VISUAL ACTIVITY CONTEXT FOR FOCUS OF ATTENTION ESTIMATION IN DYNAMIC MEETINGS

Sileye O. Ba, Hayley Hung, Jean-Marc Odobez

IDIAP Research Institute, Martigny,Switzerland
Ecole Polytechnique de Lausanne (EPFL), Lausanne, Switzerland

ABSTRACT

We address the problem of recognizing, in dynamic meetings in which people do not remain seated all the time, the visual focus of attention (VFOA) of seated people from their head pose and contextual activity cues. We propose a model that comprises the VFOA of a meeting participant as the hidden state, and his head pose as the observation. To account for the presence of moving visual targets due to the dynamic nature of the meeting, the locations of the visual targets are used as an input variables to the head pose observation model. Contextual information is introduced in the VFOA dynamics through a slide activity variable and speaking or visual activity variables that relate people's focus to the meeting activity context. The main novelty of this paper is the introduction of visual activity context for FOA recognition to account for the correlation between a person's focus and the other people's gestures, hand and body motions. We evaluate our model on a large dataset of 5 hours. Our results show that, for VFOA estimation in meetings, visual activity contextual information can be as effective as speaking context.

Index Terms— visual focus of attention, head pose, contextual information

1. INTRODUCTION

Meetings are an important aspect of human daily life. In companies most of the important decisions are taken during meetings. Nowadays due to the ubiquitous presence of recording devices such as microphones and cameras, researchers are investigating techniques for automatic meeting analysis. Research about automatic meeting analysis will allow the design of efficient tools for computer-enhanced human-to-human interaction.

Analyzing meetings requires the ability to understand the behaviors that are exhibited during human interaction. Among these behaviors, gaze plays an important role. In conversations, speakers use their gaze to specify their addressees. Listeners use their gaze to show their interest, and to request speaking turns [1]. Motivated by gaze shift studies that suggest very strong correlations between the gaze direction, defining people's visual focus of attention (VFOA), and their head pose, research has been conducted to use computer vision techniques to estimate VFOA from head pose in the case where gaze estimation directly from the eyes is impossible [2, 3, 4]. Yet, most of the studies about VFOA recognition have concentrated on situations where the people are seated during the entire meeting.

In general real life situations, participants do not remain at their seats during the entire meeting. They can stand up to use the white board or make presentations. Similar to Voit and Stiefelhagen [5], we denominate meetings where people move away from their seats *dynamic meetings*. VFOA estimation in dynamic meetings is more challenging than in the static meetings where people remain seated. The moving people, who have to be localized, increase the potential ambiguities between visual targets.

Investigations about VFOA recognition in the static meetings have shown that the use of contextual information such as speaking patterns or projection screen activity, have led to significant performance improvements [2, 6]. In dynamics meeting the use of context should also be of interest in reducing ambiguities between visual targets defined by similar head poses.

In this paper, we propose a probabilistic model for VFOA recognition from head pose in dynamics meetings making use of the dynamics of social context. Inspired by research about the strong correlation between a speaker's gestures and his speech ([7, 8]), we introduce the notion of conversation visual activity context, which accounts for people's tendencies to focus at the other persons because of their gestures or body motions. We compare VFOA recognition using visual activity context to the more classical speaking context [2, 6]. Our results show that visual activity context can be as effective as speaking context. Thus, allowing our system to deal with cases when only the video modality is available.

The remaining of this paper is organized as follows. Section 2 describes the task we address and the dataset we use for evaluation. In Section 3 we give the audio-visual features we use in the VFOA model. In Section 4, we present the model we propose. Section 5 describes the experiments we conducted to evaluate our model. Finally, in Section 6 we give conclusions.

2. DATASET AND TASK

Dataset description: The dataset we use for our study consist of 12 meetings of the AMI corpus[1], involving 4 people with real behaviors. They take notes, use laptops, make use of a white board, and a projection screen for presentations (see Fig.1). Twenty different persons were involved in the recordings. The meeting durations ranged from 15min to 35min, for a total of 5 hours. This makes this database the longest among those used for VFOA recognition studies [2, 3, 5]. With respect to the dynamic aspect of the AMI Corpus meeting, 23% of the time there was a person standing to make a presentation.

Task: Our goal is to estimate the VFOA of seated people in meetings. A person's focus can be any element of a finite set of visual targets that the person considers as interesting. For a meeting participant seated at seat k, we have identified the set of visual targets of interest $\mathcal{F}_k$ as: the 3 other participants $\mathcal{P}_k$ (e.g. for seat

This work was partly supported by the Swiss National Center of Competence in Research and Interactive Multimodal Information Management (IM2), and the European union 6th FWP IST Integrated Project AMIDA (Augmented Multi-Party Interaction with Distance Access, FP6-0033812)

[1] The dataset is available at `http://corpus.amiproject.org`

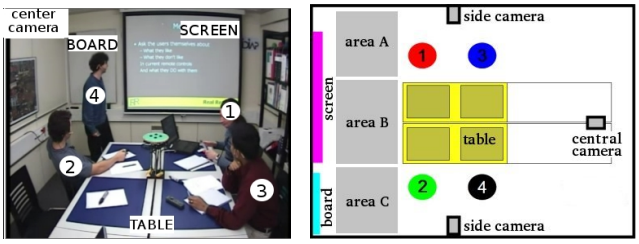

Fig. 1. Evaluation data recording setup: four people are having a meeting in a room equipped with a projection screen and a white board. Sometimes During the meeting, a person makes a presentation either seated or standing at the presentation areas A, B or C.

visual targets	people	table	board	screen	unfocused
proportion (%)	38.4	29.8	2.7	23.9	5.2

Table 1. Distribution of VFOA labels.

$1, \mathcal{P}_1 = \{$seat2,seat3,seat4$\}$), as well as a set of 4 other visual targets $\mathcal{O} = \{$table, white board, projection screen, unfocused $\}$. The latter target (unfocused) is used when the person is not visually focusing on any of the previously cited targets.

Dataset analysis: The meeting participants' VFOA were annotated based on the set of VFOA labels defined above. Tab.1 gives the VFOA statistics, where we have grouped the VFOA labels corresponding to participants into a single label 'people'. Looking at people only represents 39% of the data, while looking at the table or screen represents more than 50% of the VFOA proportion. These statistics show that classical face to face conversation dynamics where people mainly look at the speakers did not hold. Artifacts such as the table and the projection screen play an important role that has to be taken into account to understand the conversation dynamics in our meeting scenario.

3. MODEL OBSERVATIONS

The VFOA model we propose makes use of image and audio based observations: people's head locations and poses, speaking status (speaking or not), and visual activity status (visually active or not).

3.1. Head localization and pose

To estimate people's head location and pose we rely on a Bayesian formulation of the tracking problem solved through particle filtering techniques [9]. We applied our tracking method to track people when they are visible in the side cameras (see Fig.2a). At each time t, the tracker outputs the head locations in the image plane and the head poses θ_t (characterized by a pan and tilt angle) of people visible in the side view cameras.

3.2. Audio features

The audio features are extracted from close-talk microphones attached to each meeting participant. At each time t the speaking energy of participant k is thresholded to give his speaking status s_t^k which is 1 if participant k is speaking and 0 otherwise. Fig.3a) shows a sample sequence of the four person speaking statuses.

3.3. Motion features

As motion features, we used the thresholded absolute pixel differences between consecutive image frames. We denominate the motion energy of an image area, the proportion of the motion features

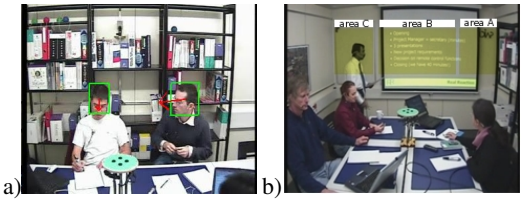

Fig. 2. a: Head location and pose tracking from one side camera view. b: Areas A, B,C in the central camera view for motion energy computation used for standing people localization.

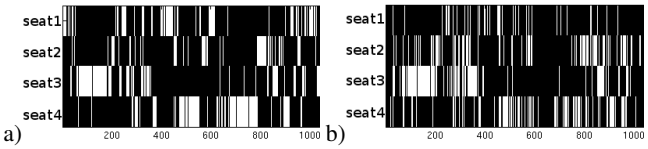

Fig. 3. a) Speaking statuses and b) motion statuses of the 4 meeting participants extracted from speaking and motion energy. Note the similarities between the speaking and motion patterns.

in that area. We made three uses of the motion energy: we detected slide changes, we localized people standing in the presentation areas, and finally we computed people's visual activity statuses.

Slide change detection: Slide changes correspond to a temporally localized peaks of the motion energy of the image area corresponding to the projection screen. To detect the slide changes we thresholded the motion energy of the projection screen image area. Given the slide change instant at t, we built a slide activity variable a_t that stores the time that has elapsed since the last slide change.

Localization of people: People are tracked from the side view camera when they are seated. The location x_t^k of a participant k is a discrete index which takes four values: seat k when he is seated, or the center of one the presentation areas A, B or C showed in Fig.1 and Fig.2b) when he is standing. This location is estimated by assuming that when people are away from their seats they are standing in one of the area A, B or C. Thus, when they are away from their seats to make presentations, they are localized from the central camera view using the motion energy. In the central view images, the motion energy corresponding to each of the standing area is computed and the location of the person standing is estimated as the area with the highest energy or his previous standing location when there is no energy is none of the standing area.

Visual activity statuses: To measure the visual activity of a person, we define the visual activity status v_t^k which is 1 when the person is visually active and 0 otherwise. When a person is visually active there is high motion energy at his location on the image. Given the four meeting participants' location, we estimate their visual activity statuses v_t by thresholding the motion energy corresponding to their locations. Fig.3b) shows a sample sequence of the four meeting participants activity statuses. The similarities between the speaking and the visual activity patterns in Fig.3 show that speaking and visual activities are strongly correlated.

4. VFOA MODELLING

The hidden state we want to estimate is the focus of a meeting participant k denoted f_t^k. His head pose θ_t^k is used as the observation for the hidden state. The people's locations $x_t = (x_t^1, x_t^2, x_t^3, x_t^4)$ are used as input variables defining the head pose observation model.

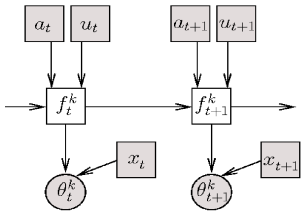

Fig. 4. VFOA graphical models.

A slide screen activity variable a_t, and an input variable u_t which, depending on the experimental conditions, is either people's speaking statuses or their visual activity statuses are used as the input variables driving the hidden state dynamics. According to independence assumptions reflected by graphical model displayed in Fig.4, the variables of our model have their joint posterior density function $p(f_{1:T}^k, \lambda|\theta_{1:T}^k, u_{1:T}, a_{1:T}, x_{1:T})$ defined as:

$$p(\lambda)\prod_{t=1}^{T} p(\theta_t^k|f_t^k, x_t)p(f_t^k|f_{t-1}^k, u_t, a_t) \quad (1)$$

where λ is the vector of model parameters, $p(\lambda)$ is a prior distribution on the model parameters, $p(\theta_t^k|f_t^k, x_t)$ is the observation likelihood relating the observed head poses to the VFOA states given the visual targets locations, and $p(f_t^k|f_{t-1}^k, u_t, a_t)$ is the state dynamics modeling the temporal VFOA states evolution subject to the screen activity and the speaking or the visual activities.

4.1. The observation models

When person k focuses at another person j, we defined the observation model as a Gaussian distribution:

$$p(\theta_t^k|f_t^k = j, x_t) = \mathcal{N}(\theta_t^k; \mu_{k,x_t^j}, \Sigma_k^j) \quad (2)$$

where μ_{k,x_t^j} is the Gaussian mean which models the mean head pose when the person at seat k looks at person j located at position x_t^j, and Σ_k^j is the covariance of the Gaussian. If the visual target j is an object (table, white board, projection), the observation model is defined as a Gaussian distribution $p(\theta_t^k|f_t^k = j, x_t) = \mathcal{N}(\theta_t^k; \mu_{k,j}, \Sigma_k^j)$. For the unfocused target, $p(\theta_t^k|f_t^k = \text{ unfocused }, x_t)$ is modeled as a uniform distribution.

4.2. The state Dynamics

We define the state dynamical model as follows:

$$p(f_t^k|f_{t-1}^k, a_t, u_t) \propto p(f_t^k|f_{t-1})p(f_t^k|a_t, u_t) \quad (3)$$

where $p(f_t^k|f_{t-1}^k)$ models the temporal transitions between focus states, $p(f_t^k|a_t, u_t)$ models the probability to observe a VFOA state given the slide activity, and the speaking or the motion activities.

VFOA temporal transitions: The VFOA temporal transition $p(f_t^k|f_{t-1}^k)$ role is to enforce temporal smoothness on the state sequence. It is modeled as a transition table with a high probability to remain in the same state and the remaining of the probability uniformly spread on the other states.

Contextual prior dynamics: It is well known in social sciences that in meetings, people's attention is attracted by activities such as a person speaking, a person gesturing, or slide changes on the projection screen [1]. Thus, we modeled the VFOA contextual prior as a probability table $p(f_t^k|u_t, a_t) = p(f_t^k|\mathcal{U}_t^k, a_t)$ where $\mathcal{U}_t^k$ denotes the set of orally or visually active persons at time t which are not person k,

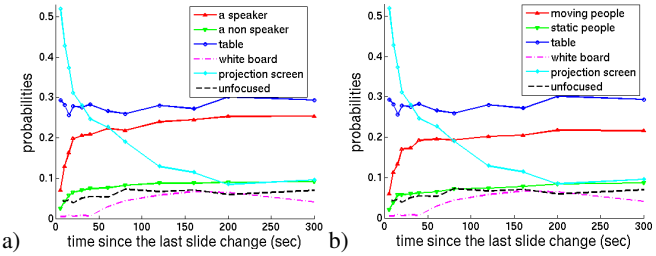

Fig. 5. Probabilities of focusing at visual targets given the time elapsed since the last slide change. In a) the probability of looking at a person is spread into looking at a speaker (red) or a listener (green). In b), the probability of looking at a person is spread between looking at a person visually active (red) or static (green).

and we further assumed that the VFOA of person k is independent of whether k is active or not. Using training data we learned the probability table $p(f_t^k|\mathcal{U}_t^k, a_t)$. Six cases are considered. The first case is $p(f_t^k = l|\mathcal{U}_t^k, a_t)$ where l is an active person ($l \in \mathcal{U}_t^k$). The second case corresponds to when l is a non-active person. The other 4 cases correspond to when l is not a person (table, white board, projection screen, unfocused). Fig.5 displays these probability tables when activity corresponds to speaking (Fig.5a) and when activity corresponds to moving (Fig.5b). These figures show that right after a slide change, the probability of looking at the projection screen is high and gradually decreases. Inversely the probability of looking at the people increases. The probability of looking at an active person, either speaking or visually active, is higher than the probability of looking at non-active people. Also, the probability of looking at a person speaking is higher than the probability of looking at a visually active person, indicating that there is a higher correlation between gaze and speaking behaviours than between gaze and visual activity behaviours. Finally, in Fig. 5 the probability of looking at the table can be considered uniform over the meetings.

4.3. Prior on the model parameters and model inference

Prior on the model parameters: The prior $p(\lambda)$ is a probability distribution over the model parameter values. In this paper we are only interested in the estimation of the values of the Gaussian means defining the observation models. This is motivated by the fact that the head pose defining gazing behaviours is more subject to variations due to people's personal way of gazing at visual targets [10]. We define the prior over each Gaussian mean as the Gaussian distribution $p(\mu_{k,x_k^j}|\tau, m_{k,x_k^j}, \Sigma_k^j)$:

$$\propto \exp -\tau(\mu_{k,x_t^j} - m_{k,x_t^j})^T \Sigma_k^j (\mu_{k,x_t^j} - m_{k,x_t^j})$$

with mean m_{k,x_t^j} and covariance matrix $\tau\Sigma_k^j$. The cognitive gaze model presented in [4], that relates people's gaze direction to their head pose can be used to predict m_{k,x_t^j}. The cognitive model assumes that a person k gazes at the direction $m_{k,x_t^j}^{gaze}$ through a head rotation m_{k,x_t^j} and eye in head rotation. The head rotation relates to the gaze direction $m_{k,x_t^j}^{gaze}$ through the linear relation $m_{k,x_t^j} = \kappa m_{k,x_t^j}^{gaze}$ where the parameter κ specifies the head pose contribution to the gaze rotation. The covariance matrix $\tau\Sigma_k^j$ is composed of two parameters, a scale factor $\tau > 0$ that is used to drive the parameter adaptation, Σ_k^j is the covariance matrix defined in Section 4.1.

Model Inference: Inference is conducted in two steps. First we apply an unsupervised maximum a posteriori (MAP) adaptation to

estimate the Gaussian means. Then, given the optimal model parameters, we use Viterbi decoding to find the optimal sequence of hidden states. Given a non-annotated test sequence, unsupervised (MAP) adaptation consists in finding the parameters μ_{k,x_l} and $\mu_{k,j}$ that allow for an optimal fit of the sequence by maximizing the posterior distribution in Eq.1. MAP adaptation can be conducted using an expectation-maximization (EM) algorithm [10].

5. RESULTS

We evaluate our VFOA model using the dataset described in Section 2. As performance measure we use the frame recognition rate (FRR) which is the percentage of video frames for which the VFOA is correctly classified. As an evaluation protocol, we adopted a leave one out approach: in turn, one meeting is left aside as test data, the remaining 11 meetings are used to learn the contextual priors.

Tab. 2 gives the results over the entire dataset. This table shows that the use of contextual activity is always beneficial. The best FRR is achieved when slide and visual activities are used as contextual cues which achieves an average (over 12 meetings, 4 seats) FRR of 53.2%. Using a pairwise T-test significance test, we tested whether the performances achieved by the method based on slide and speaking context is different from the results achieved by the other methods. The T-tests show that, at a p-values of 1% the performances of the methods based on speaking and visual activity context are not significantly different, but are significantly better than the other two methods (using pose only, or with slide context only). The correlation between speaking activity and visual activity only partly explains the similarities between the performances of the two contextual cues. On the overall dataset, the percentage of time a person speaking is visually active is 66%. This shows that visual activity captures a significant proportion of speaking activity. However, the percentage of time a person visually active is speaking which is 47%. This suggest that people may focus at a person who is not speaking but is visually active, for example when a listener is giving visual feedback as a head gesture.

Tab.2 also gives the VFOA recognition performances for the evaluation dataset split into static meetings (4 recordings) and dynamics meetings (8 recordings). For all experimental conditions the recognition rates are higher on the static meetings than on the dynamic meetings. The best performance of 55.2% over the static meetings is achieved when using slide and speaking context. This difference can be explained mainly by two factors. The first factor is that static meetings are more conversational. Thus, the speaking context is very informative. The second factor is that the VFOA recognition in the dynamic meetings is more challenging because the person standing either in front of the white board or projection screen increases the potential confusions between the visual targets: many visual targets will be defined by very similar head poses. The method based on slide and visual context achieves a FRR of 54.5% over the static meetings and 52.5% over the dynamics meetings. In dynamic meetings, visual activity context slightly outperforms the speaking context specially for the people seated at seat 4.

6. CONCLUSIONS

In this paper we proposed a model for VFOA recognition from head pose and multi-modal activity context. We introduced the concept of visual activity context for VFOA recognition which relates a person gazing behaviors to the other people's visual activities (gestures, hand and body motions). Our model, evaluated on a large database, achieved good performances for such a challenging task. In this paper we show that for VFOA recognition in meetings, the visual activity context we propose can be as efficient as the classical speaking context. This is an interesting finding knowing that in some experimental conditions such as outdoor scenes, visual activities might be easier to estimate than speaking activities. However in the situation both speaking and visual activity context are available, it might be worth investigating their combination into a single activity context to exploit the complementary information they convey.

experimental setup		seat1	seat2	seat3	seat4	mean
overall	slide-speaking	56.2	58.2	49	47.5	52.7
	slide-motion	**56.1**	**56.9**	**49.3**	**50.3**	**53.2**
	slide	53.7	55.2	43.5	48.6	50.2
	pose only	52.1	53.9	41.3	45.6	48.2
static	slide-speaking	**60.5**	**59.5**	**49.1**	**51.7**	**55.2**
	slide-motion	58.8	58.4	50.4	50.3	54.5
	slide	57.6	55.5	41.1	48.1	50.6
	pose only	58.2	56.7	40.1	44.7	49.9
dynamic	slide-speaking	54.1	57.5	49	45.5	51.5
	slide-motion	**54.7**	**56.2**	**48.8**	**50.3**	**52.5**
	slide	51.8	55	44.7	48.9	50.1
	pose only	49.1	52.5	41.9	46.1	47.4

Table 2. VFOA recognition performance over static and dynamic meetings.

7. REFERENCES

[1] S. Duncan Jr, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23(2), pp. 283–292, 1972.

[2] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Trans. on Neural Networks*, vol. 13(4), pp. 928–938, 2002.

[3] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase, "A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances," in *ICMI*, 2005.

[4] J-M. Odobez and S.O. Ba, "A cognitive and unsupervised MAP Adaptation approach to the recognition of focus of attention from head pose," in *ICME*, 2007.

[5] M. Voit and R. Stiefelhagen, "Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios," in *ICMI*, 2008.

[6] S.O. Ba and J.M. Odobez, "Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues," in *ICASSP*, 2008.

[7] P. Bull, "State of the art: non verbal communication," *The Psychologist*, 2001.

[8] J.B Bavelas, "Gestures as part of speech: Methodological implications," *Research on Language and Social Interaction*, 1994.

[9] S. O. Ba and J.-M. Odobez, "A Rao-Blackwellized mixed state particle filter for head pose tracking," in *ICMI Workshop on Multimodal Multiparty Meeting Processing*, 2005, pp. 9–16.

[10] S.O. Ba and J.-M. Odobez, "Recognizing human visual focus of attention from head pose in meetings," *IEEE Transaction on Systems, Man, and Cybernetics, Part B*, 2008.