Estimating **Visual Focus of Attention** in Multiparty Meetings using Deep Convolutional Neural Networks

Kazuhiro Otsuka, Keisuke Kasuga†, Martina Köhler‡

NTT Communication Science Labs., Japan
† Nagaoka University of Technology, Japan
‡ University of Tübingen, Germany
Abstract

Estimating **VFoA** in multiparty meetings

*VFOA* = Visual Focus of Attention

= Direction of gaze, i.e. "**who is looking at whom**"

We discovered the potential of **CNNs** in VFoA estimation

**CNNs** = Convolutional Neural Network
Communicative role of VFoA

Estimating VFoA is an essential task in multimodal analysis of multiparty meetings

Gaze interaction is key to

- Monitoring / Expressing / Turn-Taking [Kendon1967]
- Organizing multiparty conversation [Goodwin1981]
- Sharing emotions such as intimacy and empathy

VFoA is a useful cue in

- Controlling conversational robots/agents
- Discovering personal/group traits
- Assessing/training communication sills
Target

VFoA estimation in multiparty meetings + Convolutional Neural Networks (CNNs)
Research Questions

■ [Q0] CNNs vs. Bayes
  Can CNNs predict VFoA better than Bayes models can?

■ [Q1] Multimodal fusion
  What kind of multimodal features contribute to a better prediction?

■ [Q2] Multiparty fusion
  Can CNNs automatically integrate multiparty features and implicitly learn conversation structures similar to the ones explicitly embedded in Bayes models?

■ [Q3] Fusion structure
  What kind of network structure is suitable for integrating multimodal and multiparty features?

■ [Q4] Robustness
  Are CNNs robust against measurement noise in face image tracking?
What’s **CNN**?

The information flows from low-level input data such as an image, to high-level output data such as the class of an object.

Recently, CNNs have been used for **human-activity recognition** using multi-channel time series from wearable accelerometers.
Bayesian VFoA estimation

Bayesian model structures help to better predict VFoA from ambiguous observations, e.g. rough head pose in images.
Data

Four-party group meeting (NTT MM corpus 2004)

2 Groups x 2 Sessions

<table>
<thead>
<tr>
<th>Group</th>
<th>Session</th>
<th>Data Length [frames]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Session 1</td>
<td>10000</td>
</tr>
<tr>
<td></td>
<td>Session 2</td>
<td>9300</td>
</tr>
<tr>
<td>Group 2</td>
<td>Session 1</td>
<td>9100</td>
</tr>
<tr>
<td></td>
<td>Session 2</td>
<td>10000</td>
</tr>
</tbody>
</table>

1 frame = 1/30 sec
9000 frames = 5 min
## Multimodal Feature Set

### Head Pose
- **Time series of 3-DoF rotation angles**
- **Sampling rate: 30 Hz**

### Utterance Interval (Utter)
- Binary (on/off) time series of each person’s voice (mainly manually coded)

### Eyeball direction (Eye)
- **Time series of 3-class (Left/Right/Center) eyeball direction**
- **Horizontal angles of both eyes**

### MoCap Sensors
- **Polhemus’s Fastrak**

### Utterance Contour Tracking (UTT)
- **STCTracker**

### OpenFace
- **RQ4 only**
Multimodal Feature Set

all data synchronized at 30 Hz

Gaze direction is toward one of the other participants or somewhere else (averted)

Ground-truth gaze direction was created by a human coder
Proposed CNN models

**Individual Model** \( M_{\text{Indv}} \)
- It targets a single person’s VFoA.
- It is not a person-specific model, but specific to a seat position.

**Group Model** \( M_{\text{Grp}} \)
- It jointly estimates the VFoA of all participants in a meeting at the same time.
- It is intended to implicitly model group gaze behaviors among meeting participants.

Build separate model for each seat

One model for all
Proposed CNN models

Individual Model $\mathcal{M}_{\text{Indv}}$

INPUT
Length of subsequence=32

$P_i \{ H(Azi): \text{Head pose azimuth} \}
H(Ele): \text{Head pose elevation}
Utter: \text{Utterance}
$E_L: \text{Left-directed eye}$
$E_R: \text{Right-directed eye}$

temporal window ~1 sec

conv1 pool1 conv2 pool2 conv3 fc4

Each channel (feature) is separately processed until fc4, which combines all multimodal features.
Proposed CNN models

Group Model $\mathcal{M}^{Grp}$

**Fusing multiparty data at late stage**

**Stacking individual models**

- **Input**
  - Length of subsequence = 32
  - $H(Azi)$: Head pose azimuth
  - $H(Ele)$: Head pose elevation
  - $Utter$: Utterance
  - $E_L$: Left-directed eye
  - $E_R$: Right-directed eye

- **Output**
  - $N_p \times N_p$
  - $fc5$

Diagram:
- Convolution: $conv1$
- Pooling: $pool1$
- Convolution: $conv2$
- Pooling: $pool2$
- Convolution: $conv3$
- Fully Connected: $fc4$
Optimizing Hyper-Parameters

Search for optimum number of units: \( \hat{N}_1, \hat{N}_2, \hat{N}_3, \hat{N}_4 \)

1st grid search

\( N_4 \)

\( N_1 \)

150

100

50

25

25 50 75 100

2nd grid search

\( N_3 \)

\( N_2 \)

20

10

5

25 40 80

Greedy search

Other parameters (fixed)

- time window: 32 frames (~1 sec)
- number of epochs: 10
- mini-batch = 12
- width of convolution kernels: 5, 6, 3 or conv1, conv2, and conv3

OpenSource CNN Library: MatConvNet ver 1.0 beta23
Evaluation via Cross-Validation

Used **leave-one-session-out** method

<table>
<thead>
<tr>
<th>Individual Model</th>
<th>Training Data</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{\text{Indv}}$</td>
<td>persons sitting in the same seat in the other three sessions, including self and others</td>
<td>single person in a single session</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group Model</th>
<th>Training Data</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$M_{\text{Grp}}$</td>
<td>other three sessions, including same group and another group</td>
<td>one session</td>
</tr>
</tbody>
</table>

Evaluation measure: **macro F-score**

$\text{macro F-score} = \text{average over all F-scores from all people and all sessions}$
Experimental Results
[Q1] Multimodal Features

Addition of eye direction feature significantly improved performance of both models.
Example Scene
Detecting quick glances

Models with Eye feature successfully detected a quick glance.

No other models could detect it.
[Q1] Multimodal Features

In the absence of eye direction feature

The utterance feature had a positive impact only on the group model, not on the individual model.
Listener’s F-scores were higher than those of speakers, because speakers’ gaze is more active and more difficult to predict.

The group model had a greater difference in F-score between speakers and listeners than the individual model did.
### [Q2] Multiparty Fusion
Group model vs. Individual model

<table>
<thead>
<tr>
<th>When Eye direction is</th>
<th>( M_{\text{Indv}} )</th>
<th>( \geq M_{\text{Grp}} )</th>
<th><strong>The individual model</strong> better predicts speaker’s gaze than the group model can.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVAILABLE</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NOT AVAILABLE</td>
<td>( M_{\text{Grp}} )</td>
<td>( &gt; M_{\text{Indv}} )</td>
<td><strong>The group model</strong> outperformed the individual model, especially when predicting listeners’ gaze using the <strong>utterance</strong> feature.</td>
</tr>
</tbody>
</table>

This implies

The group CNN model can implicitly learn underlying group gaze behaviors; **“listeners tend to look at speaker”**

Group model showed higher gaze convergence on the speaker than the individual models did.
When using Head+Utter features, the group model outperformed the Bayes model.
When using Head+Utter+Eye features, both models outperformed the Bayes model.
[Q3] Fusion Structure 1/2

What kind of network structure is suitable for integrating multimodal and multiparty features for predicting VFoA? (Early vs. Late fusion problem)

INPUT

Individual Multimodal Fusion

Multiparty Fusion

OUTPUT

=VFoA class

2 Stage-Late (default)  Late (fc4)  Late (fc5)

Late-All  Late (fc5)

Early-Late  Early (conv1)  Late (fc5)

Early-All  Early (conv1)

Built CNN models with different fusion structures
[Q3] Fusion Structure 2/2

In the group models, 2-stage (late) fusion, integrating individual multimodal features first, followed by multiparty feature, could provide better VFoA estimates.
[Q4] Robustness 1/2

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Head pose</td>
<td>3.59 [deg] MAE (mean absolute error) in azimuth rotation</td>
<td>2.58 [deg]</td>
</tr>
<tr>
<td>Eyeball direction</td>
<td>F-score of eye direction (left, center, right) classification = 0.80</td>
<td>Horizontal angles of left and right eyes, relative to head pose</td>
</tr>
</tbody>
</table>

Head pose and eye direction from image-based tracker include a significant amount of noise and outliers.
Image trackers performed comparably to or even better than the sensor (Fastrak).

CNN models are robust against measurement noise arising from image-based tracking.
Summary
Answers to Questions

[Q0] CNNs vs. Bayes
Can CNNs predict VFoA better than Bayes models can?  
YES

[Q1] Multimodal fusion
What kind of multimodal features contribute to a better prediction?
- Eyeball direction for both models
- Utterance only for group model

[Q2] Multiparty fusion
Can CNNs automatically integrate multiparty features and implicitly learn conversation structures similar to the ones explicitly embedded in Bayes models?  
YES

[Q3] Fusion structure
What kind of network structure is suitable for integrating multimodal and multiparty features
2-stage late fusion

[Q4] Robustness
Are CNNs robust against measurement noise in face image tracking?  
YES
# Discussion

## Pros and cons of CNNs and Bayes

### Bayes
- explicit probabilistic model
- accessible to inside model
- required **handcraft model** specific to target phenomenon
- limited representation power, e.g. mixture of Gaussians
- applicable to small datasets
- relatively easy to generalize

### CNNs
- **black-box** model
- hard to interpret inside model
- automatic learning from data (with hyper-parameter tuning)
- higher representation power
- requires **large dataset**
- generalization is a big issue
Why CNNs? Not RNNs, LSTMs

Capability of modeling temporal dependency

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CNNs</td>
<td>Short-term (e.g. 1 sec in our models)</td>
</tr>
<tr>
<td>RNNs/LSTMs</td>
<td>Long-term</td>
</tr>
</tbody>
</table>

We assumed that VFoA estimation needs only short-term dynamics. So CNNs are suitable for our purpose.

Is this assumption true?
Thank you for your attention!