Multiple-gaze geometry: Inferring novel 3D locations from gazes observed in monocular video

Ernesto Brau[1 [0000−0003−0380−8630]], Jinyan Guan[1 [0000−0002−9721−6267]], Tanya Jeffries[2], and Kobus Barnard[2 [0000−0002−8568−9518]]

[1] CiBO Technologies, Cambridge MA 02141, USA
{ebrau,jguan}@cibotechnologies.com
[2] University of Arizona, Tucson AZ 85711, USA
tanyasjeffries@email.arizona.edu,kobus@cs.arizona.edu

Abstract. We develop using person gaze direction for scene understanding. In particular, we use intersecting gazes to learn 3D locations that people tend to look at, which is analogous to having multiple camera views. The 3D locations that we discover need not be visible to the camera. Conversely, knowing 3D locations of scene elements that draw visual attention, such as other people in the scene, can help infer gaze direction. We provide a Bayesian generative model for the temporal scene that captures the joint probability of camera parameters, locations of people, their gaze, what they are looking at, and locations of visual attention. Both the number of people in the scene and the number of extra objects that draw attention are unknown and need to be inferred. To execute this joint inference we use a probabilistic data association approach that enables principled comparison of model hypotheses. We use MCMC for inference over the discrete correspondence variables, and approximate the marginalization over continuous parameters using the Metropolis-Laplace approximation, using Hamiltonian (Hybrid) Monte Carlo for maximization. As existing data sets do not provide the 3D locations of what people are looking at, we contribute a small data set that does. On this data set, we infer what people are looking at with 59% precision compared with 13% for a baseline approach, and where those objects are within about 0.58m.

Keywords: 3D temporal scene understanding, 3D gaze estimation, monocular video, discovering objects, MCMC, model selection

1 Introduction

Observing people interacting with their environment can provide clues about its 3D structure. Facets of this that have been studied within computer vision include inferring functional objects as "dark matter" [64], ground plane paths [30], and modeling human-object interactions for understanding events and participants from RGB-D video [61]. 3D representations enable answering questions

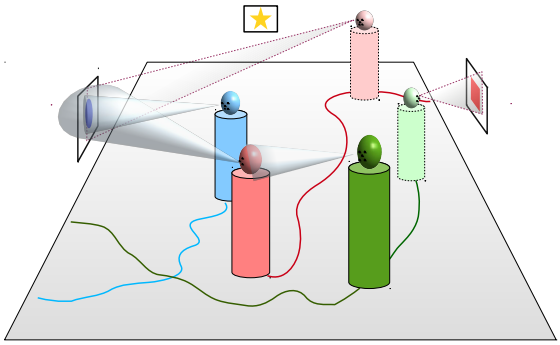

Fig. 1. Temporal 3D scene understanding through joint inference of people's locations, their head posse, and locations of what they're looking at. The gaze cones of the red person for the current (red) and previous times (faded red) intersect to help localize a target in 3D on the left wall. The hypothesis that they are looking at the same object from two different views makes this analogous to stereo vision. The blue person adds a third view. Furthermore, the hypothesis that the green person is looking at the red person enriches our understanding of the scene, and can help improve both the estimate of the green person's head pose as well as the location of the red person.

that are awkward or not accessible with 2D representations. For example, one might want to ask if there are paths that can be taken that are not visible to security cameras. In this paper, we present a system that infers 3D locations that people look at, including ones not visible to the camera, from monocular, uncalibrated video. For example, we can infer the 3D location of an interesting poster that draws people's gazes by observing the people passing by (Fig. 1).

To this end, we develop a fully 3D Bayesian modeling approach that represents where people are, their head poses (thus approximate gaze directions), and what 3D location they are looking at, which might be one of the other persons that we are tracking, or an interesting location that attract people's visual attentions in a scene. Our model further embodies the camera parameters of an assumed stationary monocular video camera, so that we can infer it rather than rely on having calibrated cameras.

Our joint inference approach is motivated by the following observations: **1)** the 3D locations of what people might be looking at can help estimate gaze direction and therefore head pose; **2)** other people in the scene are possible targets of visual attention, and if we are tracking them in 3D, joint inference of their location and gazes from others should be beneficial; and **3)** scenes often contain likely locations of visual attention (e.g., a visually interesting poster), and multiple spatio-temporal gaze cones can help pinpoint them in 3D analogously with multiple views (Fig. 1). We also make use of the following observations from Brau et al. [13] regarding tracking of people walking on a ground plane: **1)** 3D representation simplifies handling occlusions (which become evidence instead of confounds); **2)** 3D representation allows for a meaningful prior on velocity (and here, head turning angular velocity); and **3)** one can infer camera parameters

jointly with the scene, as people walking tend to maintain fixed height, and thus are like calibration probes that transport themselves to different depths.

We specify the joint probability of the latent model and the association of person detections across frames (§3). The data association implies a hypothesis for the number of people in the scene at each point in time. To compare models of differing dimensions in a principled way, we approximately marginalize out all the continuous model parameters. These include the locations of each person, their gaze angles, and the locations of the static points drawing visual attention that we are trying to discover from gazing behavior. We compute these approximate marginals using MCMC sampling to maximize the distribution, and then apply the Laplace approximation. We combine this with multiple MCMC sampling strategies to explore the space of models (§4).

Because our goals are new, we contribute a modest data set with the 3D locations of what participants are looking at, which is not available in other data sets with people walking about (see §5 for further discussion). In the contributed data set, participants recorded what they were looking at while they were walking around, and we established the ground truth 3D locations for all targets (people and other objects) using ground truth 2D detections (§6).

Our contributions include: **1)** operationalizing the observation that multiple gaze angles estimated from head pose can be used to learn 3D locations that people look at; **2)** extending the approach proposed by Brau et al. [13] to include head pose, a walking direction prior, and a more efficient sampling approach; **3)** joint inference of head pose and 3D location of what people are looking at while walking; **4)** inferring who is looking at whom or what (both anonymously defined); and **5)** a new data set for what people are looking at while they walk around, and where those objects or people are in 3D.

2 Related work

Multiple target tracking (MOT). Despite significant progress, multiple-target tracking remains a challenge due to issues such as noisy and complex evidence, occlusion, abrupt motion, and an unknown number of targets. This work is in the tracking-by-detection paradigm [13, 69, 31, 3, 44, 9, 4, 46, 17, 37, 66, 54]. Typically, these approaches first acquire the image locations of people a video sequence, and then find the tracks of each underlying target by solving the data association problem and inferring the target locations. Both 2D and 3D models have been used to represent the underlying targets. Effectively working in 2D requires explicit modeling of occluded targets (e.g., [69, 37]). Conversely, 3D models can treat occlusions and smooth motion naturally [28, 13].

Head pose estimation. There is a rich history in methods to estimate head pose from single images (e.g., [12, 22, 11, 39, 26, 33, 34, 21, 25, 38]. In video, information flow between frames has been exploited by a number of researchers (e.g., [70, 6, 65, 57]). More similar to us is model-based tracking methods that fit a 3D model to the tracked features across a video (e.g., [62, 32, 63, 45, 56]). Head and body pose have also been estimated jointly via correlations between

outputs of body pose and head pose classifiers [14, 15]. In contrast, we model this coupling through a joint distribution on 3D body and head poses.

Head pose is a strong cue for visual focus of attention (VFoA) recognition which has potential applications such as measuring the attractiveness of advertisements or shop displays in public spaces as well as analyzing the social dynamics of meetings. Much research in VFoA focuses on dynamic meeting scenarios, where people usually sit around meeting tables while being video recorded by multiple cameras [5, 7, 8, 19, 42, 43, 51–53, 58, 59]. Most of these methods exploit context-related information from speech and motion activity and the potential VFoA is a predefined discrete set with known locations. In addition, the number of people in the scene is fixed and they are considered to be seated in typically known locations, which makes sense given the application.

VFoA estimation has also been considered in surveillance settings in the context of understanding behavior [10, 27, 48, 49], where, so far, visual attention has been limited to image coordinates, and one person at a time. However, Benfold and Reid[10] use a camera calibrated to the ground plane to estimate a visual attention map representing the amount of attention received by each square meter of the ground in a town center scene. Similar to us, they identify interesting regions in the scene based on the inferred visual attention map. However, while the map can be projected into the video to visualize it, 3D location is not inferred.

Another application of estimating VFoA is human-robot interaction scenarios, which involves both person-to-person and robot-to-person interactions [36, 47, 67]. Approaches in this domain often assume known head poses (orientations and locations) of the targets (persons, robots, and objects). For example, Massé et al. proposed a switching Kalman filter formulation to jointly estimate the gaze and the VFoA of several persons from observed head poses and object locations [36]. In addition, they also assume the number of persons and objects are known and remain constant over time. In contrast, we propose simultaneously inferring the number of the targets and their locations in the scene while estimating their VFoAs using image evidence.

3 Statistical model

Figure 2 shows our generative statistical model for temporal scene understanding using probabilistic graphical modeling notation. The scene consists of **multiple people** moving on the ground plane throughout the video. At each frame, each person may have their visual attention on another person or on one of several static objects that are located in 3D space. We model the **visual focus of attention** and the **static objects** explicitly. At each frame, each person may also generate a detection box, and the **data association** groups these detection boxes by person (or noise). Finally, we model the **camera**, which projects the scene onto the image plane, generating the observed data.

We place prior distributions on each of the model variables mentioned above. Similarly, for each type of data we use, we have a likelihood function that cap-

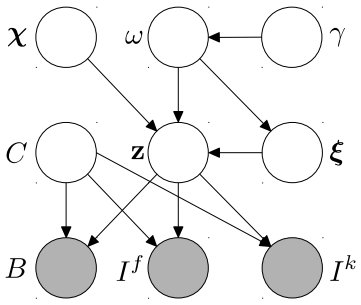

Fig. 2. Generative graphical model for temporal scene understanding. We use bold font for aggregate variables (e.g., $\mathbf{z}$ represents state vectors for each person for each frame). The data association, ω, specifies the number of people and which detections (body, face) are associated with them. ω depends on hyper-parameters collectively denoted by γ (§3.1). $\boldsymbol{\chi}$ is the set of static 3D points that people look at. The *visual focus of attention* (VFoA), $\boldsymbol{\xi}$, of each person is, for each frame, either one of these 3D points or another person. The temporal scene $\mathbf{z}$ consists of the 3D state (location, size, head pose) of each person at each frame (§3.2). $\mathbf{z}$ projects onto 2D to create *model frames* via the camera C, generating person detections, B, optical flow, I^f, and face landmarks, I^k (§3.4).

tures its dependence on the model. We combine these functions to get the posterior distribution, which we maximize (see §4).

3.1 Association

Following previous work [13], we define an *association* $\omega = \{\tau_r \subset B\}_{r=0}^{m}$ to be a partition of B, the set of all detections (body, face) for the entire video. Here, each τ_r, $r = 1, \ldots, m$, called a *track*, is the set of detections which are associated to person r, and τ_0 is the set of spurious detections, generated by a noise process [41]. The prior distribution $p(\omega)$ has hyper parameters λ_A, κ, θ, and λ_N representing the expected detections per person per frame, new tracks per frame, track length, and noise detections per frame [13].

3.2 Scene and VFoA

Our 3D scene model consists of a set of moving persons, represented using 3D cylinders and ellipsoids, which we call the temporal scene, and a set of static objects, represented by 3D points. These objects are assumed to command attention from the people in the scene, which we model explicitly for each person at each frame, and call visual focus of attention (VFoA).

Static objects. The scene contains a set of $\widehat{m}$ static objects, denoted by $\boldsymbol{\chi} = (\boldsymbol{\chi}_1, \ldots, \boldsymbol{\chi}_{\widehat{m}})$, $\boldsymbol{\chi}_r \in \mathbb{R}^3$. Since we do not have any prior information regarding their locations, we set a uniform distribution on their positions over the visible 3D space. We model *interesting locations* as independent from each other by using a joint prior of $p(\boldsymbol{\chi}) = p(\widehat{m}) \prod_{r=1}^{\widehat{m}} p(\boldsymbol{\chi}_r)$, where $p(\widehat{m})$ is Poisson.

Visual focus of attention (VFoA). The scene also contains m people, one for each association track $\tau_r \in \omega$. Each person has a VFoA at each frame that encodes who or what they are observing, if anything. We use $\xi_{rj} \in \{0, \ldots, m+\widehat{m}\}$ to denote the VFoA of person r at frame j, e.g., $\xi_{rj} = r'$ indicates person r is

looking at person or object r' at frame j, where values of $1 \leq \xi_{rj} \leq m$ indicate focus on a person, $m < \xi_{rj} \leq m+\widehat{m}$ on an object, and $\xi_{rj} = 0$ indicates no focus. A priori, people tend to focus on the same visual target in consecutive frames, and we set a simple Markov prior on $\boldsymbol{\xi}_r = (\boldsymbol{\xi}_{r1}, \ldots, \boldsymbol{\xi}_{ml_m})$, where $\xi_{rj} = \xi_{rj-1}$ with high probability. The prior for the entire VFoA set is $p(\boldsymbol{\xi} \,|\, \omega) = \prod_{r=1}^{m} p(\boldsymbol{\xi}_r)$.

Temporal scene. Each person r has temporal 3D state $\mathbf{w}_r = (\mathbf{w}_{r1}, \ldots, \mathbf{w}_{rl_r})$, where each single-frame state consists of the person's ground-plane position $\mathbf{x}_{rj} \in \mathbb{R}^2$, body yaw q_{rj}, head pitch p_{rj}, and head yaw y_{rj}, so that $\mathbf{w}_{rj} = (\mathbf{x}_{rj}, q_{rj}, p_{rj}, y_{rj})$, $j = 1, \ldots, l_r$. Importantly, the head yaw y_{rj} is measured relative to the body yaw q_{rj}, i.e., $y_{rj} = 0$ when person r at frame j is looking straight ahead. Additionally, each person has three size dimensions: width, height, and thickness, denoted by d_r^{w}, d_r^{h}, and d_r^{g}. We will denote the full 3D configuration of track τ_r by $\mathbf{z}_r = (\mathbf{w}_r, d_r^{\mathsf{w}}, d_r^{\mathsf{h}}, d_r^{\mathsf{g}})$. Conceptually, at any given frame j, this can be thought of as a $d_r^{\mathsf{w}} \times d_r^{\mathsf{h}} \times d_r^{\mathsf{g}}$ cylinder whose "front" side is oriented at angle q_{rj}, with an ellipsoid on top that has a pitch of p_{rj} and a yaw of y_{rj} (Fig. 3).

We call $\mathbf{x}_r = (\mathbf{x}_{r1}, \ldots, \mathbf{x}_{rl_r})$ the trajectory of person r, and place a Gaussian process (GP) prior on it to promote smoothness. We use analogous definitions for the body angle trajectory $\mathbf{q}_r$, the head pitch trajectory $\mathbf{p}_r$, and the head yaw trajectory $\mathbf{y}_r$ (e.g., for body angle, $\mathbf{q}_r = (q_{r1}, \ldots, q_{rl_r})$). We use similar smooth GP priors for these trajectories. Importantly, the priors on the head angle trajectories $\mathbf{p}_r$ and $\mathbf{y}_r$ depend on which objects they observe, encoded by $\boldsymbol{\xi}_r$, and their locations, which are contained in $\boldsymbol{\chi}$ and $\mathbf{x}_{-r}$ (all trajectories except $\mathbf{x}_r$); e.g., for head pitch, $p(\mathbf{p}_r \,|\, \boldsymbol{\xi}_r, \boldsymbol{\chi}, \mathbf{x}_{-r})$. We express this dependence by setting the mean of the GP prior to an angle pointing in the direction of the observed object, if any, at each frame.

The prior over a person's full physical state, $p(\mathbf{z}_r \,|\, \boldsymbol{\xi}_r, \boldsymbol{\chi}, \mathbf{x}_{-r}, \omega)$, expands to $p(d_r^{\mathsf{w}}, d_r^{\mathsf{h}}, d_r^{\mathsf{g}}) p(\mathbf{x}_r \,|\, \omega) p(\mathbf{q}_r \,|\, \omega) p(\mathbf{p}_r \,|\, \boldsymbol{\xi}_r, \boldsymbol{\chi}, \mathbf{x}_{-r}, \omega) p(\mathbf{y}_r \,|\, \boldsymbol{\xi}_r, \boldsymbol{\chi}, \mathbf{x}_{-r}, \omega)$, by conditional independence of the state variables given the context variables. We condition on ω as it encodes track length probability. Our overall state prior includes an energy function that makes trajectory intersection unlikely, which is better for inference

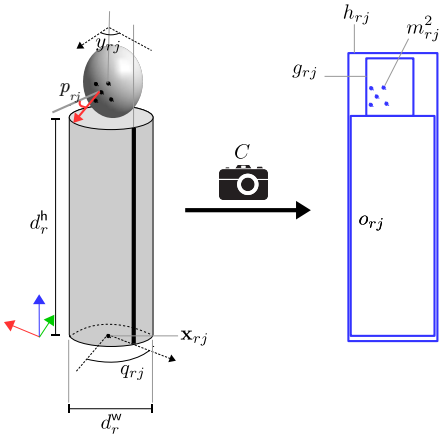

Fig. 3. 3D model for a person (left) and its projection into the image plane (right). Person r at time (frame) j consists of a cylinder at position $\mathbf{x}_{rj}$, of width d_r^{w}, height d_r^{h}, and thickness d_r^{g} (not illustrated) with body angle q_{rj} (the black stripe on the cylinder represents its "front") relative the z-axis of the world. Further, person r's head, represented by the ellipsoid, has yaw y_{rj} relative to the front of the cylinder and pitch p_{rj} indicated by the red arc. Its projection under camera C yields three boxes: model box h_{rj}, model body box o_{rj}, and model face box g_{rj}.

than a simple constraint (details omitted). Excluding the energy function, the overall prior is: $p(\mathbf{z} \mid \boldsymbol{\xi}, \boldsymbol{\chi}, \omega) = \prod_{r=1}^{m} p(\mathbf{z}_r \mid \boldsymbol{\xi}_r, \boldsymbol{\chi}, \mathbf{x}_{-r}, \omega)$, where m is the number of people in the scene.

3.3 Camera

We use a standard perspective camera model [23] with the simplifying assumptions used by Del Pero et al. [18]. Specifically, the world coordinate origin is on the ground plane (we use the xz-plane), and the camera center is $(0, \eta, 0)$, with pitch ψ, and focal length f. This simplified camera has unit aspect ratio, and roll, yaw, axis skew, and principal point offset are all zero. We denote the camera parameters as $C = (\eta, \psi, f)$ and give them vague normal priors whose parameters we set manually.

3.4 Data and likelihood

We use three sources of evidence: person detectors, face landmarks associated with person detections, and optical flow. A person detector [20] provides bounding boxes $B_t = \{b_{t1}, \ldots, b_{tN_t}\}$, $t = 1, \ldots, T$, where N_t is the number of detections at frame t. We define $B = \cup_{t=1}^{T} B_t$ to be the set of all such boxes. We parameterize each box b_{tj} by $(b_{tj}^{\mathrm{x}}, b_{tj}^{\mathrm{top}}, b_{tj}^{\mathrm{bot}})$, representing the x-coordinate of the center, and the y-coordinates of the top and bottom, respectively.

A face landmark detector [71] provides five 2D points for each face, $\mathbf{k}_{ti} = (k_{ti}^1, \ldots, k_{ti}^5)$, representing centers of the eyes, the corners of the mouth, and the tip of the nose, of the ith detection at frame t. We use $I_t^k = \{\mathbf{k}_{t1}, \ldots, \mathbf{k}_{tN}\}$ to represent all face landmarks detected at frame t, and define $I^k = \{I_1^k, \ldots, I_T^k\}$. A dense optical flow estimator [35] provides velocity vectors $I_t^f = \{v_{t1}, \ldots, v_{tN_I}\}$ for each frame $t = 1, \ldots, T-1$, where N_I is the number of pixels in the frame. We also define $I = (I^f, I^k)$.

To compute the data likelihood from evidence in 2D frames, we first convert the 3D model to 2D at each time point, by projecting the 3D scene $\mathbf{z}$ on to the image (via the camera C) as follows.

Model boxes. For each person r at frame j, we compute a set of points on the surface of their body cylinder and head ellipsoid and project them into the image. We then find a tight bounding box on the image plane, h_{rj}, called the *model box*. Similarly, using the cylinder and ellipsoid separately, we compute a *model body box*, o_{rj}, and a *model face box*, g_{rj} (see Figure 3). Using this formulation, we can reason about occlusion in 3D, as we can efficiently compute the non-occluded regions of boxes [13], denoted by $\widehat{o}_{rj}$ (body) and $\widehat{o}_{rj}$ (face).

Face features. We project five face locations on the ellipsoid representing the centers of the eyes, the nose, the corners of the mouth (see Figure 3). We denote the projected face features by $\mathbf{m}_{rj} = (m_{rj}^1, \ldots, m_{rj}^5)$, using a special value when a feature is not visible to the camera.

Image plane motion directions. We define two 2D direction vectors, called *model body vector* and *model face vector*, which represent the 3D motion of the body cylinder (respectively, face ellipsoid) projected onto the image.

To compute the model face vector for person r at its jth frame, we pick a visible point on the head ellipsoid and project that point onto the image at frames j and $j+1$. Then, the model face vector c_{rj} is given by the difference between the two projected points. We perform the analogous computation using the body cylinder to get the model body vector u_{rj}.

Likelihood. We define a likelihood function for each of the data sources discussed above, $p(B \,|\, \omega, \mathbf{z}, C)$, $p(I^f \,|\, \mathbf{z}, C)$, and $p(I^k \,|\, \mathbf{z}, C)$. Since B, I^f, and I^k are conditionally independent given $\mathbf{z}$ and C (see Figure 2), the total likelihood function is given by a product of these three functions.

Detection box likelihood. We assume each assigned detection box has i.i.d Laplace-distributed errors with respect to their assigned model box in the x-coordinate of its center and the y-coordinates of its top and bottom. Our likelihood includes video specific noise rate for box detections, and detector specific miss rate, both of which are critical for inferring the number of tracks [13].

Face landmark likelihood. We associate landmark $\mathbf{k}_{ti}$ to person r at frame t if its centroid is near the center of model face box g_{rt}. Then, we assume a Gaussian noise model around each of the model face features $\mathbf{m}_{rj}$. Specifically, for every $\mathbf{k} \in I^k$, $k^i \sim \mathcal{N}(m^i_{rj}, \Sigma^i_{I^k})$. for $i = 1, \ldots, 5$, where m^i_{rj} is the model face feature assigned to k^i. Assuming independence of all landmarks, we get a landmark likelihood of

$$p(I^k \,|\, \mathbf{z}, C) = \prod_{\mathbf{k} \in I^k} p(\mathbf{k} \,|\, \mathbf{m}(\mathbf{k})), \tag{1}$$

where $\mathbf{m}(\mathbf{k})$ is the predicted face feature for landmark $\mathbf{k}$. Because we link faces to boxes, noisy detections are not relevant. However, the probability of missing a face detection, conditioned on the model (and box) is strongly dependent on whether the face is frontal, or sufficiently in profile that only one eye is visible. Hence we calibrate miss rate for these two cases using held out data.

Optical flow likelihood. We place a Laplace distribution on the difference between the **non-occluded** model body vectors and the average optical flow in the corresponding model body box, and similarly for model face vectors [13].

4 Inference

We wish to find the MAP estimate of ω as a good solution to the data association problem. In addition, we need to infer the camera parameters C, and the association prior parameters $\gamma = (\kappa, \theta, \lambda_N)$, which we want to be video specific. We add to this block of parameters, which do not vary in dimension, the discrete VFoA variables $\boldsymbol{\xi}$. Hence, we seek $(\omega, \gamma, C, \boldsymbol{\xi})$ that maximizes the posterior

$$p(\omega, \gamma, C, \boldsymbol{\xi} \,|\, B, I) \propto p(\omega \,|\, \gamma) p(\gamma) p(C) p(\boldsymbol{\xi} \,|\, \omega) p(B, I \,|\, \omega, C, \boldsymbol{\xi}), \tag{2}$$

where the marginal data likelihood $p(B, I \,|\, \omega, C, \boldsymbol{\xi})$ is given by

$$\int p(B \,|\, \omega, \mathbf{z}, C) p(I \,|\, \mathbf{z}, C) p(\mathbf{z} \,|\, \boldsymbol{\xi}, \boldsymbol{\chi}, \omega) p(\boldsymbol{\chi}) \, \mathrm{d}\boldsymbol{\chi} \, \mathrm{d}\mathbf{z}. \tag{3}$$

4.1 Block sampling over γ, ω, C and ξ

Since expression (2) has no closed form, we approximate its maximum using MCMC block sampling, which successively draws samples from the conditional distributions $p(\gamma \,|\, \omega)$, $p(\omega \,|\, \gamma, \boldsymbol{\xi}, C, B, I)$, $p(C \,|\, \omega, \boldsymbol{\xi}, B, I)$, and $p(\boldsymbol{\xi} \,|\, \omega, C, B, I)$. During sampling, we are required to evaluate the posterior (2), which contains the integral in expression (3). Since this integral cannot be performed analytically, nor can it be computed numerically due to the high dimensionality of $(\mathbf{z}, \boldsymbol{\chi})$, we estimate its value using the Laplace-Metropolis approximation [24]. This approximation requires obtaining the best 3D scene $(\mathbf{z}^*, \boldsymbol{\chi}^*)$ with respect to the posterior distribution $p(\mathbf{z}, \boldsymbol{\chi} \,|\, B, I, \omega, C, \boldsymbol{\xi})$, which we estimate using MCMC (see §4.2), keeping track of the best scene across samples.

We use Gibbs to directly draw samples of the association parameters γ from the conditional posterior $p(\gamma \,|\, \omega)$, an extension of the MCMCDA algorithm [40] to sample values for ω from $p(\omega \,|\, \gamma, \boldsymbol{\xi}, C, B, I)$ [13], and random-walk Metropolis-Hastings (MH) to draw samples of the camera parameters η, ψ, and f from the distribution $p(C \,|\, \omega, \boldsymbol{\xi}, B, I)$.

We also use MH to sample $\boldsymbol{\xi}$ from $p(\boldsymbol{\xi} \,|\, \omega, C, B, I)$ using the following proposal mechanism. For each person r in the scene, at each frame j, we find the set of objects or persons in the current scene estimate $(\mathbf{z}^*, \boldsymbol{\chi}^*)$ that intersect (up to a threshold) with person r's gaze vector. Then, we build a distribution over these objects, which is biased towards the closer ones, as well as the VFoA in the previous frame. We draw a sample from this distribution and assign it to $\boldsymbol{\xi}_{rj}$. We then accept or reject the sample using the standard MH acceptance probability.

4.2 Estimating $(\mathbf{z}^*, \boldsymbol{\chi}^*)$

To approximate the MAP estimate of $(\mathbf{z}^*, \boldsymbol{\chi}^*)$, we alternate sample over $\mathbf{z}$ and $\boldsymbol{\chi}$ under the distribution

$$p(\mathbf{z}, \boldsymbol{\chi} \,|\, B, I, \omega, C, \boldsymbol{\xi}) \propto p(\boldsymbol{\chi}) p(\mathbf{z} \,|\, \boldsymbol{\xi}, \boldsymbol{\chi}, \omega) p(B, I \,|\, \mathbf{z}, \boldsymbol{\chi}, \omega, C). \tag{4}$$

To sample over $\boldsymbol{\chi}$, we use random-walk MH to perturb the position of each interesting point $\boldsymbol{\chi}_r$. We also perform a birth move to introduce new points in the scene. First, we construct a set of candidate points by intersecting all gaze rays across all frames using the current estimate of the temporal 3D state of the persons in the scene $\mathbf{z}$ (see Figure 4). Then, we choose a point from the candidates uniformly at random and add it to $\boldsymbol{\chi}$. We also use a death move, where we remove an element from $\boldsymbol{\chi}$ is uniformly at random.

To explore the space of $\mathbf{z}$, we use an efficient Gaussian process posterior sampling mechanism based on inducing points [55]. The basic idea is to construct a proposal distribution by drawing samples from the conditional GP prior and a set of inducing point locations that provide a low-dimensional representation of the function. We iterate over persons $r = 1, \ldots, m$ and over the different trajectories of each $\mathbf{x}_r$, $\mathbf{q}_r$, $\mathbf{p}_r$, and $\mathbf{y}_r$, drawing a sample at each iteration. More specifically,

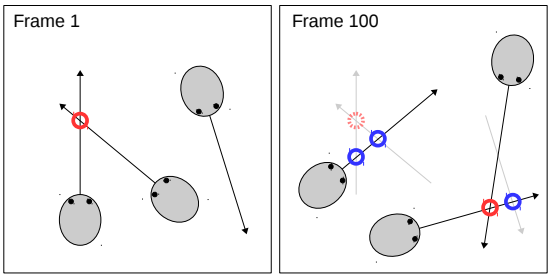

Fig. 4. Proposing static objects. On the left we show a bird's eye view of three people with their corresponding gaze vectors at frame 1. The intersection of two of them creates a candidate static object (red circle). On the right, we show frame 100 of the same video, which also contains three subjects generating four additional candidates. The three lighter lines are gazes recorded in at previous times. The red circles is a candidate generated solely by gazes in the current frame. The three blue circles are candidates generated by intersecting gaze at the current frame with gazes from the previous frames. Finally, the light red circle is the candidate from frame 1.

for a given trajectory, say $\mathbf{q}_r = (q_{r1}, \ldots, q_{rl_r})$, we arbitrarily choose a subset of $(1, \ldots, l_r)$ as *inducing points*, denoted by $(j_1, \ldots, j_{l'_r})$. Then, for each inducing point j_c, we draw a sample from the conditional GP prior $q'_{rj_c} \sim p(q_{rj} \,|\, \mathbf{q}_{rj_{-c}})$, and a sample from the predictive distribution $\mathbf{q}'_r \sim p(\mathbf{q}_r \,|\, \mathbf{q}_{rj_{-c}}, q'_{rj_c})$, where $\mathbf{q}_{rj_{-c}}$ represents $\mathbf{q}_r$ at the set of inducing points excluding j_c. The sample is accepted or rejected using the MH acceptance probability ratio using only the likelihood function $p(B, I \,|\, \mathbf{z}, \boldsymbol{\chi}, \omega, C)$.

5 Evaluation dataset and measures

Several datasets exist for evaluating VFoA recognition in meeting scenarios [7, 5, 8, 29, 58, 59]. Since most of the participants in available meeting datasets are seated throughout the videos, these datasets are not well-suited for evaluating our system, which relies on the ability to detect standing people, and is targeted for scenarios with a diversity of gaze directions in both pitch and yaw. Similarly, datasets such as the Vernissage Corpus dataset [29], which simulates an art gallery scenario, contain many frames where only the upper bodies of the participants are visible. Data sets with walking persons on the other hand uniformly do not encode 3D locations of what people are looking at. While data sets like the challenging SALSA [1], cocktail party [68] and coffee break [16], have head pose annotations, this does not suffice for our goals. Thus we created a new dataset with multiple participants moving freely about while looking at different static targets and each other.

5.1 A new dataset for 3D gaze

We captured and annotated six indoor and two outdoor video sequences. Each setting contained several static object locations, several of which were not vis-

Fig. 5. From left to right, sample frames from two outdoor videos and two indoor videos. The outdoor videos were taken on top of a garage rooftop and within a library courtyard. The indoor videos were shot in a classroom and within a hallway. Each video participant walks inside the scene and records (via an audio recorder) what they are looking at – either another person or a stationary object. All objects in the indoor videos are visible to the camera and can be seen in the frames. Some of the objects in the outdoor videos are not visible to the camera.

ible to the camera. Video participants were asked to walk around and look at each other or the stationary objects, indicating when they started and stopped focusing on each target with an audio recording device. All 8 of our videos were between 40 and 90 seconds long with 3 to 4 people and 5 to 8 objects total (including objects that were not visible). Indoor videos had an image resolution of 1920×1080. Outdoor video resolution was 1440×1080.

Annotation and ground truth. We annotated bounding boxes around each target at each frame using the VATIC annotation tool [60]. We then estimated the ground truth for the 3D positions of each target and the camera parameters in each video by minimizing the reprojection error with respect to 3D locations and heights using the tops and the bottoms of the ground truth boxes. We also used the VFoA audio annotations described above to estimate the ground truth head orientations (pitch and yaw) of each person at every frame where the person was looking at a target. To determine the locations of points not visible to the camera, we measured their locations, and locations that were visible in a shared coordinate system. We then mapped the locations of invisible points to the camera coordinate system.

5.2 Evaluation measures

Trajectory and head pose evaluation. To evaluate the 3D trajectories of the inferred targets, we first find the best match between the inferred tracks and the ground truth tracks using the Hungarian method with pairwise Euclidean distances. We then use two conventional metrics for tracking: MOTA (for accuracy of the data associations) and and MOTP (for precision of the estimated 3D tracks) [50]. Per convention, we set the MOTP threshold to 1 meter. To evaluate head pose estimation, we compute the the equivalent of MOTP for both yaw and pitch between the inferred head poses and their corresponding ground truth head poses (measured in degrees) at frames in which they are available.

To evaluate VFoA estimation, we compare the inferred VFoA of a tracked person to the ground truth VFoA at each frame where it exists. Let N_c be the number of frames where the VFoA is correctly estimated, N_m be the number of

frames where we fail to infer a VFoA (misses), and N_e be the number of frames where we infer an incorrect VFoA. We then compute the following three scores for the VFoA estimation: accuracy $= N_c/N$, mistakes $= N_e/N$, missed $= N_m/N$, where N is the total number of frames that the ground truth for that person records that they were looking at one of the scene VFoA targets. Note that this excludes evaluating the VFoA when the tracked person is transitioning from looking at one target to another. For each video, we compute the average scores over all the tracked persons.

Evaluating inferring interesting locations. Finally, we evaluate how well we can infer the interesting locations in a scene by first finding the best matching between the inferred interesting locations and the preset ground truth locations using the Hungarian method with 1 meter threshold. We then compute the recall and precision for the inferred interesting locations and their average distance to the ground truth locations.

6 Experiments and results

We ran two sets of experiments to evaluate the performance of our method. We do not compare to others on our main tasks since we are not aware of any any relevant published results. We first ran our algorithm and ablated variants on our dataset to assess the impact of different aspects of our approach. We then compare our person tracking performance against our previously published results [13] for people tracking alone to check the effect of the extensions for gaze tracking and object discovery on basic tracking on the well known TUD dataset [2].

Experiments on our dataset. We experiment with enabling and disabling inference over three different parts of the model: the 3D head pose $(\mathbf{p}, \mathbf{y})$, the VFoA $\boldsymbol{\xi}$, and the static objects $\boldsymbol{\chi}$, and replace each with a baseline algorithm. We denote the entire model MGG (for "multiple gaze geometry").

When we disable inference over $(\mathbf{p}, \mathbf{y})$, we simply set the head pose same as the walking direction at each frame (MGG-NO-HEAD). When disabling inference over $\boldsymbol{\xi}$, we set the VFoA of each person at each frame to the object or person first intersects their gaze ray (MGG-NO-VO). Finally, when turning off inference over $\boldsymbol{\chi}$, we estimate the static objects by computing a histogram of the intersections of all the 3D gaze directions of all the people across all the frames, then taking the locations of the top 5 bins with the highest votes (MGG-BASELINE).

Table 1 provides the tracking and head pose estimation results on our dataset. While MOTA and MOTP on position are comparable across all algorithms, the estimated yaw of the head is poor without head pose data. This is not surprising, as the participants in our videos often do not look straight ahead, partly due to the construction of the experiment. By jointly modeling position and pose, we maintain good performance on tracking, while obtaining reasonable accuracy of head yaw, surpassing MGG-NO-HEAD by a significant amount ($\sim 40\ ^\circ$). The gain for pitch was more modest, but the absolute error in pitch was less to begin

Table 1. Performance of different modes of our algorithm on our dataset. Numbers are averaged over eight videos. The first row shows our method with all parts enabled, while the next three rows each shows the algorithm with different aspects disabled, e.g., MGG-NO-HEAD is the stereo gaze algorithm without inferring head pose (see §6 for details). Each column shows a different evaluation measure. We evaluate using the MOTA (with 1.0m threshold) and MOTP for distance and angles. For VFoA we use the measures defined in §5.2.

Algorithm	MOTA	MOTP			VFoA		
		pos	yaw	pitch	accuracy	mistakes	missed
MGG	0.95	0.07	28.1	16.8	0.48	0.35	0.17
MGG-NO-HEAD	0.93	0.08	67.3	19.3	0.14	0.45	0.41
MGG-NO-VO	0.95	0.07	29.9	19.2	0.31	0.39	0.30
MGG-BASELINE	0.92	0.10	70.1	20.8	0.13	0.46	0.41

Table 2. Object discovery performance. Numbers are averaged over eight videos. The algorithms are the same as in Table 1, and the measures are defined in §5.2. We tabulate performance separately for objects not visible in any frame. The performance here may be favorably biased towards invisible objects because they tended to be behind the camera, and looking at them meant a more frontal image of the viewer, which entails better pose estimation.

Algorithm	All static objects			Objects in video			Objects not in video		
	recall	prec	dist	recall	prec	dist	recall	prec	dist
MGG	0.48	0.59	0.58	0.45	0.54	0.59	0.57	0.67	0.51
MGG-NO-HEAD	0.10	0.23	0.35	0.10	0.25	0.35	0.10	0.22	0.34
MGG-NO-VO	0.14	0.15	0.17	0.15	0.18	0.18	0.14	0.14	0.17
MGG-BASELINE	0.13	0.12	0.40	0.14	0.11	0.40	0.13	0.12	0.39

with, which was biased by our instructions and our environment. However, this is ecologically valid, as typical viewing angles are not that far from level.

Table 1 also provides the results for the estimated VFoA. On average, we can correctly identify the VFoA target 48% of the time, much better than the baseline (13%), and better than the ablated MGG-NO-VO version (31%). The later result suggests, perhaps not surprisingly, that learning the 3D locations that people might be looking at provides additional information beyond gaze angles determined from image data alone.

Results for object discovery are shown in Table 2. Here we define success by correctly estimating the location within one meter. We correctly identified 48% of the instances that are available to be identified across the eight videos (recall). In addition, among the ones our method proposes as interesting locations, 59% are correct (precision). The average distance error is a little more than half a meter, which is driven by the choice of the one-meter threshold. Figure 6 shows some example frames of the resulting inferred 3D scene when running the full algorithm (MGG) compared with the baseline (MGG-BASELINE).

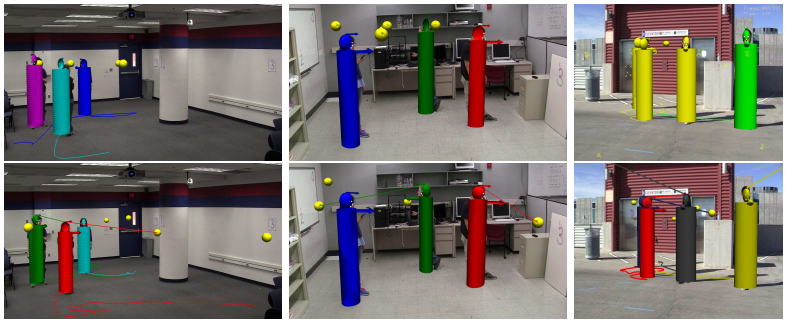

Fig. 6. Visualization of the inferred 3D targets in three scene settings. The top row shows a visualization of the results of the baseline algorithm (MGG-BASELINE), in which the yaw of the gaze direction is set based on the walking directions, and the static objects are estimated from the gaze intersections. The bottom row shows the results of the proposed method on the same frames of the same videos. The arrow on the head indicates the gaze direction and the arrow on the body cylinder indicates the body direction. A tracked person's VFoA is indicated by a line segment from their head connecting to one of the discovered 3D points (yellow spheres) or one of the other tracked people. In the last column, the objects are outside the visible image area.

Experiments on TUD benchmark videos. We compared tracking performance to a similar system for tracking only [13], to evaluate whether incorporating gaze tracking and object inference reduce the tracking performance. We found that we in fact do better on the TUD data, suggesting that the joint inference is helpful.

Table 3. Tracking results on the TUD dataset. We compare to [13], which shows that joint inference over additional scene attributes yielded a tracking performance boost as well.

video	Brau et al. [13]		Proposed	
	MOTA	MOTP	MOTA	MOTP
TUD-Campus	0.84	0.19	0.91	0.11
TUD-Crossing	0.80	0.22	0.80	0.10
TUD-Stadtmitte	0.70	0.27	0.76	0.06
mean	0.78	0.23	0.82	0.08

7 Conclusion

We demonstrated the feasibility of discovering interesting visual locations, specified in 3D, from multiple person gazes observed in monocular video. In particular, on a data set developed for the task, we found that we can infer what people are looking at 59% of the time, and where it is within about .58m. We also found that joint inference over the various scene attributes generally improved the accuracy of the individual estimates. In brief, gaze is both part of scene semantics, and can help determine other aspects of scene semantics.

References

1. Alameda-Pineda, X., Staiano, J., Subramanian, R., Batrinca, L., Ricci, E., Lepri, B., Lanz, O., Sebe, N.: Salsa: A novel dataset for multimodal group behavior analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(8), 1707–1720 (2016)
2. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)
3. Andriyenko, A., Schindler, K.: Globally optimal multi-target tracking on a hexagonal lattice. In: ECCV. pp. 466–479 (2010)
4. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. In: CVPR. pp. 1926–1933 (2012)
5. Ba, S.O., Hung, H., Odobez, J.M.: Visual activity context for focus of attention estimation in dynamic meetings. In: Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on. pp. 1424–1427. IEEE (2009)
6. Ba, S.O., Odobez, J.M.: Probabilistic head pose tracking evaluation in single and multiple camera setups. In: Multimodal Technologies for Perception of Humans, pp. 276–286. Springer (2008)
7. Ba, S.O., Odobez, J.M.: Recognizing visual focus of attention from head pose in natural meetings. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on **39**(1), 16–33 (2009)
8. Ba, S.O., Odobez, J.M.: Multiperson visual focus of attention from head pose and meeting contextual cues. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33**(1), 101–116 (2011)
9. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: CVPR. pp. 3457–3464 (2011)
10. Benfold, B., Reid, I.: Guiding visual surveillance by tracking human attention. In: BMVC. pp. 1–11 (2009)
11. Beymer, D.J.: Face recognition under varying pose. In: Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on. pp. 756–761. IEEE (1994)
12. Blanz, V., Vetter, T.: Face recognition based on fitting a 3D morphable model. Pattern Analysis and Machine Intelligence, IEEE Transactions on **25**(9), 1063–1074 (2003)
13. Brau, E., Guan, J., Simek, K., Del Pero, L., Dawson, C.R., Barnard, K.: Bayesian 3D tracking from monocular video. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 3368–3375. IEEE (2013)
14. Chen, C., Heili, A., Odobez, J.M.: A joint estimation of head and body orientation cues in surveillance video. In: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on. pp. 860–867. IEEE (2011)
15. Chen, C., Odobez, J.M.: We are not contortionists: coupled adaptive learning for head and body orientation estimation in surveillance video. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 1544–1551. IEEE (2012)
16. Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Bue, A.D., Menegaz, G., Murino, V.: Social interaction discovery by statistical analysis of f-formations. In: BMVC (2011 of Conference)
17. Dehghan, A., Assari, S.M., Shah, M.: Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In: CVPR. vol. 1, p. 2 (2015)

18. Del Pero, L., Guan, J., Brau, E., Schlecht, J., Barnard, K.: Sampling bedrooms. CVPR pp. 2009–2016 (2011)
19. Duffner, S., Garcia, C.: Visual focus of attention estimation with unsupervised incremental learning. IEEE Transactions on Circuits and Systems for Video Technology **26**(12), 2264–2272 (2016)
20. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE PAMI (2009)
21. Gee, A., Cipolla, R.: Determining the gaze of faces in images. Image and Vision Computing **12**(10), 639–647 (1994)
22. Gu, L., Kanade, T.: 3D alignment of face in a single image. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. vol. 1, pp. 1305–1312. IEEE (2006)
23. Hartley, R., Zisserman, A.: Multiple View Geometry in Computer Vision. Cambridge University Press (2000)
24. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning; data mining, inference, and prediction. Springer series in statistics, Springer-Verlag, New York (2001)
25. Horprasert, T., Yacoob, Y., Davis, L.S.: Computing 3d head orientation from a monocular image sequence. In: 25th Annual AIPR Workshop on Emerging Applications of Computer Vision. pp. 244–252. International Society for Optics and Photonics (1997)
26. Huang, J., Shao, X., Wechsler, H.: Face pose discrimination using support vector machines (svm). In: Pattern Recognition, 1998. Proceedings. Fourteenth International Conference on. vol. 1, pp. 154–156. IEEE (1998)
27. Huang, Y., Duan, D., Cui, J., Davoine, F., Wang, L., Zha, H.: Joint estimation of head pose and visual focus of attention. In: Image Processing (ICIP), 2014 IEEE International Conference on. pp. 3332–3336. IEEE (2014)
28. Isard, M., MacCormick, J.: Bramble: A bayesian multiple-blob tracker. In: ICCV. pp. 34–41 (2001)
29. Jayagopi, D.B., Sheikhi, S., Klotz, D., Wienke, J., Odobez, J.M., Wrede, S., Khalidov, V., Nguyen, L., Wrede, B., Gatica-Perez, D.: The vernissage corpus: A multimodal human-robot-interaction dataset. Tech. rep. (2012)
30. Kitani, K.M., Ziebart, B.D., Bagnell, J.A., Hebert, M.: Activity forecasting. In: ECCV (2012 of Conference)
31. Kuo, C., Huang, C., Nevatia, R.: Multi-target tracking by on-line learned discriminative appearance models. In: CVPR. pp. 685–692 (2010)
32. La Cascia, M., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3d models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **22**(4), 322–336 (2000)
33. Li, Y., Gong, S., Liddell, H.: Support vector regression and classification based multi-view face detection and recognition. In: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. pp. 300–305. IEEE (2000)
34. Li, Y., Gong, S., Sherrah, J., Liddell, H.: Support vector machine based multi-view face detection and recognition. Image and Vision Computing **22**(5), 413–427 (2004)
35. Liu, C.: Exploring New Representations and Applications for Motion Analysis. Ph.D. thesis, M.I.T. (2009)

36. Massé, B., Ba, S., Horaud, R.: Simultaneous estimation of gaze direction and visual focus of attention for multi-person-to-robot interaction. In: Multimedia and Expo (ICME), 2016 IEEE International Conference on. pp. 1–6. IEEE (2016)
37. Milan, A., Leal-Taixé, L., Schindler, K., Reid, I.: Joint tracking and segmentation of multiple targets. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5397–5406 (2015)
38. Murphy-Chutorian, E., Trivedi, M.M.: Head pose estimation in computer vision: A survey. Pattern Analysis and Machine Intelligence, IEEE Transactions on **31**(4), 607–626 (2009)
39. Niyogi, S., Freeman, W.T.: Example-based head tracking. In: Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on. pp. 374–378. IEEE (1996)
40. Oh, S.: Bayesian formulation of data association and markov chain monte carlo data association. In: Robotics: Science and Systems Conference (RSS) Workshop Inside Data association (2008)
41. Oh, S., Russell, S., Sastry, S.: Markov chain Monte Carlo data association for general multiple target tracking problems. (2004)
42. Otsuka, K., Takemae, Y., Yamato, J.: A probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. In: Proceedings of the 7th international conference on Multimodal interfaces. pp. 191–198. ACM (2005)
43. Otsuka, K., Yamato, J., Takemae, Y., Murase, H.: Conversation scene analysis with dynamic bayesian network basedon visual head tracking. In: Multimedia and Expo, 2006 IEEE International Conference on. pp. 949–952. IEEE (2006)
44. Pirsiavash, H., Ramanan, D., Fowlkes, C.: Globally-optimal greedy algorithms for tracking a variable number of objects. CVPR pp. 1201–1208 (2011)
45. Sankaranarayanan, K., Chang, M.C., Krahnstoever, N.: Tracking gaze direction from far-field surveillance cameras. In: Applications of Computer Vision (WACV), 2011 IEEE Workshop on. pp. 519–526. IEEE (2011)
46. Segal, A.V., Reid, I.: Latent data association: Bayesian model selection for multi-target tracking. In: Computer Vision (ICCV), 2013 IEEE International Conference on. pp. 2904–2911. IEEE (2013)
47. Sheikhi, S., Odobez, J.M.: Recognizing the visual focus of attention for human robot interaction. In: International Workshop on Human Behavior Understanding. pp. 99–112. Springer (2012)
48. Smith, K., Ba, S.O., Gatica-Perez, D., Odobez, J.M.: Tracking the multi person wandering visual focus of attention. In: Proceedings of the 8th international conference on Multimodal interfaces. pp. 265–272. ACM (2006)
49. Smith, K., Ba, S.O., Odobez, J.M., Gatica-Perez, D.: Tracking the visual focus of attention for a varying number of wandering people. Pattern Analysis and Machine Intelligence, IEEE Transactions on **30**(7), 1212–1229 (2008)
50. Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., Soundararajan, P.: The CLEAR 2006 evaluation. In: Proceedings of the 1st international evaluation conference on Classification of events, activities and relationships. pp. 1–44. CLEAR'06, Berlin, Heidelberg (2007)
51. Stiefelhagen, R., Yang, J., Waibel, A.: Modeling focus of attention for meeting indexing. In: Proceedings of the seventh ACM international conference on Multimedia (Part 1). pp. 3–10. ACM (1999)
52. Stiefelhagen, R., Yang, J., Waibel, A.: Modeling focus of attention for meeting indexing based on multiple cues. Neural Networks, IEEE Transactions on **13**(4), 928–938 (2002)

53. Stiefelhagen, R., Zhu, J.: Head orientation and gaze direction in meetings. In: CHI'02 Extended Abstracts on Human Factors in Computing Systems. pp. 858–859. ACM (2002)
54. Tang, S., Andres, B., Andriluka, M., Schiele, B.: Subgraph decomposition for multi-target tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 50330–5041 (2015)
55. Titsias, M.K., Lawrence, N.D., Rattray, M.: Efficient sampling for Gaussian Process inference using control variables. In: Advances in Neural Information Processing Systems. vol. 21, pp. 1681–1688. Curran Associates, Inc., Vancouver, British Columbia, Canada (2008)
56. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze estimation. Image Processing, IEEE Transactions on **21**(2), 802–815 (2012)
57. Voit, M., Nickel, K., Stiefelhagen, R.: Head pose estimation in single-and multi-view environments-results on the CLEAR07 benchmarks. In: Multimodal Technologies for Perception of Humans, pp. 307–316. Springer (2008)
58. Voit, M., Stiefelhagen, R.: Deducing the visual focus of attention from head pose estimation in dynamic multi-view meeting scenarios. In: Proceedings of the 10th international conference on Multimodal interfaces. pp. 173–180. ACM (2008)
59. Voit, M., Stiefelhagen, R.: 3D user-perspective, voxel-based estimation of visual focus of attention in dynamic meeting scenarios. In: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction. p. 51. ACM (2010)
60. Vondrick, C., Patterson, D., Ramanan, D.: Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision pp. 1–21, http://dx.doi.org/10.1007/s11263-012-0564-1, 10.1007/s11263-012-0564-1
61. Wei, P., Zhao, Y., Zheng, N., Zhu, S.C.: Modeling 4d human-object interactions for joint event segmentation, recognition, and object localization. IEEE Trans Pattern Anal Mach Intell. (2016)
62. Wu, Y., Toyama, K.: Wide-range, person-and illumination-insensitive head orientation estimation. In: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on. pp. 183–188. IEEE (2000)
63. Xiao, J., Moriyama, T., Kanade, T., Cohn, J.F.: Robust full-motion recovery of head by dynamic templates and re-registration techniques. International Journal of Imaging Systems and Technology **13**(1), 85–94 (2003)
64. Xie, D., Todorovicy, S., Zhu, S.C.: Inferring "dark matter and "dark energy from videos. In: ICCV (2013 of Conference)
65. Yang, R., Zhang, Z.: Model-based head pose tracking with stereovision. In: Automatic Face and Gesture Recognition, 2002. Proceedings. Fifth IEEE International Conference on. pp. 255–260. IEEE (2002)
66. Yi, Y., Xu, H.: Hierarchical data association framework with occlusion handling for multiple targets tracking. Signal Processing Letters, IEEE **21**(3), 288–291 (2014)
67. Yücel, Z., Salah, A.A., Mericli, C., Meriçli, T., Valenti, R., Gevers, T.: Joint attention by gaze interpolation and saliency. IEEE Transactions on cybernetics **43**(3), 829–842 (2013)
68. Zen, G., Lepri, B., Ricci, E., Lanz, O.: Space speaks: towards socially and personality aware visual surveillance. In: 1st ACM international workshop on Multimodal pervasive video analysis. pp. 37–42. ACM, Firenze, Italy (2010 of Conference)
69. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. pp. 1–8. IEEE (2008)

70. Zhao, G., Chen, L., Song, J., Chen, G.: Large head movement tracking using sift-based registration. In: Proceedings of the 15th international conference on Multimedia. pp. 807–810. ACM (2007)
71. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2879–2886. IEEE (2012)