S boost the accuracy of all models except GC, and *ICAF* performs the best in 3 out of 4 cases. ($p < 0.05$)

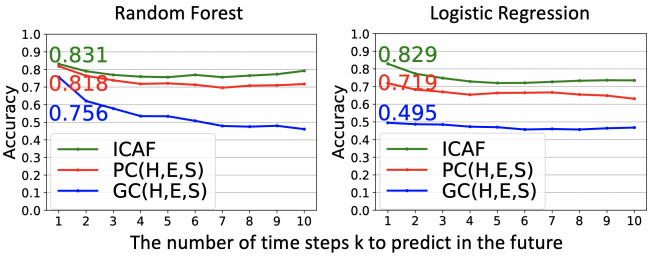

Figure 7: Experiment 5: Comparison between different (base) classifiers. In each subfigure, each of 3 colored numbers indicates the prediction accuracy of $k = 1$ in the same colored line.

results for $k = 1$, where *ICAF* outperforms the corresponding best baseline by 1.3%-11%. For $k > 1$, it outperforms the best baseline by up to 5% with RF, 12% with LR, 3% with LINSVM, and 4% with NB. Thus, we observe the generality of *ICAF*.

AMI corpus experiments. We also conducted experiments on the AMI meeting corpus [McCowan *et al.*, 2005]. 8 meetings are dynamic, where people sit around a table and upto 1 person moves to the whiteboard/screen to present. 4 meetings are static, where all people remain seated. We use people's closeup videos to extract head pose, eye gaze, and speaking probability. We followed the leave-one-out protocol as in [Ba and Odobez, 2011] and compare frame-based accuracy. *Since the 4 seats over all meetings are fixed, we train seat-specific classifiers in ICAF*. Table 4 shows that *ICAF* outperforms [Ba and Odobez, 2011] in both kinds of meetings.

6 Lightly Supervised VFOA Prediction

A major challenge in VFOA prediction is the lack of labeled data for new videos. Annotating VFOA at a second or subsecond granularity is highly time-consuming and often not clean. We now propose to generate accurate VFOA predictions without ground truth labels. The proposed technique is general and can be used to train both the baselines and *ICAF*.

The intuition is that people are highly likely to look at the person who is speaking if there is a single speaker [Stiefelhagen *et al.*, 2002]. Building on this intuition, we identify continuous clip segments where one person is speaking. This is done using the speaking prediction model **SP** described in Section 3.1. To reduce false positives, we further average over 10 frames' prediction probability around the current frame and use it as the final label to select single-speaker segments. For a segment where P_i is speaking, we assign i as the training label for all other people and the model is trained with it. To evaluate the effectiveness of this training method, we train

Model	Static meetings	Dynamic meetings
[Ba and Odobez, 2011]	0.556	0.520
ICAF	**0.568**	**0.538**

Table 4: AMI corpus experiments. Accuracy of the proposed model on static and dynamic meetings.

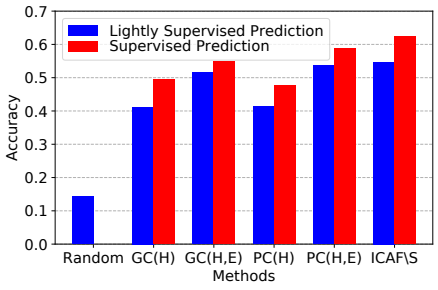

Figure 8: Lightly supervised predictions (in blue) and supervised predictions (in red): 'Random' denotes random prediction accuracy, and *ICAF*\S denotes *ICAF* without speaking feature.

all models using the introduction (by generating its speaker labels) and use the two discussion clips with the ground truth VFOA labels as test.

Figure 8 shows the results for all baselines and *ICAF* using RF as base classifier. Since the training labels are speaking labels, we remove speaking probability features from *ICAF* as well as baselines. Compared to random prediction of 14.4%, the lightly supervised training technique generates 41.2%-54.7% results. We also observe that *ICAF* performs better than the baselines. For comparison, Figure 8 shows the equivalent result with supervised training, where we train the models using the ground truth focus labels in the introduction round as well. We note that the lightly supervised prediction is comparable to supervised prediction, showing the effectiveness of the proposed training technique.

7 Conclusion

We showed that by explicitly incorporating inter-person dependencies and temporal consistency are crucial to accurately predict VFOA both in short-term future and long-term future. The *ICAF* model is, therefore, able to overcome the challenges of rapidly changing VFOA, high dynamics of the discussion, and person-person inter-dependencies. Moreover, the lightly supervised *ICAF* is crucial in making the model general to unseen videos. This opens doors to new research in efficient extraction of interaction networks from videos without any training labels.

Role of Authors. Authors Metzger and Nunamaker designed the Resistance-style game and collected the Resistance data. The remaining authors designed the machine learning algorithms and software, and designed/ran all experiments.

Acknowledgements

This work was funded in parts by ARO Grant W911NF1610342, NSF OAC-1835598, DARPA MCS, ARO MURI, JD.com, Amazon, and Stanford Data Science Initiative. JL is a Chan Zuckerberg Biohub investigator.

References

[Asteriadis *et al.*, 2014] 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.

[Ba and Odobez, 2008] Sileye O Ba and Jean-Marc Odobez. Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 2221–2224. IEEE, 2008.

[Ba and Odobez, 2009] Sileye O Ba and Jean-Marc Odobez. Recognizing visual focus of attention from head pose in natural meetings. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):16–33, 2009.

[Ba and Odobez, 2011] Sileye O Ba and Jean-Marc Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):101–116, 2011.

[Baltrusaitis *et al.*, 2018] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L. P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pages 59–66, May 2018.

[Bilgic *et al.*, 2007] Mustafa Bilgic, Galileo Mark Namata, and Lise Getoor. Combining collective classification and link prediction. In *Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on*, pages 381–386. IEEE, 2007.

[Duffner and Garcia, 2013] Stefan Duffner and Christophe Garcia. Unsupervised online learning of visual focus of attention. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, pages 25–30. IEEE, 2013.

[Duffner and Garcia, 2016] Stefan Duffner and Christophe Garcia. Visual focus of attention estimation with unsupervised incremental learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(12):2264–2272, 2016.

[Hall *et al.*, 2005] Judith A Hall, Erik J Coats, and Lavonia Smith LeBeau. Nonverbal behavior and the vertical dimension of social relations: a meta-analysis. *Psychological bulletin*, 131(6):898, 2005.

[Knapp *et al.*, 2013] Mark L Knapp, Judith A Hall, and Terrence G Horgan. *Nonverbal communication in human interaction*. Cengage Learning, 2013.

[Kong *et al.*, 2012] Xiangnan Kong, Philip S Yu, Ying Ding, and David J Wild. Meta path-based collective classification in heterogeneous information networks. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1567–1571. ACM, 2012.

[Massé *et al.*, 2017] Benoît Massé, Silèye Ba, and Radu Horaud. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[McCowan *et al.*, 2005] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al. The ami meeting corpus. In *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, volume 88, page 100, 2005.

[Rayner, 2009] Keith Rayner. Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, 62(8):1457–1506, 2009.

[Sen *et al.*, 2008] Prithviraj Sen, Galileo Namata, Mustafa Bilgic, Lise Getoor, Brian Galligher, and Tina Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.

[Sheikhi and Odobez, 2012] Samira Sheikhi and Jean-Marc Odobez. Investigating the midline effect for visual focus of attention recognition. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, pages 221–224. ACM, 2012.

[Stiefelhagen and Zhu, 2002] Rainer Stiefelhagen and Jie Zhu. Head orientation and gaze direction in meetings. In *CHI'02 Extended Abstracts on Human Factors in Computing Systems*, pages 858–859. ACM, 2002.

[Stiefelhagen *et al.*, 1999] 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.

[Stiefelhagen *et al.*, 2002] Rainer Stiefelhagen, Jie Yang, and Alex Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, 2002.

[Taskar *et al.*, 2004] Ben Taskar, Ming-Fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In *Advances in neural information processing systems*, pages 659–666, 2004.

[Voit and Stiefelhagen, 2008] Michael Voit and Rainer Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 173–180. ACM, 2008.

[Zhang *et al.*, 2008] Honggang Zhang, Lorant Toth, Jun Guo, Jie Yang, et al. Monitoring visual focus of attention via local discriminant projection. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 18–23. ACM, 2008.