Detecting Mutual Awareness Events

Meir Cohen*, Ilan Shimshoni◇, Ehud Rivlin*‡, and Amit Adam*

Abstract—It is quite common that multiple human observers attend to a single static interest point. This is known as a mutual awareness event (MAWE). A preferred way to monitor these situations is with a camera that captures the human observers while using existing face detection and head pose estimation algorithms. The current work studies the underlying geometric constraints of MAWEs and reformulates them in terms of image measurements. The constraints are then used in a method that (1) detects whether such an interest point does exist, (2) determines where it is located, (3) identifies who was attending to it, and (4) reports where and when each observer was while attending to it. The method is also applied on another interesting event when a single moving human observer fixates on a single static interest point. The method can deal with the general case of an uncalibrated camera in a general environment. This is in contrast to other work on similar problems that inherently assume a known environment or a calibrated camera. The method was tested on about 75 images from various scenes and robustly detects MAWEs and estimates their related attributes. Most of the images were found by searching the Internet.

Index Terms—Head Pose, Mutual Awareness, Social Signal Processing, Sparse 3D Structure

1 Introduction

Mutual awareness event (MAWE) is an event where multiple human observers attend to a single interest point at the same time. It is fundamental to analyzing the social behavior of a group. Since it indicates common interest and common knowledge among group members, it has been addressed by the psychological community in many contexts, e.g. [1], [2], [3]. This work on MAWEs contributes to the emerging domain of *Social Signal Processing* [4]. The detection of MAWEs and their attributes can assist various applications. The least invasive way to monitor MAWEs is by using a camera that captures the human observers. Solutions for this problem can use available software for the detection of faces and for the estimation of head poses.

The current work addresses the detection of MAWEs and their attributes. In fact, it also covers an even more general problem that takes into account the time domain: the problem of a single static interest point being the visual focus of attention (VFOA) of several human observers at different times. This generalization will be called *temporal mutual awareness event* (t-MAWE). The related interest point of both MAWE and t-MAWE will be called the *visual intersection of attention* (VIOA). The proposed method detects a MAWE with its attributes from a single image (simple MAWE) or from a video captured by a static camera. Those attributes include the spatiotemporal positions of the mutually aware observers and their VIOA. Note that a frequent but limited t-MAWE problem is addressed here, i.e. the static camera case which has the same analysis as the single frame MAWE problem. The general t-MAWE problem, which is left for future research, also includes camera motion, zooming, tracking, and moving interest points.

The analysis of a MAWE using a single camera depends on three types of information: (1) the 3D structure of the scene (the positions of the mutually aware observers and the VIOA point), (2) the calibration of the camera (internal and external), and (3) the detection of the mutually aware observers and the direction of their gaze.

Various application domains could benefit from the detection of a MAWE and its attributes. It can be used to build a human rated salience map of a 3D environment or to analyze the social behavior of a crowd. A complementary usage is the identification of people whose attention is an exception to the detected MAWEs.

The following section reviews related work. Next, Section 3 describes the proposed method, including its main ideas. The geometrical analysis is presented in Section 4. Section 5 addresses the challenges in real-world scenarios and presents an algorithm to deal with them. Finally, the last section, 6, presents simulation and real-data experiments, including the outcome of the method on many images obtained from the Internet.

2 Related Work

To the best of our knowledge, no work addresses the VIOA of a group as a whole in the general case. However, for a single human observer, the recovery of the visual focus of attention (VFOA) has been addressed. The research so far has used gaze estimation to infer the VFOA while assuming known relations between the observer, the possible interest points and the camera. One approach assumes fixed and fully calibrated cameras such that an observer and his image based estimated gaze direction will be well localized in the 3D environment [5], [6], [7]. Other works do not assume calibrated cameras but assume a known environment which

* *Technion - Israel Institute of Technology, Haifa, Israel*
Email: meirc@cs.technion.ac.il
◇ *Haifa University, Haifa, Israel*
Email: ishimshoni@mis.haifa.ac.il
‡ *Google Inc.*

includes a few cameras in fixed positions, a predefined set of a few fixed interest points and a few fixed 3D positions where observers can be located. Under those assumptions by using supervised learning, either the VIOA of an observer is detected [8], [9], [10], [11], [12] or the position of the observer is estimated [13]. The supervised learning is directly based on using a training data-set of images that was manually annotated. The learning is possible due to the fact that any given triplet of a specific camera, a specific 3D position and a specific interest point is uniquely associated with a head pose as it appears in images that are captured by the specific camera when an observer, which is located at the specific 3D position, is looking at the specific interest point.

For a general MAWE it should only be assumed that the observers and the interest point are related, i.e. the visual attention aggregates at a point from several directions. However, no prior relation should be assumed between the camera and the scene, i.e. a general solution for the MAWE problem should handle an uncalibrated camera both internally and externally. This work analyzes the general MAWE and by taking a different approach suggests a method to detect it and find its related attributes. As such, this approach does not require a training data-set and can instantly be applied on a new and unknown environment. The key idea of the approach is that the recovery of the 3D position of the VIOA is actually a triangulation problem, which shares the 3D geometric constraint of the well known problem of 3D reconstruction from multiple views. This geometric constraint is used to estimate all the MAWE's attributes only from image measurements, where the estimation is controlled by a Bayesian framework.

The current work is a significant enhancement of our previous workshop paper [14]. The improvements include: a statistical framework, an enhanced formulation of depth estimation, an enhanced accuracy of the detection algorithms, and a much richer set of experiments that test the accuracy of the method.

The above three approaches, including ours, are based on the estimation of observer's gaze direction relative to the capturing camera. Even the most accurate gaze estimation methods do not find a single point, but return a ray or a cone (centered around the ray) in the 3D world. Thus, additional cues are inherently required for the recovery of the VFOA or the VIOA. Nevertheless, an accurate estimation of gaze direction could notably improve the accuracy of the recovery. A basic finding states that the VFOA can be reasonably approximated by head pose in many cases [15]. This is important since pupil positions can be recovered only in high resolution images. The estimation of the head's angles from its image is not an easy task and the current solutions are inaccurate: for a head image in moderate resolution, the pitch angle is usually estimated with large errors. A recent survey [16], covers about 100 methods for pose estimation, including [17], which is used in our implementation. This method is mentioned in the survey as one of the most accurate. However, it estimates only 2 DOF of head pose in a continuous manner and is applicable to far-viewed heads. Still, head pose estimation remains an open research problem.

Detecting faces in an image is the first step in VFOA recovery. An efficient method, based on the boosting algorithm, is the popular method by Viola and Jones [18]. There are several extensions of it, including [19], [20], and a popular implementation [21]. We are assisted by it in depth estimation.

3 The Suggested Method

This work studies the detection of *mutual awareness events* (MAWEs), in which a single static interest point is the visual focus of attention (VFOA) of several human observers, instantly or over a time period. The interest point is called the *visual intersection of attention* (VIOA). In other words, the *attention rays* of the human observers intersect in 3D space at the VIOA, and the attention *aggregates* at the VIOA.

This work assumes a general setup for which: (1) the environment is arbitrary, (2) a static uncalibrated camera is used, (3) the location of the VIOA may be arbitrary (or even out of the camera's field of view), and (4) there is no training data for the VIOA. The 3D recovery of the VIOA may be obtained from a single arbitrary image of mutually aware observers of an interest point or from a video of fixating observers.

As such, the problem's setup consists of n human observers, the intersection point, Q, the camera (including its calibration matrix K), the head pose algorithm, and finally the measurement set $U = \{U_1, U_2, \ldots, U_n\}$, where $U_i = \begin{pmatrix} p_i & r_i & \alpha_i & \beta_i & \gamma_i \end{pmatrix}^T$. Those parameters are the head position in the image, its size and its pose angles, which are the results of the detection algorithms; see Figure 1(a).

An hypothesis of a MAWE includes a group of observers and their VIOA. Obtaining a small set of MAWE hypotheses is essential for an efficient solution. The geometry of the VIOA is used to constrain the hypothesis search. When U is accurate and the internal calibration of the camera is known, the geometry enforces a single hypothesis. However, in the real world the measurements are noisy and the calibration is not always known.

In the ideal situation all observers have the same VFOA, K is known, and the head pose algorithm detects all observers and their associated measurements accurately. The VIOA is then accurately computed by applying the geometric constraint, as explained in Section 4.

However, in real life, the attention of an observer may be related to the other observers in one of two ways: (1) she may share a VFOA with other observers, i.e., the VIOA, or (2) she might have an *independent* attention ray that does not intersect with other observers. Thus, if a MAWE occurs, its VIOA is shared by a subset of the observers in the image who have $U' \subseteq U$ as their measurements set. In addition, $U_i \in U'$ is noisy, K

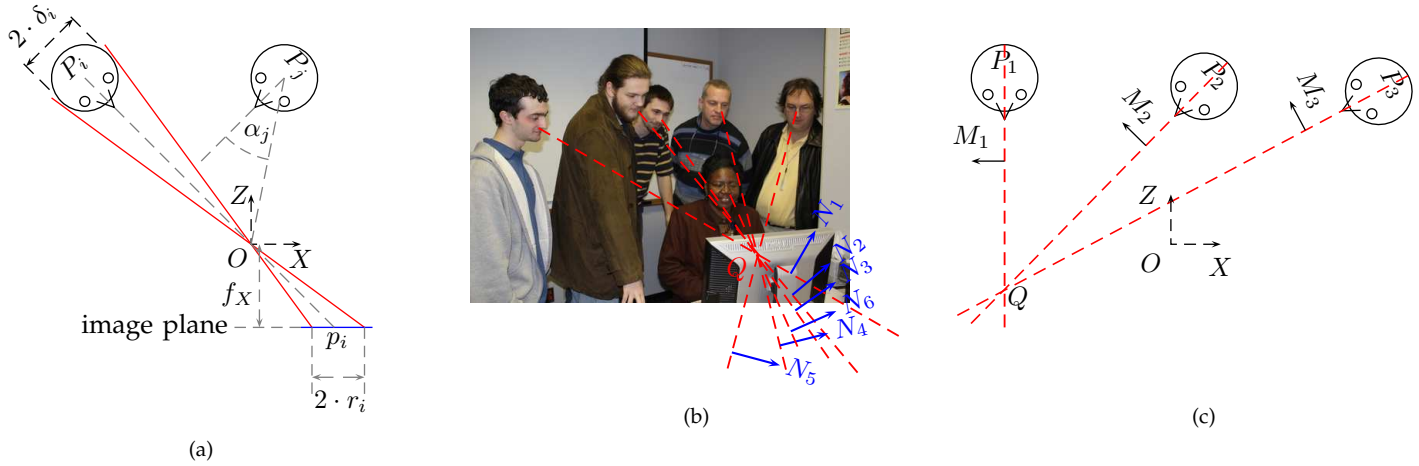


Fig. 1: (a) A top view of a scene, the XZ-plane, demonstrates the relation between the i'th observer and the camera. It can be seen that the depth of the i'th observer is proportional to the diameter of his head's projection. (b) The first type of plane (Section 4.1) associated with the i'th observer goes through the origin, the VIOA, and the nose of the observer. The normal to the i'th first type of plane is N_i. The image plane view of the scene, the XY-plane, demonstrates how planes of the first type intersect on the line from the origin to the VIOA. (c) The second type of plane (Section 4.1) associated with the i'th observer is the median-section of his head. The normal to the i'th second type of plane is M_i. A top view of a scene, the XZ-plane, demonstrates how planes of the second type intersect at the VIOA.

might be unknown, and the detection algorithms might produce false detections.

The full algorithm searches for a suitable hypothesis. First it estimates the unknown K while obtaining an initial U' and Q. Then, it refines K and Q to find the most probable U'. The detected observers are split into inliers and outliers using the log likelihood ratio threshold T_{LLR}. In the last step, the algorithm checks whether the detected MAWE is significant or whether it is probable that it has emerged from the distribution of the independent attention rays. Alg. 1 summarizes the full algorithm that handles real-life scenes. The details of this algorithm are given in Section 5. After the detection of a MAWE the method can be reapplied on the remaining outliers and detect additional MAWEs if they exist.

4 The Geometry of Mutual Awareness

The position of the VIOA is the main attribute of a MAWE. Thus, estimating the position of the VIOA is inherently coupled with the detection of such an event. The analysis of the geometry of a MAWE is a major part of this work. It enables the estimation of the VIOA from head pose and constrains the search of related parameters, such as camera calibration and scene structure.

The following notations will be used for the geometric analysis. Let $P_i = (X_i, Y_i, Z_i)^T \in R^3$, $i \in \{1, \ldots, n\}$ be the 3D positions of n heads. The pose of the i'th head is expressed by $(\alpha_i, \beta_i, \gamma_i)$, the yaw, pitch, and roll angles, respectively. The angles are given with respect to a head facing the camera (frontal view). In this situation $\alpha_i = \beta_i = \gamma_i = 0$. The attention ray of a head is a ray perpendicular to the face plane and is directed forward. Note that the roll angle does not change the direction of the attention ray. For a frontally viewed head, the attention ray is directed towards the camera. The angles are combined to create the rotation matrix, $R_i(\alpha_i, \beta_i, \gamma_i) = R_{\beta_i} \cdot R_{\alpha_i} \cdot R_{\gamma_i}$.

The attention ray of the i'th observer is rotated by the rotation matrix, R_i, from the direction towards the camera to the direction of the VIOA, Q. This can be

---

procedure IS-MUTUAL-AWARENESS(image/s, $K_{\text{range}}, \Sigma, T_{\text{LLR}}, T_{\text{miss}}, T_{\text{min}}, T_{\text{MA}}$) // **For more details see:**
- $(U, I_0) \leftarrow$ DETECT-OBSERVERS(image/s) // 6.1
- **while** $(|I_0| \geq T_{\text{min}})$ // $|I_0| = n$ observers
  - $(K, Q, U'[, L], I, \text{score}) \leftarrow$ ESTIMATE-VIOA-K($K_{\text{range}}, U, I_0, \Sigma, T_{\text{LLR}}, T_{\text{miss}}, T_{\text{min}}$) // 5.1, Alg. 2
  - $(K, Q, U'[, L], I, \text{score}) \leftarrow$ FINE-TUNE($K, Q, U, U'[, L], I, \Sigma, T_{\text{LLR}}, itr_n$) // 5.2
  - **if** IS-SIGNIFICANT($|I|, |I_0|, \Sigma, T_{\text{miss}}, T_{\text{MA}}$) // 5.3, (13) and (14)
    - **then** REPORT-MAWE($K, Q, U'[, L], I, \text{score}$) //
    - **else return** //
  - $K_{\text{range}} \leftarrow K$ // Search for an another
  - $(U, I_0) \leftarrow (U(I_0 \setminus I), I_0 \setminus I)$ // significant MAWE.

Alg. 1: Overview algorithm for detecting MAWEs in real life scenarios.

written, for the i'th observer in a MAWE, as

$$Q = P_i + \zeta_i \cdot R_i \cdot P_i, \tag{1}$$

where $\zeta_i > 0$ s.t. $\zeta_i \cdot \|P_i\|_2$ is the distance between Q and P_i. It is equivalent to say that

$$(Q - P_i) \times (R_i \cdot P_i) = 0, \tag{2}$$

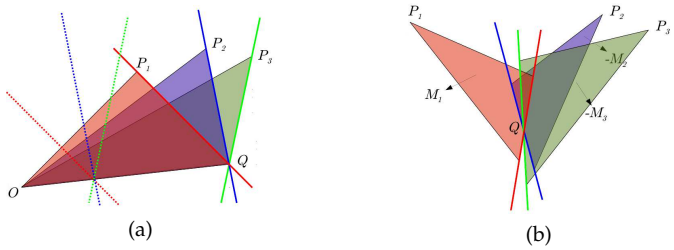
for all the participants in the MAWE. For every observer with known pose R_i and known position P_i, this constraint introduces 3 equations with a single unknown ζ_i. Thus, for n observers, there are $3 \cdot n$ equations and n unknowns. If also the VIOA, Q, is unknown there are $n + 3$ unknowns. When the scene data is collected by a single static camera, head pose and head position are estimated from the captured images. If the internal calibration of the camera is unknown, it introduces, at most, another 4 unknowns.

Note however, the special case when all the observers look directly at the camera, e.g., when a group photo is being taken, each of them appears in the image in frontal view, i.e., $R_i = I$ for $1 \leq i \leq n$. In this case the VIOA is the camera, i.e., $Q = \mathbf{0}$. Therefore, the constraint (2) is trivially true for $1 \leq i \leq n$, i.e., $\mathbf{0} - P_i \propto I \cdot P_i$. Thus, the internal calibration of the camera is not constrained by the MAWE.

As mentioned above, head pose algorithms usually do not fully estimate the rotation matrix, R_i. Quite often only one or two angles out of the three are estimated. The next section addresses this problem by using the above constraint to formulate geometric constraints that are more robust and exploit partial head pose information. This is done by intersecting planes, where the plane is spanned by the poorly estimated angle. Moreover, the geometric plane intersection in 3D is expressed by the data extracted from the images. In particular, the next section includes how the depth is estimated from the face radius.

4.1 VIOA Estimation by Plane Intersection

The following addresses the common case where the pitch angle is poorly estimated but the scheme can suit other angles too. Thus, plane intersection is used as a valid geometric constraint while bypassing the uncertainty in the pitch angle. In the following paragraphs two plane types, for every head, are considered. Each type of plane contains the intersection point, Q, as can be seen in Figures 1 and 2. The joint constraint of all the heads in the MAWE is formulated using the normals to the planes. This joint constraint will be used to find the location of the VIOA that optimizes this constraint. Each type of plane can be used separately to find Q except when the planes coincide. Nevertheless, the two plane types complement each other: the first type is robust to depth errors and the second type is robust to pitch errors. The planes of the first type coincide only when all observers are located on a straight line and the planes of the second type coincide only when all observers are located on a straight line and have the same roll angle.



(a) (b)

Fig. 2: Three observers $\{P_i\}_{i\in\{1,2,3\}}$ and their associated planes. The attention rays of the three observers intersect at Q. The VIOA, Q, sits on the intersection line (sheaf) of planes of the first type and is the intersection point of planes of the second type: (a) Plane Π_i (first type). (b) Plane Π_i^β (second type).

The first plane, Π_i, is spanned by P_i and $R_i \cdot P_i$. This plane contains the line between the observer and the origin OP_i; in particular, it contains the origin as can be seen in Figures 1(b) and 2(a). The normal to the plane Π_i is denoted N_i and expressed by $N_i = \frac{P_i \times (R_i \cdot P_i)}{\|P_i \times (R_i \cdot P_i)\|_2}$. Plane Π_i constrains the intersection point by $N_i^T \cdot Q = 0$. Arranging the normals in a matrix yields the matrix $N = \begin{pmatrix} N_1 & N_2 & \dots & N_n \end{pmatrix}$. The normals N_i determine the direction of Q regardless of the depth information, $\{r_i\}_{i=1}^n$. This can be seen by expressing P_i by its projection, $p_i \cong K \cdot P_i$, on the image plane ($p_i = \begin{pmatrix} x_i & y_i & 1 \end{pmatrix}^T$), i.e., $P_i = Z_i \cdot K^{-1} \cdot p_i$, and then writing the above as

$$\begin{aligned} N_i &= \frac{(K^{-1} \cdot p_i \cdot Z_i) \times (R_i \cdot K^{-1} \cdot p_i \cdot Z_i)}{\|(K^{-1} \cdot p_i \cdot Z_i) \times (R_i \cdot K^{-1} \cdot p_i \cdot Z_i)\|_2} = \\ &= \frac{(K^{-1} \cdot p_i) \times (R_i \cdot K^{-1} \cdot p_i)}{\|(K^{-1} \cdot p_i) \times (R_i \cdot K^{-1} \cdot p_i)\|_2}. \end{aligned} \tag{3}$$

One way to solve Q is to find the direction of Q, $\tilde{Q} = \frac{Q}{\|Q\|_2}$, by applying SVD on the homogeneous equations of the first constraint $N^T \cdot Q = 0$. This option is stable because the estimated direction of Q is not affected by the errors in the measured depth (face radius). The magnitude of Q can then be determined by using (2) which yields a set of linear equations in $\tilde{Q}$, i.e.,

$$\|Q\|_2 \cdot \begin{pmatrix} \tilde{Q} \times (R_1 \cdot K^{-1} \cdot p_1) \\ \tilde{Q} \times (R_2 \cdot K^{-1} \cdot p_2) \\ \vdots \\ \tilde{Q} \times (R_n \cdot K^{-1} \cdot p_n) \end{pmatrix} = \begin{pmatrix} Z_1 \cdot (K^{-1} \cdot p_1) \times (R_1 \cdot K^{-1} \cdot p_1) \\ Z_2 \cdot (K^{-1} \cdot p_2) \times (R_2 \cdot K^{-1} \cdot p_2) \\ \vdots \\ Z_n \cdot (K^{-1} \cdot p_n) \times (R_n \cdot K^{-1} \cdot p_n) \end{pmatrix}.$$

Depth information is required for the last step. It will be shown in Section 4.2 that the depth of the i'th head, Z_i, can be estimated from the image by measuring the radius r_i of the i'th head. That is because the size of a head corresponds to its distance from the camera.

The second plane, Π_i^β, is the YZ plane of the i'th face that includes the line between the middle of the nose and the middle of the chin ("Median Section"). This plane contains the line, Q-P_i, between the observer and the VIOA, as can be seen in Figures 1(c) and 2(b). The normal to the plane Π_i^β is denoted M_i and expressed by $M_i = R_i \cdot \frac{P_i \times (0 \quad 1 \quad 0)^T}{\|P_i\|_2}$. Plane Π_i^β constrains the intersec-

tion point by $M_i^T \cdot (Q - P_i) = 0$. Arranging the normals in a matrix yields the matrix M.

Combining the above two constraints also solves for Q directly by

$$Q = \begin{pmatrix} M^T \\ N \end{pmatrix}^{\dagger} \cdot \begin{pmatrix} \mathrm{mp} \\ 0 \end{pmatrix}, \tag{4}$$

where $\mathrm{mp}_i = M_i^T \cdot P_i$. The depth information is required to compute mp. The following section describes how to estimate the head's depth from its radius.

4.2 Depth Estimation

The suggested method requires a depth estimation for each head, which is computed from the 2D radius of the head in the image. Specifically, a circle that is centered on the nose and covers the eyes and mouth will be estimated as shown in Figure 3. The estimated radius is the same for any head pose such as if the face was captured in a frontal view. Let δ_i be the 3D physiological radius of the i'th face and r_i be the radius of that face as measured in the image by the face detector. In addition, it is assumed that the average 3D physiological face size, δ, is a good approximation such that $\delta_i \approx \delta$, e.g. see [22]. The estimated depth is given by:

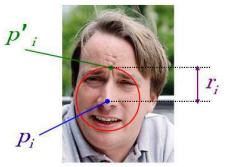


Fig. 3: Face radius.

$$Z_i = \frac{\delta}{C_1} \cdot \left(\frac{1}{r_i} \cdot \frac{C_3}{C_2} + C_2 \cdot \left(\frac{C_3}{r_i} + 1 \right) \right), \tag{5}$$

where $C_1 = \left\| K^{-1} \cdot p_i \right\|_2$, $C_2 = \left\| D_2 \left(K^{-1} \cdot p_i \right) \right\|_2$, $C_3 = \left\| D_2(p_i) \right\|_2$, and $D_2 : \mathbb{R}^3 \to \mathbb{R}^2$ such that $D_2(X, Y, Z) = (X, Y)$; see the Appendix for a proof.

The expression in (5) agrees with similar expressions in [23] and [24] that were developed for special cases of camera calibration. In the case where $K = \left(\begin{smallmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{smallmatrix} \right)$ the expression in (5) simplifies to $Z_i = \frac{\delta}{C_1} \cdot \left(\frac{f}{r_i} + \frac{C_3}{f} \cdot \left(\frac{C_3}{r_i} + 1 \right) \right)$. When p_i is near the origin, i.e. $C_3 \approx 0$ and $C_1 \approx 1$, the approximated depth near the origin will be $Z_i = f \cdot \frac{\delta}{r_i}$.

Planar Constraint. Images that contain MAWEs include multiple observers that in many cases are well organized in the 3D space. In some cases the positions of the observer heads can be approximated to lie on a plane. In such cases the plane can be estimated from the depth information and be used to correct back the depth for each head.

5 Mutual Awareness in the Wild

Applying the face detection and head pose estimation algorithms on the images of the scene produces a set of n measurements, U (see Section 3). In general, nothing is assumed on how the detection algorithms work internally and usually U will be noisy and include false detections. A MAWE can be detected by recovering its related state of the world, $\mathcal{W} = \{K, Q, \hat{U}\}$, when the image was captured, i.e., image calibration, VIOA position, and the position and head pose of the observers. The optimal world state minimizes some preselected loss function w.r.t. U. In this work, the Bayesian setup, that also models outliers, is preferred and used as the loss function. Thus, the loss is measured w.r.t. image measurements rather than w.r.t. the 3D reconstructed world, similar to the *optimal correction* strategy taken by [25] for the triangulation problem in stereo vision. The optimal world state is the MAP estimator, i.e., the one that maximizes $\mathfrak{p}\left(K, Q, \hat{U} \,\middle|\, U \right)$. Note that the loss function is highly non-linear w.r.t. the hypothesis, i.e., w.r.t. the calibration matrix K, and the VIOA Q. A general optimization algorithm can be used to find the best world state estimator w.r.t. the loss function. It is expected to converge to the global optima if its initial guess is close enough. The Bayesian setup makes it possible to decide whether a detected MAWE is significant. Those are the three main components which compose Alg. 1 in Section 3. The details of the three components are given bellow: (1) the search for an initial hypothesis (Section 5.1), (2) the optimization (Section 5.2), and (3) the significance decision (Section 5.3).

5.1 Initial Hypothesis Search

A search conducted over all possible hypotheses can obtain an initial guess of the world's state. If the calibration matrix is unknown, then various candidates for it are tested. For a given candidate calibration matrix, the number of initial guesses for the VIOA, Q, is bounded by $\binom{n}{2}$, which is the size of the 3D set of points $Q_{ij} \in \mathbb{R}^3$, where Q_{ij} is the middle of the shortest line segment that connects the attention rays of the i'th and j'th observers.

We use a RANSAC algorithm (MLESAC flavor, [26]) to restrict the hypothesis search even more; see Alg. 2. When the calibration matrix is unknown, the hypothesis search is repeated for several initial guesses of K and the best one is selected. An initial random guess of an inlier subset, I, is used in a RANSAC iteration to produce an estimation for the MAWE's world state from $\{\hat{U}_i\}_{i \in I}$. The estimation includes the VIOA (Q) that is computed by plane intersection (Section 4.1) and possibly, if the planar constraint is assumed, also the head's plane parameters L; see Section 5.1.3. If the planar constraint is assumed, then for the i'th observer $\hat{r}_i$ is computed from K, L and, $\hat{U}_i$ ($\hat{p}_i$). In addition, $\hat{\alpha}_i$ and $\hat{\beta}_i$, are computed from K, Q and the other members of $\hat{U}_i$ ($\hat{\gamma}_i$, $\hat{r}_i$ and, $\hat{p}_i$). This is done using a closed form solution that is shown in Section 5.1.2. The most probable observer subset, I, and the score for the MAWE hypothesis are computed according to the theory in Section 5.1.1. The subset, I, is selected by thresholding the probability such that at least 95% of the inliers will be included, i.e., $T_{\mathrm{LLR}} = 1.96$ standard deviations. The resulting inlier subset is then reused to re-estimate Q, L, $\hat{\alpha}$, $\hat{\beta}$, I and

procedure ESTIMATE-VIOA-K($K_{\text{range}}, U, I_0, \Sigma, T_{\text{LLR}}, T_{\text{miss}}, T_{\text{min}}$) // **For more details see:**
  for $K \in K_{\text{range}}$ //
    for 1 **to** MLESAC-ITERATION-NO($\Sigma, T_{\text{LLR}}, T_{\text{miss}}, T_{\text{min}}$) // sample and inliers
      $I \leftarrow$ GET-RANDOM-SUBSET(I_0, T_{min}) // (13) and (15)
      repeat //
        $U' \leftarrow U$ //
        $[(L, U'.r) \leftarrow$ EVAL-PLANAR-GROUND$(K, U, I)]$ // 5.1.3 and (12)
        $Q \leftarrow$ ESTIMATE-VIOA(K, U', I) // 4.1, (4), and (5)
        $(U'.\alpha, U'.\beta) \leftarrow$ GET-YAW-PITCH(K, Q, U') // 5.1.2, (10), and (11)
        $(\text{score}, I) \leftarrow$ GET-SCORE$(K, Q, U, U', \Sigma, T_{\text{LLR}})$ // 5.1.1, (9)
      until not (UPDATE-BEST$(Q, U'[, L], I, \text{score}, \underbrace{Q', U''[, L'], I', \text{score}'}_{\text{best so far}})$) //
      // Optimizing
    $(Q', U''[, L'], I', \text{score}') \leftarrow$ GLOBAL-TUNE$(K, Q', U, U''[, L'], I', \Sigma, T_{\text{LLR}})$ // over Q [and L]
    UPDATE-BEST$(K, Q', U''[, L'], I', \text{score}', \underbrace{K', Q'', U'''[, L''], I'', \text{score}''}_{\text{best so far}})$ //
  output $(K', Q'', U'''[, L''], I'', \text{score}'')$ //

Alg. 2: Hypothesis search algorithm.

to compute a new score. This enhancement iteration is repeated as long as the score improves. After a sufficient number of RANSAC iterations (determined according to Section 5.3) a general optimization is applied to refine the parameters that are shared by all inliers, i.e. Q and L. The result is a MAWE hypothesis with the maximal probability.

5.1.1 The Bayesian Setup

For simplicity, the Bayesian setup will address only a single MAWE, i.e. the most significant MAWE, $\text{MAWE}_{1/1}$. It can be generalized to address multiple MAWEs at once but the goal here is to decide whether a MAWE occurred. The MAP estimator of the world state regarding the largest MAWE is:

$$\mathcal{W}_{\text{MAP}} = \arg\max_{K,Q,\hat{U}} \mathfrak{p}\left(K, Q, \hat{U} \middle| U\right), \tag{6}$$

where $\hat{U}$, K, and Q are estimated such that (2) is satisfied (and optionally (12). The formulation in Section 5.1.2 is used to guarantee that (2) is satisfied and the formulation in Section 5.1.3 is used to guarantee that (12) is satisfied. Assuming that U_i is independent of U_j, for every selection of i and j such that $i \neq j$, the above becomes:

$$\mathfrak{p}\left(K, Q, \hat{U} \middle| U\right) = \mathfrak{p}\left(K, Q, \hat{U}\right) \cdot \prod_i \frac{\mathfrak{p}\left(U_i \middle| K, Q, \hat{U}_i\right)}{\mathfrak{p}\left(U_i\right)}. \tag{7}$$

The measurements may include heads that do not participate in the MAWE, i.e., outliers. There are a few reasons for the existence of outliers, including: false head detections, blocked attention rays, and observers who do not participate in the MAWE. In order to model this, let I_i be a binary random variable that is 1 if the i'th observer participates in the MAWE, i.e., that observer is an inlier. By conditioning w.r.t. I_i we can write $\mathfrak{p}\left(U_i \middle| K, Q, \hat{U}_i\right)$ in (7) as:

$$\begin{aligned}&\mathfrak{p}\left(U_i \middle| K, Q, \hat{U}_i, I_i = 1\right) \cdot \mathfrak{p}\left(I_i = 1 \middle| K, Q, \hat{U}_i\right) + \\ &\mathfrak{p}\left(U_i \middle| K, Q, \hat{U}_i, I_i = 0\right) \cdot \mathfrak{p}\left(I_i = 0 \middle| K, Q, \hat{U}_i\right).\end{aligned} \tag{8}$$

Thus, the data is represented as a mixture of distributions: distributions which correspond to MAWEs and the background distribution. The distribution of the measurements, U_i, of the i'th participant in a MAWE, $\mathfrak{p}\left(U_i \middle| K, Q, \hat{U}_i, I_i = 1\right)$, is modeled as having random noise normally distributed around $\hat{U}$, i.e. $U \sim N\left(\hat{U}, \Sigma_i\right)$ with a diagonal covariance matrix Σ_i. The diagonal of Σ_i is $\left(\sigma_{i\alpha}^2 \quad \sigma_{i\beta}^2 \quad \sigma_{i\gamma}^2 \quad \sigma_{ir}^2 \quad \sigma_{ix}^2 \quad \sigma_{iy}^2\right)$, where $\sigma_{i\star}$ are the standard deviations of the errors of the detection algorithms w.r.t. image head pose, image face radius and image face position respectively. The standard deviations for the i'th observer can be estimated empirically by applying the selected algorithms for face detection and pose estimation on a reference data-set with known ground-truth values. If one of the head pose angles cannot be measured by the head pose algorithm, it is selected as zero with a large standard deviation.

Heads that do not participate in the MAWE are considered as background heads. The distribution of the measurements of those heads is modeled as randomly selected from a uniform distribution, i.e. $\mathfrak{p}\left(U_i \middle| K, Q, \hat{U}_i, I_i = 0\right) = 1/\upsilon$, where υ is just a constant that reflects the possible ranges of the quantities in $\hat{U}_i$.

We assume that being an inlier or an outlier is independent w.r.t. K, Q and, $\hat{U}$, i.e., $\mathfrak{p}\left(I_i = 0/1 \middle| K, Q, \hat{U}_i\right) = \mathfrak{p}\left(I_i = 0/1\right)$. The two priors $\mathfrak{p}\left(I_i = 1\right)$ and $\mathfrak{p}\left(I_i = 0\right)$ are generally unknown. Following [26], the priors can be selected to maximize $\mathfrak{p}\left(U_i \middle| K, Q, \hat{U}_i\right)$. Let the MAWE prior be denoted by $\mathfrak{p}\left(I_i = 1\right) = \lambda$ and the background prior by $\mathfrak{p}\left(I_i = 0\right) = 1 - \lambda$. The optimal value for λ can be found by the EM algorithm suggested by [26]. Thus, the data is explained by a mixture of a multivariate normal distribution and a multivariate uniform distribution. Each multivariate normal distribution corresponds to a significant MAWE w.r.t. to the background distribution. The MAP estimator for the world state of the largest

MAWE, $\mathcal{W}_{\text{MAP}}$, is given by

$$\arg\max_{K,Q,\hat{U}} \sum_i \log\left(\lambda \cdot \frac{e^{-\frac{1}{2}\cdot(U_i-\hat{U}_i)^T\cdot\Sigma_i^{-1}\cdot(U_i-\hat{U}_i)}}{(2\cdot\pi)^{n/2}\cdot|\Sigma_i|^{1/2}} + \right. \\ \left. + \frac{1-\lambda}{\upsilon}\right) + \log\left(\mathfrak{p}\left(K,Q,\hat{U}\right)\right). \tag{9}$$

The prior $\mathfrak{p}\left(K,Q,\hat{U}\right)$ represents reasonable values of the calibration matrix and reasonable values for Q and $\hat{U}$. Note that in the context of a specific scene type, a different model, that represents the priors of the specific scene, can be used instead of the uniform distribution.

5.1.2 Estimating the Yaw and Pitch Angles from Q

Given an intersection point Q and an estimated head position, P_i, the corrected head pose (the yaw and pitch angles) can be computed. The coordinate system of the i'th head when directed towards the camera is represented by the unit vectors, $\begin{pmatrix}\bar{\mathfrak{X}}_i^O & \bar{\mathfrak{Y}}_i^O & \bar{\mathfrak{Z}}_i^O\end{pmatrix}^T$. They are computed as,

$$\bar{\mathfrak{Z}}_i^O = -\frac{P_i}{\|P_i\|_2} = -\frac{K^{-1}\cdot p_i}{\|K^{-1}\cdot p_i\|_2}, \bar{\mathfrak{X}}_i^O = R_{\text{roll},i}\cdot\frac{\bar{\mathfrak{Z}}_i^O\times\bar{\mathfrak{Y}}}{\|\bar{\mathfrak{Z}}_i^O\times\bar{\mathfrak{Y}}\|_2},$$

and

$$\bar{\mathfrak{Y}}_i^O = \bar{\mathfrak{Z}}_i^O\times\bar{\mathfrak{X}}_i^O,$$

where $\bar{\mathfrak{Y}}$ is a unit vector in the direction of the Y axis of the main coordinate system (the camera coordinate system), i.e., $\bar{\mathfrak{Y}} = \begin{pmatrix}0 & -1 & 0\end{pmatrix}^T$. Rotating $\bar{\mathfrak{Z}}_i^O$ yields

$$\bar{\mathfrak{Z}}_i^Q = \frac{Q-P_i}{\|Q-P_i\|_2} = R_{\text{pitch},i}\cdot R_{\text{yaw},i}\cdot R_{\text{roll},i}\cdot\bar{\mathfrak{Z}}_i^O.$$

The estimators for the yaw and pitch angle, $\hat{\alpha}_i(K,Q,p_i,r_i)$ and $\hat{\beta}_i(K,Q,p_i,r_i)$, are given by

$$\hat{\alpha}_i = \tan^{-1}\left(\frac{\left\langle\bar{\mathfrak{Z}}_i^Q,\bar{\mathfrak{X}}_i^O\right\rangle}{\left\langle\bar{\mathfrak{Z}}_i^Q,\bar{\mathfrak{Z}}_i^O\right\rangle}\right) \tag{10}$$

and

$$\hat{\beta}_i = \tan^{-1}\left(\frac{\left\langle\bar{\mathfrak{Z}}_i^Q,\bar{\mathfrak{Y}}_i^O\right\rangle}{\left\|\begin{pmatrix}\left\langle\bar{\mathfrak{Z}}_i^Q,\bar{\mathfrak{Z}}_i^O\right\rangle & \left\langle\bar{\mathfrak{Z}}_i^Q,\bar{\mathfrak{X}}_i^O\right\rangle\end{pmatrix}^T\right\|_2}\right). \tag{11}$$

In the above, the inner product $\mathfrak{V}_1^T\cdot\mathfrak{V}_2$ is denoted by $\langle\mathfrak{V}_1,\mathfrak{V}_2\rangle$.

5.1.3 Planar Approximation

The depth can be corrected when the heads are assumed to lie approximately on a plane in $\mathbb{R}^3$. The plane will be represented by the vector $L = \begin{pmatrix}a & b & c\end{pmatrix}^T$. A point $P = \begin{pmatrix}X & Y & Z\end{pmatrix} \in \mathbb{R}^3$ is on the plane iff $P^T\cdot L = 1$. Dividing by the depth and using $P = Z\cdot K^{-1}\cdot p$, the constraint can be written as $\frac{(K^{-1}\cdot Z\cdot p)^T\cdot L}{Z} = \frac{1}{Z}$. Let matrix $\mathcal{P} = \begin{pmatrix}p_1 & p_2 & \dots & p_n\end{pmatrix}$ be the image positions of all the heads. Then, the following is obtained:

$$(K^{-1}\cdot\mathcal{P})^T\cdot L = \begin{pmatrix}\frac{1}{Z_1} & \frac{1}{Z_2} & \dots & \frac{1}{Z_n}\end{pmatrix}. \tag{12}$$

By applying the pseudo-inverse, $\left((K^{-1}\cdot\mathcal{P})^T\right)^\dagger$, on (12), the parameters of the plane are estimated and then a new estimation for the depths is obtained. The plane and the corrected depths are reevaluated whenever the estimates for K or U change.

5.2 The Optimization

The optimization is performed in two steps, a global optimization and a single head based optimization.

In the global optimization, Q, K and the plane of all heads (if assumed) are fine-tuned to maximize (9). In the head based optimization phase, the image position, roll angle, and face radius (if no plane is assumed on all heads) are fine-tuned to maximize (9) for every head separately. This phase tries to compensate for an assumption that is made in the earlier MLESAC phase from efficiency reasons, i.e. those head measurements are used in the MLESAC phase as if they were accurate. If the detected MAWE is found to be significant, then it is reported with all its attributes; otherwise, no MAWE is reported.

When the VIOA is known, e.g., in the special case when the camera is the VIOA, all that is required is to find the observers subset, perform fine tuning, and check the significance of the event.

5.3 Significance Level

Possible causes for false MAWE detections are false head detections, wrong head pose estimations, and independent observers who are not part of a true MAWE. Such events should be discarded while only a significant MAWE with a high confidence should be reported. The significance of the selected observers subset, of size k, is evaluated w.r.t. the background distribution. First, the probability that a detected observer randomly drawn from the uniform distribution is looking at the VIOA at Q will be evaluated, followed by an evaluation of the distribution of a false MAWE detection. Using this distribution, the significance level of a detected MAWE will be evaluated.

The cone of attention of a head intersects with the sphere of radius r around the head in a spherical cap. The uniform distribution assumption implies that every point on the sphere has the same probability. The ratio between the spherical cap area and the entire sphere area is the probability that an attention ray will pass through a specific cone of attention,

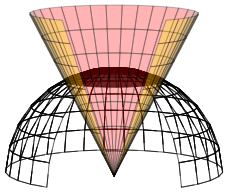


Fig. 4: Spherical cap

$$\text{hit}_{\text{Pr}} = \frac{1}{4\cdot\pi\cdot r^2}\cdot\int_{-T_{\text{LLR}}\cdot\sigma_\beta}^{T_{\text{LLR}}\cdot\sigma_\beta}\int_{-T_{\text{LLR}}\cdot\sigma_\alpha}^{T_{\text{LLR}}\cdot\sigma_\alpha} r^2\cdot\cos(\alpha)\,d\alpha\,d\beta = \\ = \frac{\sin(T_{\text{LLR}}\cdot\sigma_\alpha)\cdot T_{\text{LLR}}\cdot\sigma_\beta}{\pi}.$$

As mentioned above, the pitch angle β is often poorly estimated, i.e., $\sigma_\beta = \pi$. In this case the probability is:

$$\text{hit}_{\text{Pr}} = \frac{1}{2\cdot\pi\cdot r}\cdot\int_{-T_{\text{LLR}}\cdot\sigma_\alpha}^{T_{\text{LLR}}\cdot\sigma_\alpha} r\,d\alpha = \frac{T_{\text{LLR}}\cdot\sigma_\alpha}{\pi}. \tag{13}$$

Note that the vast majority of head detection and head pose estimation algorithms are applicable only on half of the sphere (frontal to profile). In this case the probability hit_{Pr} above is multiplied by 2.

The above is used to evaluate the distribution of a false MAWE detection. The number of participants in a MAWE with VIOA at Q is a binomial random variable, $k \sim B(n, \text{hit}_{\text{Pr}})$. The normal distribution $N(n \cdot \text{hit}_{\text{Pr}}, n \cdot \text{hit}_{\text{Pr}} \cdot (1 - \text{hit}_{\text{Pr}}))$ is a reasonable approximation for the distribution of k. Thus, in the case where Q is known, e.g., $Q = 0$, $n \cdot \text{hit}_{\text{Pr}}$ heads are expected to be facing Q, with standard deviation of $\sqrt{n \cdot \text{hit}_{\text{Pr}} \cdot (1 - \text{hit}_{\text{Pr}})}$. Using this normal approximation, the significance of a detected MAWE with k_0 participants can be estimated. For example, if k_0 is larger than $n \cdot \text{hit}_{\text{Pr}} + 1.96 \cdot \sqrt{n \cdot \text{hit}_{\text{Pr}} \cdot (1 - \text{hit}_{\text{Pr}})}$, then the probability of a false detection is less than 5%.

Determining the confidence of the detected MAWE, when Q is unknown, is not an easy task. Following are the arguments that support this claim and a possible solution. When Q is unknown, the algorithm will find a position with the highest likelihood. The likelihood increases when more observers have their attention ray near Q. Thus, when the measurements come from a uniform distribution, the expected number of observers in the detected MAWE will be larger than the expected number for a known Q, which is $n \cdot \text{hit}_{\text{Pr}}$. However, it is not easy to compute this expected number since the different events for different VIOA positions are statistically dependent. Furthermore, the dependency between possible VIOA positions increases when the standard deviation in the pose increases. A related problem which demonstrates the analytic difficulty in defusing the dependency among possible VIOA positions is studied by [27], [28]. Nevertheless, a rule of thumb regarding the normal distribution can be used to decide whether a detected MAWE is significant: the number of observers in $\text{MAWE}_{1/1}$ is likely to be larger than the average number by not more than $T_{\text{MA}} = 3$ times the standard deviation, i.e.,

$$k_{\max} \approx n \cdot \text{hit}_{\text{Pr}} + T_{\text{MA}} \cdot \sqrt{n \cdot \text{hit}_{\text{Pr}} \cdot (1 - \text{hit}_{\text{Pr}})}. \quad (14)$$

This observation is empirically demonstrated in Section 6.2 and illustrated in Figure 6. Note that for a false MAWE, that is generated at random, the number of inliers is much smaller than in the case of a true MAWE. The two cases can be seen in Figure 6(b) for a false MAWE and Figure 6(d) for a true MAWE.

The minimal number of MLESAC iterations such that will guarantee a miss probability that is less than T_{miss} was computed according to [26], yielding:

$$\frac{\log(1 - T_{\text{miss}})}{\log\left(1 - \left(\frac{\sin(T_{\text{miss}} \cdot \sigma_\alpha) \cdot T_{\text{miss}} \cdot \sigma_\beta}{\pi}\right)^{T_{\min}}\right)}, \quad (15)$$

while in every MLESAC iteration $T_{\min}$ observers are tested. For example, if significant MAWEs are those with at least 35% of the detected heads and the required miss probability is not less than 99%, then 105 MLESAC iterations are required while using $T_{\min} = 3$.

6 Experiments

In this section the implementation of the method is presented together with an analysis of its results when applied on real and simulated data. This analysis includes qualitative and quantitative evaluation of the results and their accuracy. First the method's implementation details are given, including an evaluation of the accuracy of the selected head pose algorithm the details of our test dataset; see Section 6.1. An evaluation of method's accuracy in detecting MAWE is presented in Section 6.2. Section 6.3 discuss the accuracy of the method in estimating MAWE's attributes. The last section presents the output of the method when applied on several specific cases; see Section 6.4.

6.1 Implementation

The method was implemented in C++ using the OpenCV library [21] and OCTAVE under Linux. First the head pose algorithm of [17] was applied to the image and its mirror. In addition, the OpenCV implementation of the Viola-Jones algorithm was applied (profile and frontal head poses) to the image (and its mirror) to refine the estimate of the size of each head. False detections of heads or gross errors in head pose estimation were observed. To address this, the algorithm in Section 5 was used, where the size of a subset drawn by the RANSAC algorithm is $T_{\min} = 3$. To detect MAWEs with at least 45% inliers, 30 RANSAC iterations were used. The optimization phase (global, local) was repeated twice with 250 iterations in both local and global optimizations. The covariance matrix of the various attributes was estimated using a reference database of known head poses, the INRIA head pose database [29]; see Figure 5 for examples. Specifically: $\sigma_\alpha = 11°$, $\sigma_\beta = 30°$, $\sigma_\gamma = 4°$,$\sigma_x = ((1 - \cos(\alpha)) \cdot 0.25 + 0.1) \cdot r$,$\sigma_y = ((1 - \cos(\alpha)) \cdot 0.45 + 0.07) \cdot r$, and $\sigma_r = ((1 - \cos(\alpha)) \cdot 0.15 + 0.05) \cdot r$. The threshold for rejecting an observer from a MAWE due to measurement noise was set to 5% $(1.96 \cdot \sigma)$. In order to presents our results in meters we used the approximated face radius $\delta = 72mm$; see the Appendix for more details.

6.1.1 Data-set

We used both real and simulated data to test our method. The real data includes images obtained from the Internet of scenes which contain a single MAWE. The simulated data includes simulated measurements that come from one of the two distributions: a uniform distribution for background heads and for the heads in a MAWE a normal distribution around the correct ones. The method was applied to both data sets, reporting whether a significant MAWE took place and estimating its attributes. Those attributes include: (1) the inliers, i.e., the subset of heads participating in the event, and (2) scene structure in meters (the VIOA position and inlier positions). For

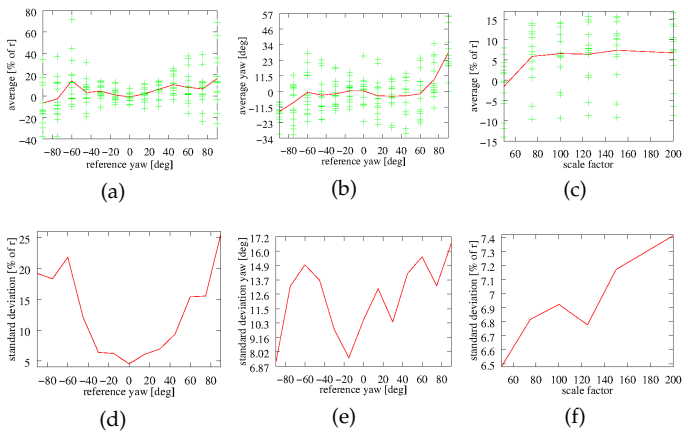


Fig. 5: Several examples from the statistical characterization of the measurements over the INRIA head pose database. The first column shows the statistics of the measured head radius w.r.t. the known yaw angle. The second column shows the statistics of the measured yaw angle w.r.t. the known yaw angle. The third column shows the statistics of the measured head radius w.r.t. a known scale factor (percent of the original images). The first row shows the values of a certain measurement for the 15 people in the database and their average. The second row shows the corresponding standard deviations.

the case where the VIOA is not the camera, the method was used to estimate the camera's focal length and compare it with the known one. All images were tested with and without the planar constraint to estimate its contribution.

We tested the method's accuracy in detecting a significant MAWE and its accuracy in estimating the event's attributes. The simulated data was used to test how the method's accuracy is affected by the accuracy of its input, by simulating different inliers-outliers ratios and different levels of added noise to the head pose angles.

The set of images includes 26 images with the camera as their VIOA (known Q) and 59 images whose VIOA is not the camera (unknown Q). Common assumptions on the perspective model were made, i.e., no skew, the principal point is at the center of the image, and the ratio between f_x and f_y is known (in our case 1). In order to test both known and unknown K, only images with a known focal length (from the *exif* data of the images) were used.

6.2 MAWE Detection Accuracy

The method reports on a MAWE if the percentage of detected inliers out of all heads is above a threshold. Thus, the accuracy in the detection of a MAWE is mainly affected by the number of outliers. The noise in the inliers' measurement is much less dominant than the number of outliers in the detection of significant MAWEs. The above is demonstrated in Figure 6, on simulated data (a-d) and in Figure 7 on real data. Figures 6(a,b, and c) show that the portion of inliers in $\text{MAWE}_{1/1}$ matches the binomial distribution for the case where all the measurements come from a uniform distribution (only false detections of insignificant MAWEs are possible). These graphs, together with the fact that the normal distribution approximates well the binomial distribution, motivates us to determine the significance threshold for a detected MAWE: the number of observers in $\text{MAWE}_{1/1}$ is likely to be larger than the average number by not more than 3 times the standard deviation; see Section 5.3 for more details. For the case of a true MAWE with outliers, the portion of inliers in $\text{MAWE}_{1/1}$ is significantly larger than the above mentioned case of insignificant MAWEs. This portion approaches 100% when there are no outliers; see Figure 6(d). On real data Figure 7 shows, under different assumptions, the ROC curve that quantitatively summarizes the accuracy in which an observer might be detected as part of a significant MAWE. A MAWE was classified as significant when the number of its participants is above $k_{\max}$ in (14) for $T_{\text{MA}} \in [0, 20]$. Note that the ROC curve do not account for the thresholds that are involved with head and head pose detection and the T_{LLR} threshold. We used $T_{\text{MA}} = 1.96$ to determine whether a MAWE is significance. With the current head pose estimation algorithm and its error, true MAWEs in images with less than 30 observers might be classified as insignificant.

The ground truth of the MAWEs in the real images and their corresponding participants were manually classified. Most of the events were detected as significant MAWEs with a small portion of false inliers ($\approx 5\%$) and with a high portion of inliers ($\approx 70\%$ from the ground truth and $\approx 85\%$ w.r.t. the detected heads in the head detection phase). The MAWEs that appear as insignificant are mostly due to poor performance of the head detection phase on those specific images.

Allowing the method to search for the optimal focal length (ignoring the *exif* data) slightly improves the significance of the detected MAWE. Another test case was when enforcing the planar constraint on the observers. The results of this case show a better scene structure reconstruction while maintaining a similar detection level. Allowing the method to search for the optimal VIOA position, for images whose VIOA is at the camera, slightly improves the significance of the detected MAWE.

6.3 Attribute Estimation Accuracy

One way to evaluate the accuracy and quality of the estimated attributes is by their dispersion. If the dispersion is low then the estimation can be reliably used. In the previous section it is shown that the estimated subset of inliers have a low dispersion and thus the method reliably reports whether a head is part of a MAWE or not. The dispersion of K and Q can be evaluated by numerical computation of their covariance matrix using uncertainty propagation, e.g. [30]. However, this kind of dispersion evaluation is required for every instance specifically and does not give a strong intuition for new instances.

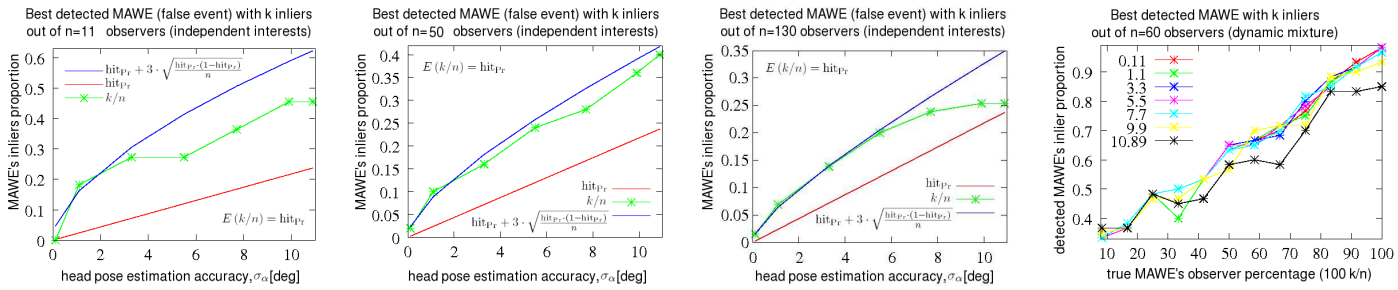


Fig. 6: The percentage of inliers in $\text{MAWE}_{1/1}$: (a-c) Simulation of false detection of MAWEs. All head attributes were generated from a uniform distribution. A MAWE emerging from this data is a false event. The algorithm was applied to the data multiple times while using different values for the stddev of the yaw angle, $0^\circ < \sigma_\alpha \leq 11^\circ$. The x axis is the stddev of the yaw angle. There are 11 heads in (a), 50 in (b) and 130 in (c). The average and variance of the number of inliers decrease with the number of heads. (d) Simulation of a single MAWE with varying proportions between inliers and outliers, where the sum of the number of inliers and outliers was 60. The simulation begins with 5 inliers; that number increases by jumps of 5 until 60. For a selected number of inliers, 7 different stddevs of the yaw angle (0.11,1.1,...,9.9,10.98) were used to generate the attributes of inliers from a normal distribution with the stddev in turn. The attributes of outliers were generated from a uniform distribution. The method was applied to all the 84 events with a fixed stddev estimation of $\sigma_\alpha = 11^\circ$. The x axis is the percentage of inliers out of the 60 heads.

When considering only the observers and their attention rays then the estimation of Q is a triangulation problem, similar to 3D reconstruction from multiple views. The accuracy of a similar situation is analyzed in [31], in the context of the triangulation problem in stereo vision. The uncertainty in estimating Q is the same as the uncertainty of reconstruction in the triangulation context; see Figure 12.6 in [31][pp.321]. In our context it implies that the VIOA, Q, is less precisely localized along an attention ray as the attention rays become more parallel. This suggests an important observation regarding the geometrical structure of the dispersion of Q estimation. The estimation dispersion of Q should be viewed relative to the observers and not relative to the position of the capturing camera. This is because the angle between the attention rays becomes larger when Q come closer to the observers and becomes smaller when it is distant.

We illustrate the uncertainty region in a similar figure, Figure 8. The dispersion of Q estimation has a shape that depends on the distance of the true Q from the centroid of the inliers $\{P_i\}_{i=1}^n$, i.e. $EP = \frac{1}{n}\sum_{i=1}^n P_i$. A sequence of areas one inside the other with a growing probability to contain Q, is shown in Figure 8 for increasing values of $\|Q - EP\|_2$ demonstrating this. When Q is close to EP the area is a small polygon (almost symmetric). When Q is farther from EP the area becomes a narrow polygon around the ray from EP to Q. Thus, the estimated Q has a high probability to be in the right direction from EP but the distance from EP is reliable only when Q is close to EP. When Q is very far from EP the area still seems symmetric around the Q-EP ray but is unbounded. Our other experiments strongly agree with this observation.

The dispersion of attention rays around Q is affected by the calibration matrix, K. For accurate measurements all attention rays intersect at Q only for the correct value of K. Other values for K yield multiple pairwise intersection points. Similarly, when the measurements are noisy the area of dispersion of the pairwise intersection points is the smallest near the correct K and becomes larger with the diversity level in the estimated K. The distance of Q from EP has a strong effect here also because when Q is close to EP the dispersion of the points of pairwise intersections have clear boundaries while when Q get farther from EP the area is unbounded and the correct K cannot be estimated well.

The implication of the above discussion is that if Q is close to EP we expect accurate estimation of Q and K. Using those estimations a reliable correction can be applied back to image measurements. However,

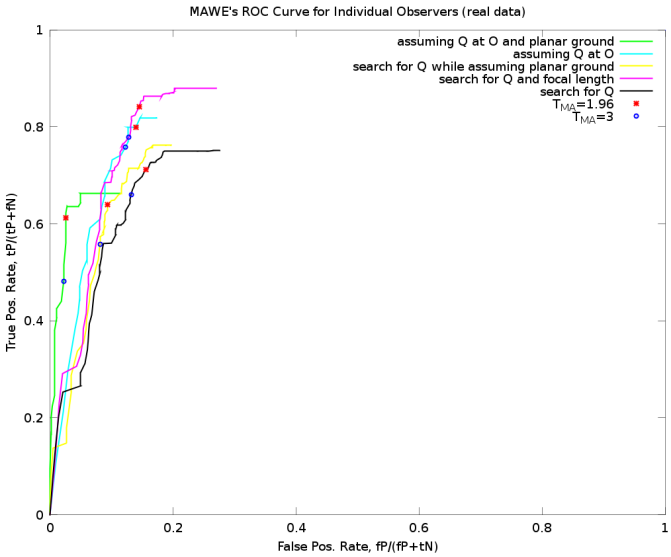
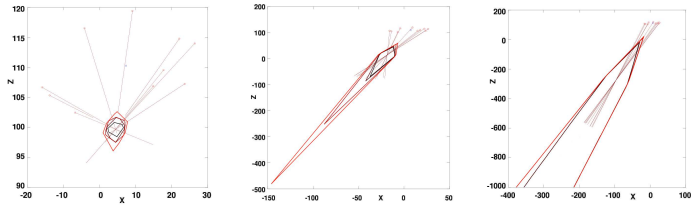


Fig. 7: A ROC that summarizes the accuracy in which an observer might be detected as part of a significant MAWE. The ROC was obtained by applying the method on the real image data-set, which contains more than 1000 observers, using various significance levels, i.e. $T_{\text{MA}} \in [0, 20]$.



(a) $\|\cdot\|_2 = 11$ (b) $\|\cdot\|_2 = 100$ (c) $\|\cdot\|_2 = 400$

Fig. 8: A sequence of areas one inside the other with a growing probability to contain Q. Four cases are given for different values of $\|Q - EP\|_2$, i.e. 11, 100, and 400 (K and P_i are fixed).

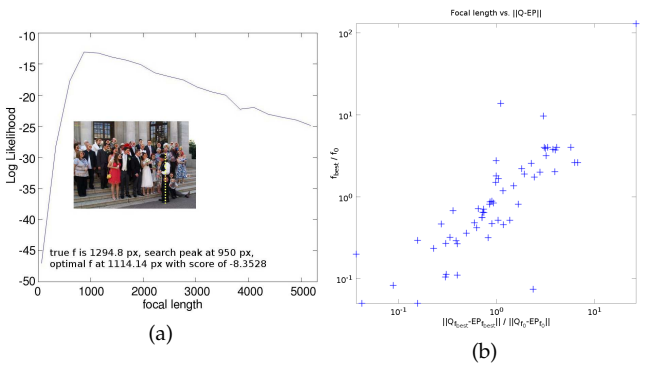
when Q is farther from EP there are only two reliable estimations: the direction of Q along the Q-EP ray and the inliers-outliers dichotomy.

Therefore, we have found it intuitive to consider a single virtual observer located at EP, while virtually observing Q. In this way, the virtual observer has an equivalent attention ray $EA = \frac{1}{n}\sum_{i=1}^{n} \zeta_i \cdot R_i \cdot P_i$ that by (1) satisfies $Q = EP + EA$. On real images we demonstrate the above intuition by measuring the error in Q-EP using the angle between the estimated equivalent attention ray and the true equivalent attention ray. It was found that the standard deviation of this mutual error in the yaw angle was 3° as opposed to 11° for a single observer. In the simulations, the error in Q-EP was much smaller and the difference might be explained by a larger average number of inliers and the fact that the noise in real images is correlated (statistically dependent) due to the contribution of pupil corrections.

The focal length was estimated in two steps. The first step was a crude search over 20 possible values of the focal length ($f \in [80, 80000]$). The second step was a fine tuning using the global optimization phase. The typical behavior of the first step can be seen in Figure 9(a). It can be seen that the log likelihood graph peaks notably near the true focal length. It is clear that running time is much faster with a known focal length.

The estimations of the focal length, f, and the distance of Q from EP are highly correlated; see Figure 9(b). Selecting a larger estimation for f will result in a larger Q-EP distance. In addition, the estimation of the focal length is more accurate for smaller Q-EP distances; see Figure 8. As a consequence, it is expected to reconstruct the scene more accurately for smaller Q-EP distances.

There was no ground truth for the position of the observers but the results seem qualitatively accurate, especially when using the planar constraint; see Figures 10 and 11.



(a) (b)

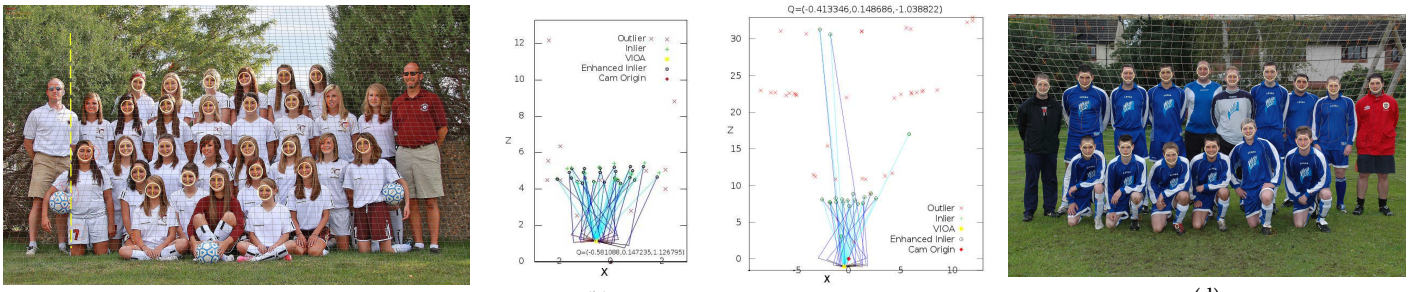
Fig. 9: (a) Typical behavior of the score for different values of the focal length. It is common to get a unique maximum near the true focal length when the measurements noise is small. (b) Focal length estimation accuracy on real images. The accuracy of the estimated focal length and the accuracy of the distance between Q and EP are highly correlated.

6.4 The General Results

The method was tested on images of scenes with a single MAWE obtained from the Internet. A small fraction of the results is shown in Figures 10, 11, 12, 13, and 14. The entire set of results can be found online; see [32]. In each image, each detected person's face is indicated by a red circle. The radius of the circle indicates the size of the face, which is proportional to its distance from the camera. The enclosed triangle indicates the yaw, pitch and roll angles. A face is indicated by an additional white circle if it is an inlier. Since it is quite common that observers are captured in a vertical posture, the rotations w.r.t. the yaw and pitch angles are aligned with the x and y axes of the camera coordinate system. As the pitch angle is not estimated well by the head pose estimation algorithm, the detected Q has an inaccurate Y position. Thus, a vertical dashed yellow line was placed along the x position of the projected VIOA, and two other dotted yellow parallel lines mark the size of a virtual head that is placed at the estimated Q. When the estimated Q was accurate enough then its projection on an image was marked by a bright circle whose size is the size of a virtual head that is placed at Q. Near each image is a

(a) (b) (c)

Fig. 10: (a) An image of a MAWE of a large group (83 observers) with a VIOA outside the field of view of the camera. The image was taken with a focal length $f = 22.2$mm. (b) When no plane is assumed, the method detects a $\text{MAWE}_{1/1}$ with 71 inliers. The estimated focal length is $f = 18.4$mm and the estimated position of the VIOA is at $Q = \begin{pmatrix} -7.453\text{m} & -0.281\text{m} & 14.115\text{m} \end{pmatrix}$. (c) When a plane is assumed, the method detects a $\text{MAWE}_{1/1}$ with 66 inliers. The estimated focal length is $f = 18.3$mm and the estimated position of the VIOA is at $Q = \begin{pmatrix} -6.736\text{m} & -0.235\text{m} & 16.796\text{m} \end{pmatrix}$. See text for more details.

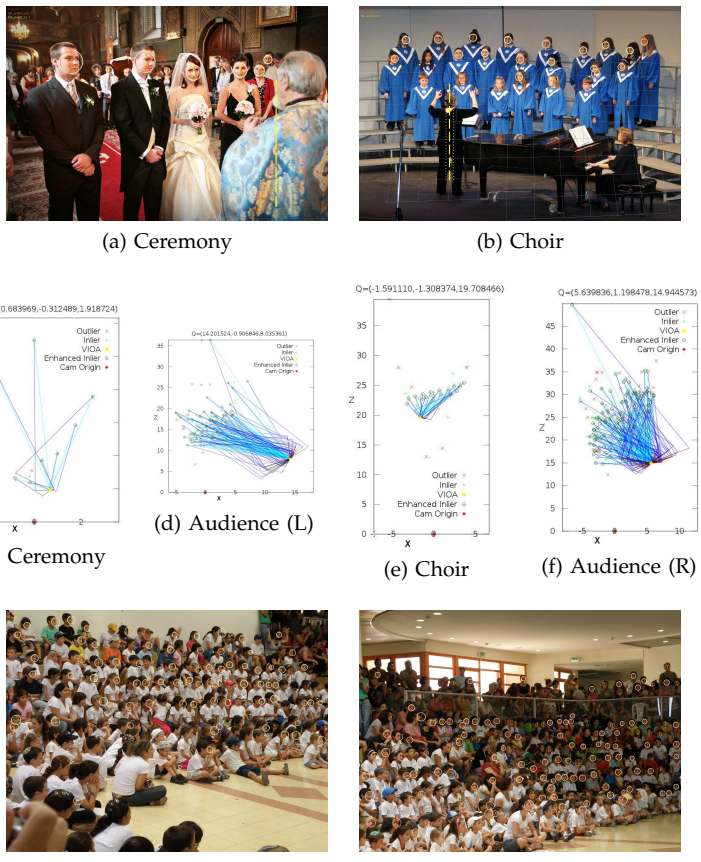


(a) (b) (c) (d)

Fig. 11: MAWE images with a VIOA close to the camera. (a-b) At first glance it might appear that this group is mutually aware of the camera that took this image. But, on second glance, they seem to be looking to the left of that camera. The method estimates the VIOA position about 60cm to the left, which suggests that another person (camera?) is very close to the one who took the picture. (c-d) The group members in this image look into the camera (their VIOA). The method correctly segregates inliers and outliers even when the head pose detector finds a large number of false heads.

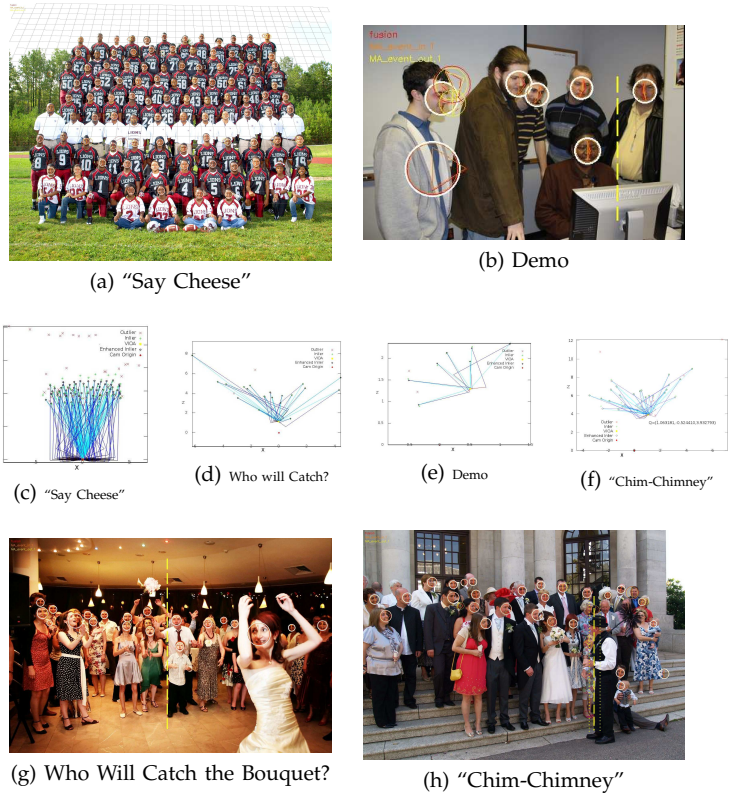
scheme of the structure of the reconstructed scene, where all units are in meters. For clarity, only the projection on the XZ plane is shown for most images. In each box an inlier is marked by a black circle at its corrected 3D position and a cyan '+' at its measured 3D position. The blue and cyan lines are the measured and corrected attention rays respectively. A yellow square was placed at Q and is the intersection of the red lines that connect it with the nearest point on each blue attention ray. The red rhombus at the XZ origin $(0, 0)$ is the position of the camera. An outlier is marked by a red 'x' at its measured 3D position.

VIOA at the Origin. Figure 11 shows two images of a group photo of a soccer team. At first glace it seems that in both images all group members are looking at the camera. But careful observation shows that in the first image (Figure 11(a-b)) all group members are looking slightly to the left, probably to another camera. Consider a virtual observer located at EP as above. The angular difference between the virtual equivalent attention ray



(a) Ceremony (b) Choir (c) Ceremony (d) Audience (L) (e) Choir (f) Audience (R) (g) Audience (left) (h) Audience (right)

Fig. 12: Four examples: (a,b,g,h) different cases of MAWEs.

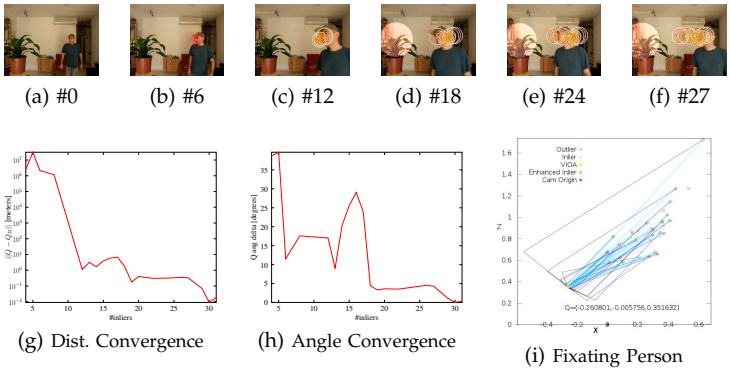


(a) "Say Cheese" (b) Demo (c) "Say Cheese" (d) Who will Catch? (e) Demo (f) "Chim-Chimney" (g) Who Will Catch the Bouquet? (h) "Chim-Chimney"

Fig. 13: Four examples: (a,b,g,h) different cases of MAWEs. (c,d,e,f) scene reconstruction for those cases (in meters). (a,c) the VIOA is the camera, the MAWE was detected, and by using the focal length information in the image file with a planar assumption, the structure of the scene was reconstructed. (b) the VIOA is the computer screen. (g) the VIOA is the bouquet in the air. (h) the VIOA is the chimney sweep.

when facing the estimated VIOA, Q, and when facing the presumed VIOA, the origin O, is used as a confidence measure whether Q deviates from O or is a different point. For the first image the algorithm detected an angle of $7.56°$, which is about 2.5 times the standard deviation that was found empirically, as reported above. Thus, it is less probable that Q is indeed the origin. However, in the second image (Figure 11(c-d)) the angle is $2.74°$, which is less than the standard deviation. Thus, it is reasonable to assume that the group members in the second image are indeed looking directly into the camera that took the image. Note that in the second image most of the faces are outliers that were mistakenly detected on the net behind the group. Nevertheless, $\text{MAWE}_{1/1}$ was detected near the origin, and only a few outliers were mistakenly detected to have joined it.

Plane Enhancement. In Figure 10 a group photo of contestants in a beauty pageant was taken by two cameras (similar to Figure 11(a-b), except that the distance between the two cameras is much larger). The observers were looking towards one camera, while the algorithm was applied on the image taken by the other camera. The detected VIOA is the estimated position of the camera the observers were looking at. The algorithm was applied to the image in Figure 10 twice: the first time while assuming that the heads are on a plane and the second time without that assumption. A grid on the detected plane was projected and drawn on the image. The result when the planar assumption is not made is shown in Figure 10(b), while the result when the planar assumption is made is shown in Figure 10(c). A comparison of the two boxes shows that the planar assumption significantly enhances the recovery of the scene structure, as the four rows of people can be seen clearly. It appears from the scene structure that this group is facing another camera, not seen in the image.



(a) #0 (b) #6 (c) #12 (d) #18 (e) #24 (f) #27

(g) Dist. Convergence (h) Angle Convergence (i) Fixating Person

Fig. 14: A fixating person is walking while keeping his gaze on the plant. The algorithm was applied to 28 images from a video clip of 30 seconds. Several frames are shown in (a-f). The algorithm was applied on every prefix of the 28 images. The attributes of the detected MAWE were estimated with an increasing accuracy as more frames were used, e.g., (g) and (h). After 28 images the attributes are quite accurate. The projection on the XZ plane of the recovered scene when processing the final frame is shown in (i); the units are meters. This 2D projection is a sufficient view of the 3D scene since all attention rays lie approximately on a plane that is parallel to the XZ plane, because the person was walking on a planar floor and image plane was orthogonal to the floor.

VIOA Near the Group. In Figures 12 and 13 there are 8 examples. In Figure 12 the detected VIOA in the first image (a,c) is the minister conducting the ceremony. Due to head pose errors, the depth of the groom was erroneously corrected. The other distances (in meters) seem reasonable without the planar constraint, e.g., a distance of about 2 meters between the minister and the bride and groom (ground truth was not available). In the second image (b,e) a choir is shown and the detected VIOA, which is exactly on the conductor, seems to be about 2.5 meters from the first row. The third (g,d) and fourth (h,f) images show audiences at a show. A subset of the observers appears in both images. However, the detected VIOA is not the same for both and the true position is probably at $Q = \begin{pmatrix}10 & 0 & 10\end{pmatrix}$, which is, for each image, on the Q-EP ray. In Figure 13 a group photo of a football team is shown in the first image (a,c). The method was applied on this image while assuming that the VIOA is known and is the origin. The result is that most of the inliers are indeed part of the group and most of the outliers are not human faces. In the second image (b,e) a group is mutually observing a computer screen. The result on this image yields the exact horizontal and depth position of the VIOA . In the third image (d,g) a bouquet is thrown by the bride and the group is tracking it. The algorithm estimates the position of the bouquet near its correct location. We postulate that if the method was applied on a video of the same event rather than a single image, the result would be much more accurate (see also Figure 14). In the fourth image (h,f) a group is standing on the steps observing a chimney sweep. Applying the method to this image results in the exact horizontal and depth position of the VIOA.

t-MAWE Example. In Figure 14, the method is applied on a video sequence showing a person walking while fixating on a plant. After 18 images, the exact 3D position of the plant (VIOA) was found. The convergence to the correct VIOA, while attention rays are accumulating, is demonstrated in Figure 14(g-h). The convergence can be measured by the distance between the correct VIOA and its current estimation, Figure 14(g), or by the angle between the correct equivalent attention ray, EA, and its current estimation, Figure 14(h). The equivalent attention ray begins at the average position, EP, of the observers who have been seen so far. The position of the VIOA position does not change much when it is near the correct position.

Failure Example. The result of applying the method on an image of a poster session is shown in Figure 15. In this image there are a few separate observer groups (≈ 4), each observing a different poster. The method found a single 3D point can be considered as the mutual virtual interest point of the union of all groups. As the quality of the measurements increases these kinds of failures will become less likely.

REFERENCES

[1] S. Baron-Cohen, "The empathizing system: a revision of the 1994 model of the mindreading system," *a chapter appearing in the book "Origins of the Social Mind" by Ellis, B and Bjorklund, D, (eds) , Guilford Publications Inc.*, 2005.

[2] N. Emery, "The eyes have it: the neuroethology, function and evolution of social gaze," *NBR*, vol. 24, pp. 581–604, 2000.

[3] M. Frank, E. Vul, and S. Johnson, "Development of infants' attention to faces during the first year," *Cognition*, vol. 110, no. 2, pp. 160–170, 2009.

[4] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *IVC*, vol. 27, no. 12, pp. 1743–1759, 2009.

[5] X. L. C. Brolly, C. Stratelos, and J. B. Mulligan, "Model-based head pose estimation for air-traffic controllers," *ICIP*, vol. 2, pp. 113–116, 2003.

[6] M. Farenzena, A. Tavano, L. Bazzano, D. Tosato, G. Paggetti, G. Menegaz, V. Murino, and M. Cristani, "Social interactions by visual focus of attention in a threedimensional environment," in *Workshop on Pattern Recognition and Artificial Intelligence for Human Behaviour Analysis*, 2009.

[7] B. Benfold and I. Reid, "Guiding visual surveillance by tracking human attention," *BMVC*, vol. 1, 9 2009.

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

[9] S. O. Ba and J.-M. Odobez, "A study on visual focus of attention recognition from head pose in a meeting room," *MLMI*, pp. 75–87, 2006.

[10] ——, "Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues," *ICASSP*, 2008.

[11] ——, "Multi-person visual focus of attention from head pose and meeting contextual cues," *PAMI*, vol. 33, no. 1, pp. 101–116, 2011.

[12] L. Dong, H. Di, L. Tao, G. Xu, and P. Oliver, "Visual focus of attention recognition in the ambient kitchen," *CV*, pp. 548–559, 2010.

[13] 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," *PAMI*, vol. 30, no. 7, pp. 1212–1229, 2008.

[14] M. Cohen, I. Shimshoni, E. Rivlin, and A. Adam, "Detecting mutual awareness events," in *ECCV Workshop on Face Detection: Where we are, and what next?*, 2010.

[15] R. Stiefelhagen, M. Finke, J. Yang, and A. Waibel, "From gaze to focus of attention," *VIIS*, pp. 761–768, 1999.

[16] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: A survey," *PAMI*, vol. 31, no. 4, pp. 607–626, 2009.

[17] M. Osadchy, Y. LeCun, and M. Miller, "Synergistic face detection and pose estimation with energy-based models," *JMLR*, vol. 8, pp. 1197–1215, 2007.

[18] P. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features," *CVPR*, vol. 1, pp. 511–518, 2001.

[19] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," *ICIP*, vol. 1, pp. 900–903, 2002.

[20] Y. Wang, Y. Liu, L. Tao, and G. Xu, "Real-time multi-view face detection and pose estimation in video stream," in *ICPR*, 2006, pp. 4:357–360.

[21] "http://www.sourceforge.net/projects/opencvlibrary," 2009.

[22] N. Dodgson, "Variation and extrema of human interpupillary distance," *SDVRS*, vol. 5291, no. 6, pp. 36–46, 2004.

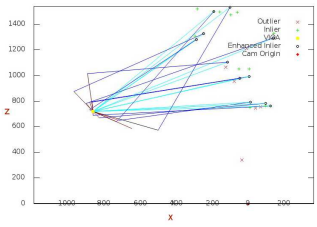
[23] C. BenAbdelkader, R. Cutler, and L. Davis, "Person identification using automatic height and stride estimation," in *ICPR*, vol. 4, no. 16, 2002, pp. 377–380.

[24] A. Gallagher, A. Blose, and T. Chen, "Jointly estimating demographics and height with a calibrated camera," in *ICCV*, 2009, pp. 1187–1194.

[25] K. Kanatani, Y. Sugaya, and H. Niitsuma, "Triangulation from two views revisited: Hartley-sturm vs. optimal correction," in *BMVC*, no. 19, 2008, pp. 173–182.

[26] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *CVIU*, vol. 78, pp. 138–156, 2000.

[27] A. Desolneux, L. Moisan, and J. Morel, "Meaningful alignments," *IJCV*, vol. 40, no. 1, pp. 7–23, 2000.



(a)

(b)

Fig. 15: A case when the method can fail. A specific configuration of a few separate observer groups, each with its own interest point, such that a single 3D point can be accounted as the mutual virtual interest point of the union of all groups.

[28] ——, "Maximal meaningful events and applications to image analysis," *Stat. Ann.*, vol. 31, no. 6, pp. 1822–1851, 2003.

[29] N. Gourier, D. Hall, and J. Crowley, "Estimating face orientation from robust detection of salient facial structures," in *FG Net Workshop on Visual Observation of Deictic Gestures*, 2004, pp. 1–9.

[30] R. Haralick, "Propagating covariance in computer vision," in *ICPR*, 1994, pp. 493–498.

[31] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*, 2nd ed. Cambridge Univ Press, 2004.

[32] M. Cohen, I. Shimshoni, E. Rivlin, and A. Adam, "http://mis.hevra.haifa.ac.il/~ishimshoni/MAW_Events," 2010.

Meir Cohen received the BSc degree in mathematics and computer science from the Hebrew University in Jerusalem and the MSc degree in computer science from the Weizmann Institute of Science, Rehovot, Israel, in 1999. Currently, he is a PhD candidate in the Computer Science Department at the Technion-Israel Institute of Technology. His current research interests are computer vision and machine learning.

Ilan Shimshoni received the BSc degree in mathematics and computer science from the Hebrew University in Jerusalem, the MSc degree in computer science from the Weizmann Institute of Science, Rehovot, Israel, and the PhD degree in computer science from the University of Illinois at Urbana-Champaign. Currently, he is a professor in the Department of Information Systems at Haifa University. His research interests are computer vision, robotics, and computer graphics. He is a member of the IEEE.

Ehud Rivlin received the BSc and MSc degrees in computer science and the MBA degree from the Hebrew University in Jerusalem and the PhD from the University of Maryland. Currently, he is an associate professor in the Computer Science Department at the Technion-Israel Institute of Technology. His current research interests are in machine vision and robot navigation.

Amit Adam obtained his PhD from the Technion-Israel Institute of Technology, in 2001, for a thesis on vision-based navigation. Since his graduation he has been employed as a computer vision researcher and developer, working in areas such as video surveillance and recognition. His interests cover most areas of computer vision.