Visual Focus of Attention estimation with unsupervised incremental learning

Stefan Duffner, Christophe Garcia

Visual Focus of Attention estimation with unsupervised incremental learning

Stefan Duffner and Christophe Garcia

***Abstract*—In this paper, we propose a new method for estimating the Visual Focus Of Attention (VFOA) in a video stream captured by a single distant camera and showing several persons sitting around table, like in formal meeting or video-conferencing settings. The visual targets for a given person are automatically extracted on-line using an unsupervised algorithm that incrementally learns the different appearance clusters from low-level visual features computed from face patches provided by a face tracker without the need of an intermediate error-prone step of head-pose estimation as in classical approaches. The clusters learnt in that way can then be used to classify the different visual attention targets of the 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.**

***Index Terms*—Unsupervised learning, pattern clustering, image sequence analysis**

I. 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 (see Fig. 1). The automatic estimation of the VFOA of a person from video is of great importance in many applications, such as human-computer interaction, video-conferencing, smart meeting rooms, or human behaviour analysis in general, and much research has been conducted in this area in the past years.

A. Related Work

Principally, the VFOA of a person is defined by the person's eye gaze direction. Many studies about automatic gaze estimation from video exist [1], [2], [3], [4], [5], [6], 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. Other works [7], [8], [9] rely on the fusion of information from several cameras. But often the spatial camera configuration is very constrained or a preceding calibration step is required, which can be difficult or even impossible depending on the application and environment. Also depth sensors, like Kinect, have been used for head pose and eye gaze estimation [6]. Although, their precision depends highly on the distance of the person from the sensor, this is an interesting direction for future research beyond the scope of this work. In this paper, we will focus on (non-intrusive) scenarios where a single 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 (as illustrated in Fig. 1).

S. Duffner and C. Garcia are with the Université de Lyon, CNRS, INSA-Lyon, LIRIS, UMR5205, F-69621, France

Fig. 1. Graphical illustration of VFOA estimation and the type of setting that is used. Targets 1 to 4 are persons, 5 corresponds to the table.

Previous work on VFOA analysis in such open spaces has mostly been based on the estimation of head pose as a surrogate for gaze [10], [11], [12], [13], [14], [15], [16], [8], [17], [18], [19], [20], [21], [22], [23]. 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 [24], [25] and by geometrically and statistically inferring the global orientation, or a combination of the two [22] (see [26] for a literature survey). However, these algorithms mostly require the person(s) to face the camera more or less and be rather close to it in order to have a relatively high image resolution of the face. Using video, head pose estimation can be included in a joint head and pose *tracking* algorithm [27], [15], [28], [29]. Early works of Stiefelhagen and Zhu [30], 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. Note that, in this paper, we will concentrate on methods that are relying on *visual* information, although there are previous works that use audio, actions or or types of cues to infer the VFOA [30], [14], [31], [32]. Otsuka and Yamato [16] 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*. [33] for people walking by an outdoor advertisement and by Ba and Odobez [18] for analysing meeting videos. 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 [34], [35]) combining head orientation and eye gaze direction. Voit and Stiefelhagen [8], [36] built on this geometrical model and presented VFOA recognition results on a dynamic dataset with multiple cameras. Recently, Ba and Odobez [37] 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. Dong *et al*. [9] proposed an approach also based on a DBN which is similar to ours in the fact that they recognise VFOA by comparing tracked face image patches with a set of clusters modelling the face appearance for each attention target. However, the difference to our approach is that the clusters in their algorithm are trained before the tracking and in a supervised way. Thus, the number of targets and the targets itself are known in advance.

The recent work of Benfold *et al*. [20] is similar to ours in that they also perform *unsupervised* training on head images in order to determine where people look at in a given video. However, their approach is not incremental (although they claim that it could be extended) and needs an initial training of prior models using hand-labelled ground plane velocities and gaze directions of persons in a given video. They do not extract VFOA but head orientation (using a given number of classes), and they apply their approach to video surveillance data where they take advantage of people moving, which is different from our indoor scenario. On the one hand, the advantage of their probabilistic model – a conditional random field (CRF) – is that a more powerful discriminative head pose classifier can be learnt taking into account several hidden variables (walking speed, angle velocities etc). On the other hand, the complexity of learning and inference is increased, and the model is also independent from the head tracking as opposed to our approach that allows for a purely sequential and joint inference.

B. Motivation

As experimental results of these previous 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 variations 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 head pose estimation results, a relatively precise localisation of the head, the face, or facial features – commonly called face alignment – is crucial but challenging in unconstrained application scenarios.

Another difficulty in automatic VFOA estimation is to determine the number of semantic visual targets for a given person in a video and to map them to given head pose or eye gaze angles. A preceding supervised training step is commonly performed on separate video data, and in some approaches the model (*e.g*. a GMM) is adapted on-line to a given video. However, it is desirable to avoid this scene-dependant training step or in some applications it might even be impossible. Further, the subsequent model adaptation can in many cases not cope with a different number of focusing targets or when the persons' locations differ too much from those in the training data.

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.

C. Contribution

In [38], we introduced a basic algorithm for unsupervised incremental learning of VFOA clusters from low-level features. We have improved this work in the following way:

- The VFOA recognition has been changed from a simple frame-by-frame classification into a Hidden Markov Model (HMM) with full Bayesian inference on an automatically learnt observation model and state transition distribution.
- The VFOA estimation process has been integrated in the Particle Filter framework that now performs a *combined* face and VFOA tracking.
- The cluster management of the incremental learning algorithm has been extended, notably to allow also for merging two clusters into one during the training.
- Finally, more extensive experiments have been performed varying different parameters of our approach, showing an improved classification accuracy compared to [38] as well as a classical supervised VFOA recognition approach.

The outline of the paper is as follows. Section II briefly describes the overall VFOA estimation procedure. In section III, the overall face and VFOA tracking algorithm is explained. Section IV describes the unsupervised incremental learning algorithm. Finally, experimental results are presented in section V, and in section VI we draw our conclusions.

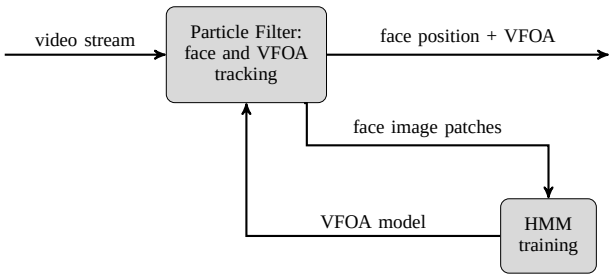


Fig. 2. Principal procedure of the VFOA learning and tracking approach.

II. Overall Procedure

The principal procedure of our approach is illustrated in figure 2. 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 are computed to initialise the VFOA model at the first video frame and to update it at each subsequent frame during the training phase (see section IV). An incremental clustering algorithm on these low-level features is used to learn face appearances corresponding to attention targets of the person. At the same time a matrix modelling the transition probabilities between the different targets is learnt, and together with the clusters forms a continuous HMM. Note that this incremental learning is done *on-the-fly* and does not require any prior knowledge on head pose or room configuration.

After a given number of iterations (a couple of minutes from the beginning of a video) the training phase stops and the Particle Filter continues to jointly track face position and VFOA of a person using the learnt HMM model, *i.e.* the transition probabilities and the face clusters. (see section III).

Note also that we assume that the VFOA targets are not moving, and our *incremental* algorithm learns the underlying "static" distributions in a sequential manner as opposed to *on-line learning* algorithms which are also sequential but can further adapt to non-stationary distributions.

In order to facilitate understanding, before describing the main contribution of the paper, *i.e.* the unsupervised VFOA learning, we will first explain the underlying tracking framework in the following section.

III. Face and VFOA Tracking

For tracking the face position and VFOA of a person, we used the Sequential Monte Carlo algorithm, commonly known as Particle Filter (*c.f.* [39], [29], [40]). 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 $\mathbf{X}_t = (\hat{\mathbf{X}}_t, v)$ is composed of the state of the face $\hat{\mathbf{X}}_t = (x, y, s)$ with x, y being its position and s being its bounding box scale factor, as well as the current VFOA target index $v \in 1..V$.

The dynamics of the face state $p(\hat{\mathbf{X}}_t|\hat{\mathbf{X}}_{t-1})$ are defined by a first-order auto-regressive model with Gaussian noise:

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

The dynamics of the discrete VFOA target index v are defined by transition probability matrix

$$\boldsymbol{A} := [a_{ij}], \quad i,j = 1..V \quad \text{with}$$
$$a_{ij} := p(v_t = j | v_t = i) \quad (3)$$

being the transition probability from VFOA target i to j. The co-variance matrix $\Sigma_p = \mathrm{diag}(\sigma_{px}, \sigma_{py}, \sigma_{ps})$ of the auto-regressive model is fixed, whereas the matrix $\boldsymbol{A}$ is learnt online during the tracking of a person in a given video stream. Details on how $\boldsymbol{A}$ is learnt are presented in section IV-B.

The observations likelihood is defined as the product of a colour likelihood and texture likelihood::

$$p(\mathbf{Y}_t|\mathbf{X}_t) = p(\mathbf{Y}_t^C|\mathbf{X}_t) p(\mathbf{Y}_t^T|\mathbf{X}_t)\ , \quad (4)$$

where the colour likelihood is used to track the position and size (x, y, s) of the face bounding box, and the texture likelihood is mainly used to track the VFOA target v. We define:

$$p(\mathbf{Y}_t^C|\mathbf{X}_t) \propto \exp\left(-\lambda_1 \sum_{r=1}^{9} \left(D_C^2[h_r^*, h_r(\mathbf{X}_t)]\right)\right)\ , \quad (5)$$

where λ_1 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_C is the Bhattacharyya distance. As in [39], 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 colour observation vector size of $9 \cdot (8 \cdot 8 + 8 + 4 \cdot 4 + 4) = 828$.

The texture likelihood is defined similarly:

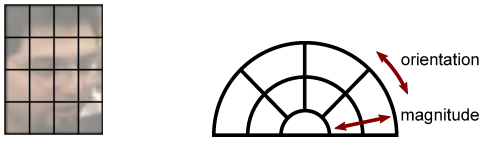
$$p(\mathbf{Y}_t^T|\mathbf{X}_t) \propto \exp\left(-\lambda_2 \sum_{r=1}^{16} \left(D_T[\boldsymbol{\mu}_{r,v}, \mathbf{t}_r(\mathbf{X}_t)]\right)\right)\ , \quad (6)$$

where λ_2 is a constant, $\mathbf{t}_r(\mathbf{X}_t)$ are Histograms of Oriented Gradients (HOG) (see description below) extracted (similarly to h_r) from a grid of 16 cells (indexed by r) centred at $\mathbf{X}_t$, and $\boldsymbol{\mu}_{r,v}$ are the reference histograms corresponding to the VFOA target index v in $\mathbf{X}_t$. The overall texture model is composed of a set of N-dimensional clusters with means $\boldsymbol{\mu}_{r,i}$ where each cluster $i \in 1..V$ corresponds to a VFOA target. D_T is the normalised Euclidean distance:

$$D_T(\boldsymbol{\mu}_{r,i}, \mathbf{t}_r(\mathbf{X}_t)) = \sqrt{\sum_{j=1}^{N} \frac{(t_{r,j}(\mathbf{X}_t) - \mu_{r,i,j})^2}{\sigma_j^2 + \epsilon}}\ , \quad (7)$$

with ϵ being a small constant avoiding division by zero.

The feature vectors $\mathbf{t}_r(\mathbf{X}_t)$ constitute the visual observations used for recognising the VFOA targets of a person in a video by means of $p(\mathbf{Y}_T|\mathbf{X}_t)$. They are computed on a 4



(a) (b)

Fig. 3. Visual feature extraction for the VFOA model. a) HOG 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.

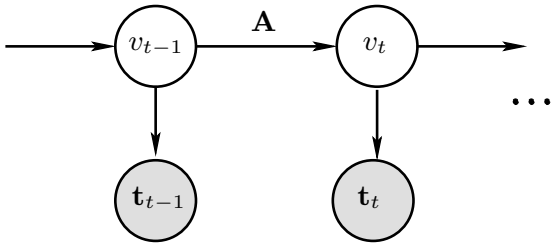


Fig. 4. The Hidden Markov Model used to estimate the hidden discrete variable v (the VFOA target) from the observations $\mathbf{t}_t$ (feature vectors) using the learnt transition probability matrix $\boldsymbol{A}$.

by 4 grid of non-overlapping cells on a face image patch as illustrated in Fig. 3(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. 3(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.

IV. Learning VFOA

The VFOA model can be regarded as a dynamic HMM estimating the hidden variable v, the VFOA target index, from the observed features $\mathbf{t}_r(\mathbf{X}_t)$, illustrated in Fig. 4. It consists of two main parts. First, the data model that is used for the likelihood computation in Eq. 6 and that contains the k cluster means $\boldsymbol{\mu}_i$ and a global co-variance matrix $\boldsymbol{\Sigma}$, and second, the matrix $\boldsymbol{A}$ (Eq. 3) defining the transition probabilities from one cluster to another. All, these parameters are learnt on-line during the training phase, and used subsequently in the tracking (*c.f.* section III). After training, the learnt parameters $\boldsymbol{\mu}_i$, $\boldsymbol{\Sigma}$, and $\boldsymbol{A}$ of the HMM are used in the Particle Filter framework explained in the previous section to jointly estimate the posterior probability of the state $\mathbf{X}_t$ at each time step. In the following, the training procedures are described in more detail.

A. VFOA clustering

The visual feature vectors $\mathbf{t}_r(\bar{\mathbf{X}}_t)$ computed on the image region corresponding to the mean state of the current distribution 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 $\boldsymbol{\mu}_{r,i}\,(i = 1..k)$ and a global diagonal co-variance matrix $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$. For better readability, in the following notation, we drop the indexes for the cell r and the time step t, denoting the current cluster means as $\boldsymbol{\mu}_i$ and the current feature vector as $\mathbf{t}$. Algorithm 1 summarises the main learning procedure. At each time step the observed feature vector $\mathbf{t}$ is computed, and the closest cluster c is determined using the normalised Euclidean distance (Eq. 7). Also, the mean distance $\bar{D}_T$ 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 $\theta_c \bar{D}_T$, where θ_c is a parameter of our algorithm (set to 2 in our experiments). Then, the mean vector $\boldsymbol{\mu}_c$ of the closest cluster as well as the global covariance matrix $\boldsymbol{\Sigma}$ are incrementally updated using the current feature vector $\mathbf{t}$. In the previous version [38] of the algorithm, the closest cluster means of

Algorithm 1 Incremental VFOA cluster learning algorithm

$k = k_{ini}$
$\boldsymbol{\mu}_i = \mathbf{t}_0 \quad i = 1..k$
$n_i = 0 \quad i = 1..k$
$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_{ini}$
for $t = 1$ to T **do**
  $c = \mathrm{argmin}_i(D_T(\mathbf{t}, \boldsymbol{\mu}_i))$ ▷ get closest cluster
  $\bar{D}_T = \frac{2}{N(N+1)} \sum_{i=1}^{k} \sum_{j=i+1}^{k} (D_T(\boldsymbol{\mu}_i, \boldsymbol{\mu}_j))$
  if $D_T(\mathbf{t}, \boldsymbol{\mu}_c) > \theta_c \bar{D}_T$ **then** ▷ add new cluster
    $k \leftarrow k + 1$
    $n_k = 1$
    $\boldsymbol{\mu}_k = \mathbf{t}$
  else
    $n_c \leftarrow n_c + 1$ ▷ update closest cluster
    $\boldsymbol{\mu}_c \leftarrow \boldsymbol{\mu}_c + \frac{1}{n_c}(\mathbf{t} - \boldsymbol{\mu}_c)$
  end if
  incrementally update $\boldsymbol{\Sigma}$
  for each cluster pair (i,j) **do** ▷ merge clusters
    if $D_T(\boldsymbol{\mu}_{c_i}, \boldsymbol{\mu}_{c_j}) < \theta_d \bar{D}_T$ **then**
      $\boldsymbol{\mu}_i = (n_i \boldsymbol{\mu}_i + n_j \boldsymbol{\mu}_j)/(n_i + n_j)$
      $n_i = n_i + n_j$
      remove cluster j
      $k \leftarrow k - 1$
    end if
  end for
end for

the previous time steps and neighbouring clusters were also updated. However, with the integration of the model into the Particle Filter framework, this did not improve the overall performance significantly. Thus, we removed this step. Finally, pairs of clusters are merged together if the distance of their means are below the threshold $\theta_d \bar{D}_T$ (with $\theta_d = 0.01$ in our experiments). At each time step, the algorithm classifies the observed features $\mathbf{t}$ from the mean state of a face into one of the k clusters: c, and, as we will show in the following experimental results, the learnt classes correspond to a large degree to specific targets of VFOA.

B. VFOA Transition Model

The transition probability matrix $\boldsymbol{A}$ of equation 3 is learnt on-line during the training phase at the same time as the cluster centres. The main procedure is the following. The visual feature vectors $\mathbf{t}$ of the image patch corresponding to the current mean state are extracted, and the closest cluster c_t according to the normalised Euclidean distance (Eq. 7) is computed. Then, the transition probabilities $a_{c_{t-1},j} := p(v = j | v = c_{t-1})$ are linearly updated, using the following equation:

$$a_{c_{t-1},j} = \gamma \mathbb{1}_{j=c_t} + (1-\gamma) a_{c_{t-1},j} \qquad \forall j \in 1..k, \quad (8)$$

where $\mathbb{1}_x$ denotes the indicator function, and the constant $\gamma = 0.001$. Thus, the transition probability from c_{t-1} to c_t is increased, and from c_{t-1} to any other cluster j is decreased. Also, a new row and column is added if a new cluster is created and inversely if a cluster is removed. At the end of each iteration, the row c_{t-1} that has been updated is normalised to sum up to 1.0. Algorithm 2 summarises the overall procedure. In many cases, the learnt transition matrix will have high values on the diagonal (staying in the same state most of the time) and low values elsewhere. Of course, this depends on the dynamics of the scene. In our formal meeting setting, people are interacting frequently and changing their attention targets quite often. Thus, this seems not to be a limitation. But even in more static settings (*e.g.* a person giving a talk), this model is still appropriate. And we can observe this with less active persons in some videos in our experiments. Clearly, transitions with very low probabilities can still be "triggered" if the observation likelihood of the target state is high enough. Nevertheless, to prevent extreme cases where a transition probability becomes zero and thus a state inaccessible, in our experiments, we set a very small lower boundary (10^{-3}) for the transition probabilities.

Algorithm 2 Incremental learning of the transition matrix

initialise $\boldsymbol{A}$ to uniform distribution: $a_{ij} = \frac{1}{k} \quad i,j \in 1..k$
for $t = 1$ to T **do**
  adapt the size of $\boldsymbol{A}$ to $k \times k$
  $c_t = \text{argmin}_i \, D(\mathbf{t}, \boldsymbol{\mu}_i)$
  $a_{c_{t-1},j} = \gamma \mathbb{1}_{j=c_t} + (1-\gamma) a_{c_{t-1},c_t}$
  normalise row c_{t-1} to sum up to 1.0
end for

Fig. 5. Example frames from the three datasets that have been used for evaluation. Top TA2 dataset, middle: PETS 2003 dataset, and bottom IHPD dataset. (Faces have been blurred artificially in this figure).

V. Evaluation

A. 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. 5). Note that we do not evaluate the accuracy of face or head pose tracking, as this is not the main contribution of the paper. Our main goal is to correctly estimate the VFOA of a person, which requires a robust face tracking system. The VFOA targets are different for each datasets, due to the scenario and the layout of the room. The three datasets are:

TA2[1][41]**:** 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. The defined VFOA targets are the table, the camera, and the other persons, *i.e.* 5 targets for the first video and 3 for the second. For each person, 7 500 frames (5 minutes 30 minutes in total, have been annotated.

IHPD[2][42]**:** this dataset consists of the "meeting" part of the Idiap Head Pose Database. It contains 8 meeting recordings with four persons, where each video shows two participants behind a table. The annotated targets are the table, the slide screen, and the other persons (the whiteboard 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.

[1]https://www.idiap.ch/dataset/ta2
[2]https://www.idiap.ch/dataset/headpose

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 I
THE THREE DATASETS AND ANNOTATION USED FOR 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 (∼30 minutes) in total has been used.

Table I 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.

B. Qualitative evaluation

First, we will show some qualitative results on the clustering that is obtained on some of the videos. Fig. 6 illustrates the result of the proposed on-line clustering algorithm (Alg. 1) for six different persons and videos. Each point represents a 2D projection of the 416-dimensional gradient feature vectors $\mathbf{t}_r(\bar{\mathbf{X}}_t)$ extracted from the mean state at time t (after the training phase). The linear embedding has been performed by applying multi-dimensional scaling with Euclidean distance measure on the whole data. Different colours (and point shapes) correspond to different labels produced by a k-Nearest Neighbour classifier using the normalised Euclidean distance, Eq. 7, and the learnt cluster means $\boldsymbol{\mu}_i$ as references. Note that the clusters means have been trained during the training phase, *i.e.* the first few minutes of a video. There are two difficulties that we want to emphasise here: first, the *test* data might be distributed slightly differently (*e.g.* the person's main focus changes), and second, the training data arrives sequentially and in a non-random order, *i.e.* a person's focus changes slowly and might be static for long periods. Note also, that the 2D projection of all points suggests that clustering is difficult in many cases, like in the top middle, bottom left, and bottom right example where cluster centres and frontiers are not so clear. Nevertheless, the output of the algorithm looks reasonable, apart from the bottom right example.

In order to experimentally verify if the learnt clusters correspond to different VFOA targets, we saved for each tracking run the face image regions corresponding to the feature vectors $\mathbf{t}_r(\bar{\mathbf{X}}_t)$ that were closest to the cluster centres $\boldsymbol{\mu}_i$. Fig. 7 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. 7 (corresponding to the left-most person in the top-left image of Fig. 5) because the three

[3]http://www.cvg.rdg.ac.uk/slides/pets.html

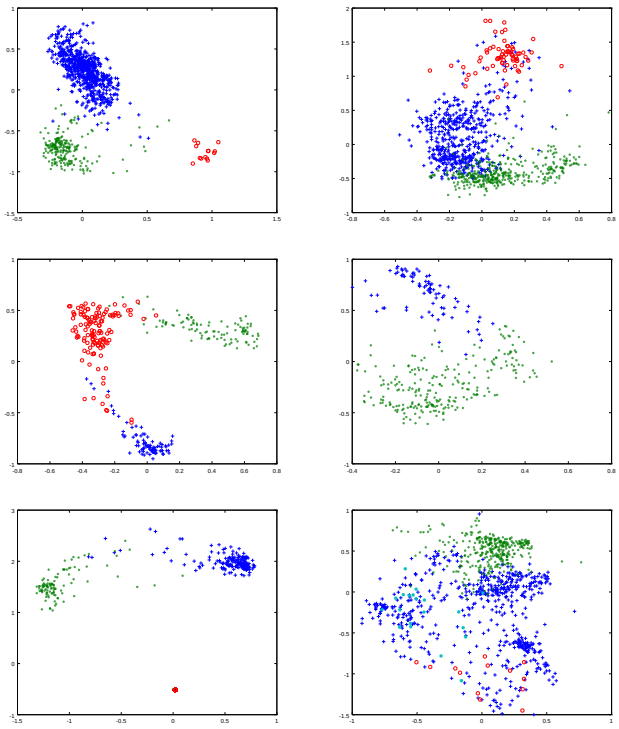


Fig. 6. Visualisation of the clustering of low level features produced by the proposed incremental learning algorithm (Alg. 1) for some examples. (Best viewed in colour.) From left to right, top to bottom: TA2 room 1, TA2 room 2, 2 x IHPD, PETS, and the last example shows a poor clustering result for one TA2 example.

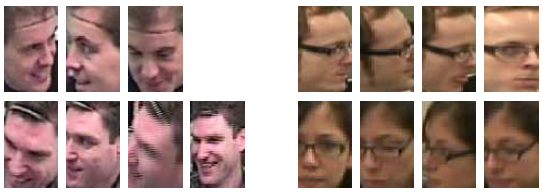


Fig. 7. Four examples of face images that are closest to the corresponding learnt cluster centres. Upper-left and lower-left: from IHPD dataset. Upper-right and lower right: from TA2 (first video).

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.

Finally, Fig. 8 visualises the VFOA tracking over time for two example videos. At each time step t, the red circles show the ground truth targets, and the blue crosses the output of the tracking algorithm, *i.e.* the v component of the mean state $\bar{\mathbf{X}}_t$. Some targets have not been learnt by the initial clustering and thus are not recognised. This is because these targets are less frequent (apart from target 4 in the bottom example), and thus, potential clusters would be sparsely represented and difficult to estimate. In the bottom example, from the IHPD dataset, targets 2 and 4 corresponds to persons that are sitting very close to each other. Here, the person's VFOA is mostly controlled by his eye gaze, and the head pose is almost the same. Thus, only target number 2 is recognised.

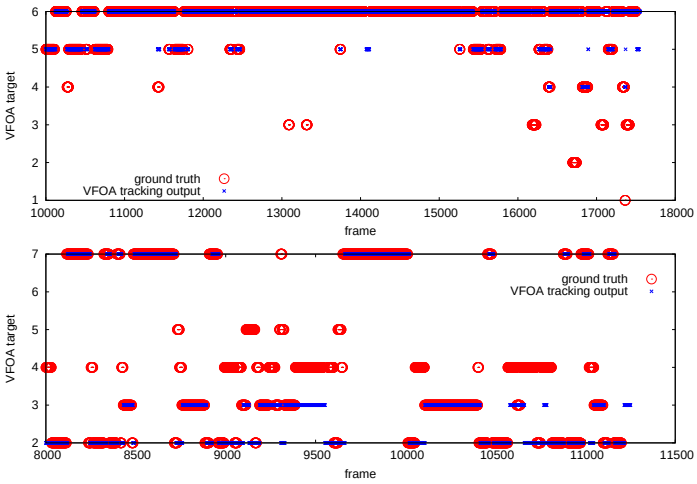


Fig. 8. Visualisation of VFOA classification over time for two example videos from TA2 and IHPD (best viewed in colour.) Red circles show the ground truth VFOA targets and blue crosses the output of a k-Nearest Neighbour classifier with the proposed clustering algorithm. Clusters 1, 2, 3 from the top example and 4 and 5 from the bottom have not been found by the initial on-line learning algorithm, and thus are not recognised.

Note that, as our algorithm is unsupervised, we do not have the actual estimated VFOA targets (*i.e.* meaningful labels) that we can directly compare to the ground truth. For evaluation purposes, after running our method on a whole video, we therefore assign to each cluster the target that maximises VFOA accuracy, *i.e.* we assume that we know which target label 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.

C. Quantitative evaluation

Additionally, we quantitatively evaluated the complete VFOA tracking algorithm on the three datasets described above by initialising it manually with a bounding box around the face and measuring the VFOA recognition accuracy by assigning a label to each cluster as described above, and counting the Frame-based Recognition Rate (FRR) of the VFOA for all the videos and averaging it over each dataset and over several runs. The FRR is simply the proportion of frames with correctly recognised VFOA:

$$FRR = \frac{N_c}{N_t} \ , \tag{9}$$

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 do not 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. A more detailed analysis on this parameters is given below (Fig. 9).

	TA2	IHPD	PETS 2003	average
$k = 4$	0.7242	**0.5248**	0.4597	0.5696
$k = 5$	0.693	0.4816	**0.4747**	0.5498
$k = 6$	0.6219	0.4777	0.4639	0.5212
variable k	**0.7663**	0.5163	0.4437	**0.5754**

TABLE II
VFOA RECOGNITION RATE OF THE PROPOSED ALGORITHM WITH FIXED AND VARYING NUMBER OF CLUSTERS.

	TA2	IHPD	PETS 2003	average
Euclidean (D'_T)	0.5713	0.5013	0.4308	0.5011
Bhattacharyya (D''_T)	0.608	0.5047	0.4033	0.5053
normalised Euclidean (D_T)	**0.7915**	**0.5282**	**0.4668**	**0.5955**

TABLE III
VFOA RECOGNITION RATE WITH DIFFERENT DISTANCE MEASURES REPLACING D_T IN EQ. 7.

In our first experiment, we studied the influence of using a *variable* number of clusters k, *i.e.* dynamic cluster creation and merging in Alg. 1, compared to using a *fixed* k. Table II shows the results for the three datasets. It can be seen that the k with the highest average FRR depends on the dataset. However, on average, a *variable* cluster number gives the best average FRR over all datasets. Therefore, we used a variable k for the following experiments (as presented in Alg. 1).

Further, we replaced the normalised Euclidean distance D_T of Eq. 7 with the simple Euclidean distance:

$$D'_T(\boldsymbol{\mu}_{r,i}, \mathbf{t}_r(\mathbf{X}_t)) = \sqrt{\sum_{j=1}^{N} (t_{r,j}(\mathbf{X}_t) - \mu_{r,i,j})^2} \ , \tag{10}$$

as well as the Bhattacharyya distance:

$$D''_T(\boldsymbol{\mu}_{r,i}, \mathbf{t}_r(\mathbf{X}_t)) = \sqrt{1 - \sum_{j=1}^{N} \sqrt{t_{r,j}(\mathbf{X}_t)\mu_{r,i,j}}} \ . \tag{11}$$

The results summarised in table III show that the *normalised* Euclidean distance D_T (Eq. 7) largely outperforms the other distances in terms of the FRR on all the dataset.

In another set of experiments, we varied the number of iterations for our incremental training algorithm in order to understand the impact of this parameter on the overall VFOA recognition rate. Figure 9 shows the results. As for the other experiments, the VFOA recognition rates are averaged over several runs and always computed on the same number of frames (irrespective of the number of training iterations). It can be seen that using very few training iterations, *i.e.* below 1000-2000, deteriorates the performance for all datasets. And beyond around 4000 iterations, the performance stays relatively stable.

Finally, we compared the proposed approach with three other approaches:

- **supervised**: a state-of-the-art supervised approach, that uses a specific face detection and tracking algorithm, a head pose estimator as in [29], and Gaussian Mixtures Models (GMM) to model different VFOA targets in terms of head pose pan and tilt angles as in [30], [18]. In this

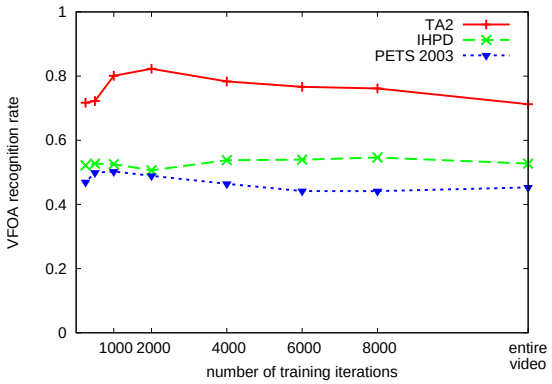


Fig. 9. VFOA recognition rate for the three datasets with varying number of training iterations.

	TA2	IHPD	PETS 2003	average
supervised [37]	0.59	0.49	0.26	0.4489
no PF	0.7663	0.5163	0.4437	0.5754
PF, fixed TM	0.6577	0.5235	0.4379	0.5397
PF, learnt TM	**0.7915**	**0.5282**	**0.4668**	**0.5955**

TABLE IV
VFOA RECOGNITION RATE OF THE PROPOSED ALGORITHM WITH AND WITHOUT PARTICLE FILTER INTEGRATION, AND WITH FIXED OR LEARNT TRANSITION PROBABILITY MATRIX $\boldsymbol{A}$ COMPARED TO A CLASSICAL SUPERVISED APPROACH.

approach, the head pose model is trained beforehand in a supervised way, and the GMM parameters have been partly trained and partly defined manually.

- **no PF**: a variant of the proposed approach that does not integrate the VFOA estimation into the Particle Filter tracking, *i.e.* v is not included in the state vector and is estimated frame-by-frame by a k-NN classifier using the feature vectors $\mathbf{t}(\bar{\mathbf{X}}_t)$ of the mean state and the cluster means $\boldsymbol{\mu}_i$, as in our previous work [38].
- **PF, fixed TM**: a variant of the our approach with Particle Filter VFOA tracking and a fixed, uniform transition probability matrix $\boldsymbol{A}$.
- **PF, learnt TM**: the proposed approach as presented in this paper, *i.e.* with Particle Filter VFOA tracking and learnt transition matrix $\boldsymbol{A}$

Table IV shows the average FRR for these different approaches. One can see that the proposed approach outperforms the supervised method with an average FRR of $\sim 60\%$ compared to $\sim 45\%$. Tracking the VFOA with a Particle Filter, as opposed to a frame-by-frame estimation, and learning the transition probability matrix on-line also improves the recognition performance on the three tested datasets. 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 contextual information like speaking status or other external events 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 $\sim 80-90$fps on a 3.6GHz processor for a 720×576 video, where around $\sim 11\%$ of CPU time is spent on feature extraction for VFOA (gradient computation on the whole image), and less than 1% on the VFOA learning and classification.

VI. CONCLUSION

We presented a VFOA tracking algorithm that incrementally, and in an unsupervised way, directly learns a VFOA model from low-level features extracted from a stream of face images coming from a tracking algorithm. The VFOA estimation is based on an HMM whose parameters are learnt incrementally and which is tightly integrated into a global Particle Filter framework that is used for face tracking. 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 two 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, more dynamic scenarios with moving persons and the possibility of automatically assigning meaningful labels to the clusters. Also, it would be interesting to 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. Finally, a combination of supervised and unsupervised learning might be beneficial and improve the overall performance of VFOA recognition. Especially when there are attention targets that are focused on only rarely and thus might not be captured by the proposed clustering algorithm.

REFERENCES

[1] J.-G. Wang and E. Sung, "Study on eye gaze estimation," *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics*, vol. 32, pp. 332–350, 2002.
[2] J.-G. Wang, E. Sung, and R. Venkateswarlu, "Eye gaze estimation from a single image of one eye," in *Proceedings of the International Conference on Computer Vision*, 2003, pp. 136–143.
[3] C. Morimoto and M. Mimica, "Eye gaze tracking techniques for interactive applications," *Computer Vision and Image Understanding*, vol. 98, pp. 4–24, 2005.
[4] U. Weidenbacher, G. Layher, P. Bayerl, and H. Neumann, "Detection of head pose and gaze direction for human-computer interaction," *Perception and Interactive Technologies*, vol. 4021, pp. 9–19, 2006.
[5] J. J. Magee, M. Betke, J. Gips, M. R. Scott, and B. N. Waber, "A human-computer interface using symmetry between eyes to detect gaze direction," *IEEE Transactions on Systems, Man, and Cybernetics. Part A*, vol. 38, no. 6, pp. 1248–1261, 2008.
[6] K. A. Funes Mora and J.-M. Odobez, "Gaze estimation from multimodal kinect data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Workshop on Gesture Recognition*, Jun. 2012.
[7] M. Voit and R. Stiefelhagen, "Tracking head pose and focus of attention with multiple far-field cameras," in *Proceedings of the IEEE Conference on Multimodal Interfaces (ICMI)*, 2006, pp. 281–286.
[8] ——, "Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios," in *Proceedings of the ICMI*, 2008, pp. 173–180.
[9] L. Dong, H. Di, L. Tao, G. Xu, and P. Oliver, "Visual focus of attention recognition in the ambient kitchen," in *Proceedings of the Asian Conference on Computer Vision*, 2009, pp. 548–559.
[10] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing," in *Proceedings of the ACM Multimedia*, 1999.

[11] S. Langton, R. Watt, and I. Bruce, "Do the eyes have it? cues to the direction of social attention," *Trends in cognitive sciences*, vol. 4, no. 2, pp. 50–59, Feb. 2000.

[12] R. Stiefelhagen, J. Yang, and A. Waibel, "Tracking focus of attention for human-robot communication," in *IEEE-RAS International Conference on Humanoid Robots - Humanoids*, Tokyo, Japan, 2001.

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

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

[15] 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)*, 2005, pp. 9–16.

[16] 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*, 2006, pp. 949–952.

[17] H. Zhang, L. Toth, W. Deng, J. Guo, and J. Yang, "Monitoring visual focus of attention via local discriminant projection," in *Proceeding of the International Conference on Multimedia Information Retrieval*, 2008.

[18] 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*, vol. 39, no. 1, pp. 16–33, Feb. 2009.

[19] C. Peters, S. Asteriadis, and K. Karpouzis, "Investigating shared attention with a virtual agent using a gaze-based interface," *Journal on Multimodal User Interfaces*, vol. 3, no. 1–2, pp. 119–130, 2009.

[20] B. Benfold and I. Reid, "Unsupervised learning of a scene-specific coarse gaze estimator," in *Proceedings of the International Conference on Computer Vision*, 2011.

[21] C. Chen and J.-M. Odobez, "We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video," in *Proceedings of the Computer Vision and Pattern Recognition*, 2012.

[22] S. Asteriadis, K. Karpouzis, and S. Kollias, "Visual focus of attention in non-calibrated environments using gaze estimation," *International Journal of Computer Vision*, vol. 107, no. 3, pp. 293–316, 2013.

[23] L. Bazzani, D. Tosato, M. Cristani, M. Farenzena, G. Pagetti, G. Menegaz, and M. V., "Social interactions by visual focus of attention in a three-dimensional environment," *Expert Systems*, vol. 30, no. 2, pp. 115–127, 2013.

[24] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proceedings of the Computer Vision and Pattern Recognition*, 2013.

[25] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[26] E. Murphy-Chutorian and M. M. Trivedi, "Head pose estimation in computer vision: a survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, Apr. 2009.

[27] 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 the IEEE Workshop on Models versus Exemplars in Computer Vision (CVPR-MECV)*, Dec. 2001.

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

[29] E. Ricci and J.-M. Odobez, "Learning large margin likelihoods for real-time head pose tracking," in *Proceedings of the International Conference on Image Processing*, Nov. 2009, pp. 2593–2596.

[30] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Transactions on Neural Networks*, vol. 13, no. 4, pp. 928–938, Jul. 2002.

[31] J. Ou, L. M. Oh, S. R. Fussell, T. Blum, and J. Yang, "Predicting visual focus of attention from intention in remote collaborative tasks," *IEEE Transactions on Multimedia*, vol. 10, no. 6, pp. 1034–1045, 2008.

[32] S. O. Ba, H. Hung, and J.-M. Odobez, "Visual activity context for focus of attention estimation in dynamic meetings," in *Proceedings of the ICME*, 2009, pp. 1424–1427.

[33] 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*, vol. 30, no. 7, pp. 1212–1229, Jul. 2008.

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

[35] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *TRENDS in Cognitive Sciences*, vol. 9, no. 4, pp. 188–194, 2005.

[36] M. Voit and R. Stiefelhagen, "3D user-perspective, voxel-based estimation of visual focus of attention in dynamic meeting scenarios," in *Proceedings of the ICMI-MLMI*, 2010.

[37] 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*, vol. 33, no. 1, pp. 101–116, 2011.

[38] S. Duffner and C. Garcia, "Unsupervised online learning of visual focus of attention," in *Proceedings of the IEEE Conference on International Conference on Advanced Video and Signal-Based Surveillance*, 2013.

[39] P. Pérez, C. Hue, J. Vermaak, and M. Gangnet, "Color-based probabilistic tracking," in *Proceedings of the European Conference on Computer Vision*, 2002, pp. 661–675.

[40] S. Duffner and J.-M. Odobez, "A track creation and deletion framework for long-term online multiface tracking," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 272–285, Jan. 2013.

[41] S. Duffner, P. Motlicek, and D. Korchagin, "The TA2 database – a multimodal database from home entertainment," in *International Conference on Signal Acquisition and Processing*, Feb. 2011.

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

Stefan Duffner received his PhD degree from University of Freiburg, Germany, in 2008, after doing his dissertation research at Orange Labs in Rennes, France, on face analysis with statistical machine learning methods. He then worked 4 years at Idiap Research Institute in Martigny, Switzerland, in the field of computer vision and multi-object tracking. As of today, Stefan Duffner is an associate professor in the IMAGINE team of the LIRIS research laboratory at the National Institute of Applied Sciences (INSA) of Lyon, France. His main research interests lie in machine learning for computer vision, and more specifically, on-line visual object tracking, and face image analysis.

Christophe Garcia is Full Professor at INSA de Lyon, and head of the IMAGINE research team of the LIRIS laboratory. His current technical and research activities are in the areas of deep learning, pattern recognition and computer vision. He holds 17 industrial patents and has published more than 140 articles in international conferences and journals. He has served in more than 30 program committees of international conferences and is an active reviewers in 15 international journals where he co-organized several special issues. He has served as an associate editor of the International Journal of Visual Communication and Image Representation (Elsevier), Image and Video Processing (Hindawi) and Pattern Analysis and Application (Springer-Verlag).