Unsupervised Online Learning of Visual Focus of Attention

Stefan Duffner, Christophe Garcia

Unsupervised Online Learning of Visual Focus of Attention

Stefan Duffner and Christophe Garcia
Université de Lyon, CNRS
INSA-Lyon, LIRIS, UMR5205, F-69621, France
stefan.duffner@liris.cnrs.fr, christophe.garcia@liris.cnrs.fr

Abstract

In this paper, we propose a novel approach for estimating visual focus of attention in video streams. The method is based on an unsupervised algorithm that incrementally learns the different appearance clusters from low-level visual features extracted from face patches provided by a face tracker. The clusters learnt in that way can then be used to classify the different visual attention targets of a given person during a tracking run, without any prior knowledge on the environment and the configuration of the room or the visible persons. Experiments on public datasets containing almost two hours of annotated videos from meetings and video-conferencing show that the proposed algorithm produces state-of-the-art results and even outperforms a traditional supervised method that is based on head orientation estimation and that classifies visual focus of attention using Gaussian Mixture Models.

1. Introduction

Generally, the Visual Focus of Attention (VFOA) of a person denotes the target – an object or another person – the person is looking at, at a given point in time. The automatic estimation of the VFOA of a person from video is of great importance in many applications, such as human-computer interaction, advanced video-conferencing, smart meeting rooms, or human behaviour analysis in general, and much research has been conducted in this area in the past.

Principally, the VFOA of a person is defined by the person's eye gaze direction. Many studies about automatic gaze estimation from video exist [10, 20], but their use is mostly limited to close-up and near-frontal views of a person's face, for example in Human-Computer Interaction applications. In this paper, we will focus on more challenging scenarios where the camera is fixed at a few meters from the filmed persons and where the persons stay roughly at the same places, like in formal meetings or video-conferencing applications. Previous work on VFOA analysis in such open spaces has mostly been based on the estimation of head pose as a surrogate for gaze [18, 15, 2, 12, 3, 19, 21]. This is done either globally, *e.g*. by learning to classify image patches of the head at different angles based on low-level visual features or locally, *i.e*. by localising certain facial features and by geometrically and statistically inferring the global orientation (see [11] for a literature survey). Further, with video, head pose estimation can be included in a joint head and pose *tracking* algorithm [9, 2, 8, 14]. Early works of Stiefelhagen *et al*. [17], for example, used a Gaussian Mixture Model (GMM) on head pose angles to estimate VFOA. The model is initialised with k-means and further updated with an Expectation-Maximisation algorithm. They also showed that using the other participant's speaking status increases the VFOA performance. Otsuka *et al*. [12] proposed a method based on a Dynamic Bayesian Network that also analyses the group behaviour and detects certain conversational patterns. A GMM and Hidden Markov Model (HMM) approach for modelling and recognising VFOA was proposed by Smith *et al*. for people walking by an outdoor advertisement [16] and by Ba and Odobez for analysing meeting videos [3]. In the latter work, the authors also presented a MAP adaptation method to automatically adapt the VFOA model to the individual persons as well as a geometrical model (based on findings from [6, 7]) combining head orientation and eye gaze direction. Voit and Stiefelhagen [19] built on this geometrical model and presented VFOA recognition results on a dynamic dataset with multiple cameras. Recently, Ba and Odobez [4] extended their approach on VFOA estimation for meetings with a Dynamic Bayesian Network (DBN) that incorporates contextual information, like speaking status, slide change, and modelling conversation behaviour.

As the experimental results of these works show, head pose can be used effectively to estimate the VFOA of a group of people, *e.g*. in a meeting room, to a certain extend. However, there are certain drawbacks of this approach: for example, in uncontrolled environments it is difficult to estimate head pose reliably because it often requires a large amount of annotated training data of head appearances or shapes beforehand in order to model all the possible varia-

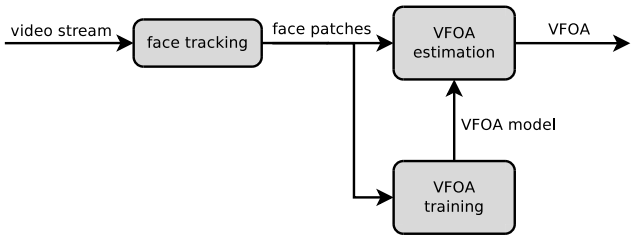


Figure 1. Process of the proposed approach

tions of a head and face among different people as well as for a given individual. These data are often not available, or too time-consuming to produce. Further, for accurate results, a relatively precise localisation of the head, the face, or facial features is crucial but challenging in unconstrained application scenarios.

In this paper, we propose a novel approach that alleviates these problems. Our algorithm, given a video stream from a single camera and the rough 2D position estimation of a person's head, incrementally learns to automatically extract the VFOA of the person *without explicitly estimating head pose or gaze and without any prior model of the head, face, the room configuration, or other external conditions*. The proposed method learns *on-the-fly* the different classes of targets in an unsupervised way directly from the low-level visual features. This means also that, as opposed to supervised algorithms, it will not assign labels to the different targets (*e.g.*'table', 'screen', 'person 1'). However, we will experimentally show that the proposed unsupervised approach is able to identify and estimate the (unlabelled) targets with higher accuracy than a classical supervised approach. The fact that no pre-trained model is needed makes this approach especially interesting for applications where the specific environment, as well as the configuration of the room and the filmed persons is not known a priori, and where an explicit training phase is not possible.

2. Principal approach

The overall process of our approach is illustrated in figure 1. First, a basic tracking algorithm is initialised and tracks a rectangular face region throughout the video stream. The image patch inside the tracked face region is extracted and visual features, Histograms of Oriented Gradients (HOG), are calculated to initialise the VFOA model at the first video frame and to update it at each subsequent frame. At the same time, the model that has been learnt so far is used to estimate the person's VFOA target (*e.g.* table, other person) from the face patch. Note that the learning is done *on-the-fly* and does not require any prior knowledge on head pose or room configuration. Another advantage of this approach is that the VFOA estimation is completely independent from the face tracking algorithm, which means that the standard tracking method used here can easily be replaced by a different, possibly more accurate, algorithm.

For tracking a face, we used a standard particle filter approach, very similar to the one presented in [13]. It provides a solution for the classical recursive Bayesian model, where, assuming we have the observations $\mathbf{Y}_{1:t}$ from time 1 to t, we estimate the posterior probability distribution over the state $\mathbf{X}_t$ at time t:

$$p(\mathbf{X}_t|\mathbf{Y}_{1:t}) = \frac{1}{C}\, p(\mathbf{Y}_t|\mathbf{X}_t) \times \int_{\mathbf{X}_{t-1}} p(\mathbf{X}_t|\mathbf{X}_{t-1}) p(\mathbf{X}_{t-1}|\mathbf{Y}_{1:t-1})\, \mathrm{d}\mathbf{X}_{t-1}\ , \quad (1)$$

where C is a normalisation constant. In our experiments, the state $\mathbf{X}_t \in \mathbb{R}^3$ is composed of the position and scale of the face. The state dynamics $p(\mathbf{X}_t|\mathbf{X}_{t-1})$ is defined by a first order auto-regressive model with Gaussian noise:

$$p(\mathbf{X}_t|\mathbf{X}_{t-1}) = \mathcal{N}(\mathbf{X}_{t-1}; 0, \Sigma_p)\ . \quad (2)$$

The observations likelihood is defined as:

$$p(\mathbf{Y}_t|\mathbf{X}_t) \propto \exp\left(-\lambda \sum_{r=1}^{9} \left(D_B^2[h_r^*, h_r(\mathbf{X}_t)]\right)\right)\ , \quad (3)$$

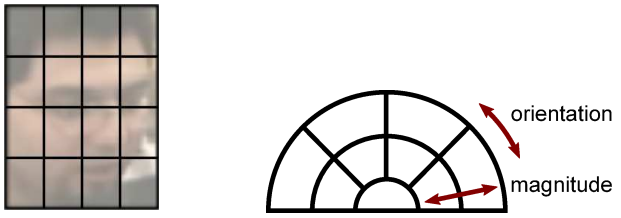
where λ is a constant, $h_r(\mathbf{X}_t)$ are HSV colour histograms extracted from a grid of $r = 9$ cells centred at $\mathbf{X}_t$, h_r^* is the reference histogram initialised from the face region in the first frame, and D_B is the Bhattacharyya distance. As in [13], the histogram bins for the H and S channels are decoupled from the V channel. Also the quantisation is applied at two different levels, *i.e.* 4 bins and 8 bins, to improve the robustness under difficult lighting conditions. This leads to an overall observation vector size of 828.

3. Learning Visual Focus of Attention

3.1. Overview

The tracking algorithm described in the previous section provides us at each video frame with a bounding box of the tracked face. The principal idea is to compute visual features from the underlying image patch described by this rectangle and directly classify the VFOA of the person at that point in time.

Here, we make use of the fact that the head orientation of a person in a given scenario (*e.g.* a meeting) is not completely random. There are a few stationary angles that a head takes most of the time. And in the scenarios we are dealing with, these head poses mostly correspond to VFOA targets. Using an unsupervised clustering algorithm we model these stationary head poses directly in the low-level feature space. Thus, we estimate VFOA from these low-level features without an error-prone intermediate step



(a) (b)

Figure 2. Visual features extraction for the VFOA model. a) HOG-like features are computed on a grid of 4×4 cells placed on the tracked face. b) To compute the histograms, gradient orientation is quantised into 4 bins (respectively 8 bins) and magnitude into 2 bins.

of head pose estimation. The proposed algorithm *incrementally* learns these clusters from the texture features extracted on-the-fly from the video stream, which makes it appropriate for applications that do not require (or even allow) prior training or adaption of different room configurations, lighting conditions, or other scene-dependant variables.

3.2. Learning Algorithm

Let us denote $\mathbf{O}_t \in \mathbb{R}^N$ the observed feature vector of a face image patch at time t. Here, we propose to use HOG-like features computed on an array of 4 by 4 non-overlapping cells (see Fig. 2(a)). For each cell, two normalised two-dimensional histograms of unsigned oriented gradients and magnitudes are computed using a specific quantisation scheme illustrated in Fig. 2(b). The gradient orientation is quantised in 4 bins and the magnitude in 2 bins. An additional bin (with no orientation) is used for very weak gradients (in the centre of the half circle in the diagram). Also, to improve the overall robustness and discriminative power, we compute *two* histograms at different quantisation levels for orientation: 4 and 8, and normalise each of them separately. Thus, the dimension N of the feature vector is: $16 \cdot (4 \cdot 2 + 1 + 8 \cdot 2 + 1) = 416$. One advantage of these histogram features is that they are relatively robust to small spatial shifts of the overall bounding box, which frequently occur with common face tracking methods.

The visual feature vectors $\mathbf{O}_t$ at time t are used to incrementally learn the VFOA classes. To this end, we propose a specific sequential k-means clustering algorithm with an adaptive number of clusters. The algorithm constructs a model of k clusters corresponding to the VFOA classes and described by their mean feature vectors μ_i and a global diagonal co-variance matrix $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$. Algorithm 1 summarises the main learning procedure. At each point in time t the observed feature vector $\mathbf{O}_t$ is computed, and the closest cluster c_t is determined using the normalised

Algorithm 1 Proposed incremental learning algorithm

$k = k_{ini}$
$\boldsymbol{\mu}_i = \mathbf{O}_0 \quad i = 1..k$
$n_i = 0 \quad i = 1..k$
$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{ini}$
for $t = 1$ to T **do**
  $c_t = \mathrm{argmin}_i(D(\mathbf{O}_t, \boldsymbol{\mu}_i))$
  $\bar{d} = \frac{2}{N(N+1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} (D(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j))$
  if $d(\mathbf{O}_t, c_t) > \theta \bar{d}$ **then**
    $k \leftarrow k + 1$
    $\boldsymbol{\mu}_k = \mathbf{O}_t$
  end if
  $n_c \leftarrow n_c + 1$
  $\boldsymbol{\mu}_{c_t} \leftarrow \boldsymbol{\mu}_{c_t} + \frac{1}{n_{c_t}} (\mathbf{O}_t - \boldsymbol{\mu}_{c_t})$
  $\boldsymbol{\mu}_{c_{t-1}} \leftarrow \boldsymbol{\mu}_{c_{t-1}} + \frac{\alpha}{n_{c_{t-1}}} (\mathbf{O}_t - \boldsymbol{\mu}_{c_{t-1}})$
  $\boldsymbol{\mu}_i \leftarrow \boldsymbol{\mu}_i + \frac{1}{n_i^2} (\mathbf{O}_t - \boldsymbol{\mu}_i) \quad i \in \Omega$
  incrementally update $\boldsymbol{\Sigma}$
end for

Euclidean distance:

$$D(\mathbf{O}_t, \boldsymbol{\mu}_i) = \sqrt{\sum_{j=1}^{N} \frac{(o_{t,i} - \mu_{i,j})^2}{\sigma_j^2 + \epsilon}}, \qquad (4)$$

with ϵ being a small constant avoiding division by zero.

Also, the mean distance $\bar{d}$ between each of the k clusters is calculated, and a new cluster is created if the distance of the current feature vector to the closest cluster is greater than $\theta \bar{d}$, where θ is a parameter of our algorithm (set to 2 in our experiments). Existing clusters are not removed or merged in this algorithm.

Eventually, we incrementally update the mean $\boldsymbol{\mu}_{c_t}$ of the closest cluster, the mean of the previous closest cluster $\boldsymbol{\mu}_{c_{t-1}}$ (with a lower factor $\alpha = 0.3$ here), and also the means in a "neighbourhood" Ω (*i.e.* with adjacent indexes) but with a quadratically decreasing update factor. Updating the previous closest cluster enforces a certain temporal continuity of the model. Updating the neighbouring means (inspired by self-organising maps) ensures a certain topographical coherence in the feature space and limits outliers in the initial learning phase.

Finally, the global covariance matrix $\boldsymbol{\Sigma}$ is incrementally updated using the current feature vector $\mathbf{O}_t$. At each time t, the algorithm classifies the observed features $\mathbf{O}_t$ of a face into one of the k clusters: c_t, and, as we will show in the following experimental results, the learnt classes correspond to a large degree to specific targets of VFOA.

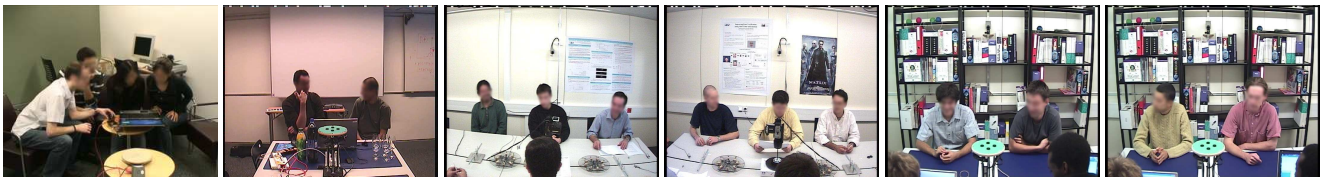


Figure 3. Example frames from the three datasets that have been used for evaluation. First two: TA2 dataset, middle two: PETS 2003 dataset, and last two: IHPD dataset. (Faces have been blurred manually.)

4. Experiments

4.1. Data

We evaluated the proposed approach on three public datasets from different scenarios, each containing a certain number of persons sitting around a table and filmed roughly from the front (see Fig. 3). The VFOA targets are different for each dataset, due to the scenario and the layout of the room. The three datasets are the following:

TA2[1][5]**:** in this set there are two videos from two different rooms where people communicated over a video-conferencing system and performed a shared task on a laptop in front of them. In the first video there are four persons and in the second there are two, and the defined VFOA targets are the table, the camera, and the other persons, resulting in 5 targets for the first video and 3 for the second. For each person, the targets over 7 500 frames (5 minutes), 30 minutes in total, have been annotated.

IHPD[2][1]**:** this dataset consists of the "meeting" part of the Idiap Head Pose Database. It contains eight meeting recordings with four persons, where one video shows two participants behind a table. The annotated VFOA targets here are the table, the slide screen, and the other persons (the white-board target has not been used here). 110 040 frames ($\sim$ 1 hour 13 minutes) with VFOA annotation have been used in total for the evaluation.

PETS 2003[3]**:** this dataset contains two videos from a formal meeting of six participants (scenario D), where each video shows 3 of the persons roughly from the front (similar to IHPD). Here, the VFOA targets that have been annotated for each participant are the other five participants. The provided VFOA annotation for the 6 persons and 47 000 frames ($\sim$30 minutes) in total has been used.

Table 1 summarises the properties of these datasets.

Annotation has been done manually and frame-by-frame, where frames with ambiguous visual focus and transition phases have not been annotated.

[1]https://www.idiap.ch/dataset/ta2
[2]https://www.idiap.ch/dataset/headpose
[3]http://www.cvg.rdg.ac.uk/slides/pets.html

dataset	number of videos	number of persons	VFOA targets	annotated frames
TA2	2	6	5 / 3	7500
IHPD	8	16	5	110040
PETS 2003	2	6	5	47000
total	12	28	5/3	164540 (110 min.)

Table 1. The three datasets and annotation used for evaluation.

4.2. Evaluation

As our algorithm is unsupervised, we don't have the actual estimated VFOA targets (the labels) that we can directly compare to the ground truth. For evaluation purposes, *i.e.* after running our method on the whole video, we therefore assign to each cluster the VFOA target that maximises VFOA accuracy. That means, we assume that we know which target each cluster corresponds to. We believe that this is not a very restrictive assumption, as the labels could be assigned in a separate processing step, for example by incorporating a more general discriminative classifier trained beforehand.

As an accuracy measure, we used the Frame-based Recognition Rate (FRR) of the VFOA for all the videos and averaged it over each datasets and over several runs. The FRR is simply the proportion of frames with correctly recognised VFOA:

$$FRR = \frac{N_c}{N_t} \,, \tag{5}$$

where N_c is the number of correct classifications, and N_t is the total number of annotated video frames. As our algorithm is learning the VFOA model *incrementally*, we need to account for a certain training phase, which we don't include in the evaluation. We used 8 000 ($\sim$ 5 min.) training frames in the beginning of the videos (not annotated), and evaluated the FRR on the following sequence with annotation. This length has been chosen in order to have enough training data for *all* the VFOA targets of a person, as sometimes a target is focused for the first time only after several minutes.

We compared the proposed approach to a variant that uses a fixed number of clusters. We also compared it to a classical supervised approach, that uses a specific face detection and tracking algorithm, a head pose estimator as in

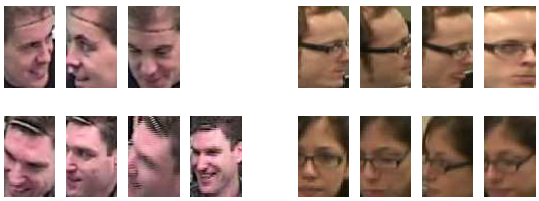


Figure 4. Four examples of face images that are closest to the corresponding learnt cluster centres. Upper-left and lower-left: from IHPD dataset. Upper-right and lower right: from TA2 (first video).

	TA2	IHPD	PETS 2003	average
supervised	58.8	49.4	26.5	46.5
fixed $k = 5$	45.7	46.8	35.2	44.1
variable k	**61.5**	**51.0**	**45.0**	**52.0**

Table 2. Average frame-based recognition rate of VFOA (in %).

[14], and Gaussian Mixtures Models (GMM) to model different VFOA targets in terms of head pose pan and tilt angles as in [17, 3]. In this approach, the head pose model is trained beforehand in a supervised way, and the GMM parameters have been partly trained and partly defined manually.

4.3. Results

First, we will show some qualitative results on the clustering that is obtained on some of the videos. To illustrate this, we saved for each tracking run the face image regions corresponding to the feature vectors that were closest to the cluster centres. Fig. 4 shows some examples. We can see that the images come from different head poses mostly corresponding to real VFOA targets. Clearly, some targets might not be captured by the model, as in the top right example of Fig. 4 (corresponding to the left-most person in the top-left image of Fig. 3) because the three other persons are almost seated in the same gaze direction. In the bottom right example, two clusters (corresponding to the second and fourth image) have been created for the same VFOA target: the table. Apart from these errors, the results mostly make sense.

Additionally, we evaluated the proposed algorithm using the metric and the three datasets described above. The tracking algorithm has been initialised manually with a bounding box around the face. Table 2 shows the FRR for each dataset, for the proposed method with fixed cluster size $k = 5$, variable cluster size, and the classical supervised approach described above. One can see that the proposed approach outperforms the supervised method with an average FRR of 52% compared to 46%. The variant with fixed cluster size is generally worse than the baseline method, except for the PETS 2003 dataset. This is probably due to the fact that the cluster centres are all initialised at the beginning and not when they are actually required as in the variable case. These results are comparable or superior to those published in the literature, although the evaluation protocols are not exactly the same due to the unsupervised and incremental nature of our method. Note that we do not include any temporal modelling (*e.g.* a HMM) or contextual information in the VFOA estimation process as in other existing work. This may additionally improve the overall performance.

The overall tracking algorithm, implemented in C++, runs at around 80-90 fps on a 3.6GHz processor for a 720×576 video (face size $\sim 50 \times 80$), where around 11% of CPU time is spent on feature extraction for VFOA, i.e. gradient computation on the whole image, and less than 1% on the VFOA learning and classification.

5. Conclusion

We presented an algorithm that incrementally, and in an unsupervised way, learns a VFOA model directly from low-level gradient histogram features extracted from face images coming from a tracking algorithm. In a meeting room or video-conferencing setting, the proposed method is able to automatically learn the different VFOA targets of a person without any prior knowledge about the number of persons or the room configuration. By assigning a VFOA label to each cluster, a posteriori, we evaluated the VFOA recognition rate for three different datasets and almost 2 hours of annotated data. The obtained results are very promising and show that this type of unsupervised learning can outperform traditional supervised approaches.

Future work will investigate different types of visual features, cluster merging, and the possibility of automatically assigning meaningful labels to the clusters. Also, we will study the generalisation capability of the algorithm to unseen videos (same room with different persons) as this might enable a broader range of practical applications.

References

[1] S. O. Ba and J.-M. Odobez. Evaluation of multiple cue head pose estimation algorithms in natural environements. In *Proceedings of ICME*, pages 1330–1333, 2005.

[2] S. O. Ba and J.-M. Odobez. A Rao-Blackwellized mixed state particle filter for head pose tracking. In *Proceedings of the ICMI Workshop on Multi-modal Multi-party Meeting Processing (MMMP)*, pages 9–16, Trento, Italy, 2005.

[3] S. O. Ba and J.-M. 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, Feb. 2009.

[4] S. O. Ba and J.-M. 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, Jan. 2011.

[5] S. Duffner, P. Motlicek, and D. Korchagin. The TA2 database – a multi-modal database from home entertainment. In *International Conference on Signal Acquisition and Processing*, Feb. 2011.

[6] E. G. Freedman and D. L. Sparks. Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys. *Journal of Neurophysiology*, 77:2328–2348, 1997.

[7] M. Hayhoe and D. Ballard. Eye movements in natural behavior. *TRENDS in Cognitive Sciences*, 9(4):188–194, 2005.

[8] O. Lanz and R. Brunelli. Joint bayesian tracking of head location and pose from low-resolution video. In *Multimodal Technologies for Perception of Humans*, pages 287–296, 2007.

[9] L. Lu, Z. Zhang, H. Shum, Z. Liu, and H. Chen. Model and exemplar-based robust head pose tracking under occlusion and varying expression. In *Proceedings of the IEEE Workshop on Models versus Exemplars in Computer Vision (CVPR-MECV)*, Dec. 2001.

[10] C. Morimoto and M. Mimica. Eye gaze tracking techniques for interactive applications. *CVIU*, 98:4–24, 2005.

[11] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: a survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626, Apr. 2009.

[12] K. Otsuka and J. Yamato. Conversation scene analysis with dynamic bayesian network based on visual head tracking. In *Proceedings of the International Conference on Multimedia and Expo*, pages 949–952, 2006.

[13] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *Proceedings of ECCV*, pages 661–675, 2002.

[14] E. Ricci and J.-M. Odobez. Learning large margin likelihoods for realtime head pose tracking. In *Proceedings of ICIP*, pages 2593–2596, Nov. 2009.

[15] M. Siracusa, L.-P. Morency, K. Wilson, J. Fisher, and T. Darrell. A multi-modal approach for determining speaker location and focus. In *ICMI*, pages 77–80, 2003.

[16] K. Smith, S. O. Ba, J.-M. Odobez, and D. Gatica-Perez. Tracking the visual focus of attention for a varying number of wandering people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(7):1212–1229, July 2008.

[17] R. Stiefelhagen, J. Yang, and A. Waibel. Modeling focus of attention for meeting indexing based on multiple cues. *IEEE Transactions on Neural Networks*, 13(4):928–938, July 2002.

[18] R. Stiefelhagen and J. Zhu. Head orientation and gaze direction in meetings. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*, 2002.

[19] M. Voit and R. Stiefelhagen. Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In *Proceedings of ICMI*, pages 173–180, 2008.

[20] J.-G. Wang and E. Sung. Study on eye gaze estimation. *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, 32:332–350, 2002.

[21] H. Zhang, L. Toth, W. Deng, J. Guo, and J. Yang. Monitoring visual focus of attention via local discriminant projection. In *Proceeding of the 1st ACM international conference on Multimedia Information Retrieval*, 2008.