Detecting Mutual Awareness Events

Meir Cohen[1], Ilan Shimshoni[2], Ehud Rivlin[1], and Amit Adam[1]

[1] Dept. of Computer Science, Technion, Haifa, Israel
[2] Dept. of Management Information Systems, Haifa University, Haifa, Israel

Abstract. It is quite common that multiple observers attend to a single interest point or that a single moving observer is fixating on a static interest point. The current work analyzes those situations and suggests to use it in various applications, including surveillance, social robots and more. The analysis includes the detection that such an interest point does exist, where it is located, who was attending to it and where and when each observer was while attending it.
The least invasive way to monitor those situations is by using a camera that captures the observers while using existing face detection and head-pose estimation algorithms. The current work reviles a surprising relation between fundamental problems in vision and human (social) behavior. This relation is a novel constraint that enables the analysis for the general case of an uncalibrated camera in a general environment. This is in contrast to other works on similar problems that inherently assume a known environment and a calibrated camera.
The only assumption that is made is that the visual attention is aggregating at an interest point from several directions. In addition to the detection of a mutual awareness event the suggested method is able to recover the calibration of the camera, the 3D locations of the observers and the 3D location of the interest point where the attention is aggregating.

1 Introduction

Mutual awareness (MA) is the event where multiple observers attend to a single interest point at the same time. Mutual awareness is a fundamental event when analyzing the social behavior of a group. It indicates a common interest and a common knowledge of the group, thus it was addressed by the psychological community in multiple contexts, e.g. [1–3].

The detection of mutual awareness events and their attributes can assist various applications. The least invasive way to monitor those situations is by using a camera that captures the 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 MA events and their attributes. In fact, it covers also an even more general problem that takes into account the time domain. That is when a single static interest point is the visual focus of attention (VFOA) of several observers at different times. This generalization will be noted by *temporal mutual awareness* (TMA). The related interest point

of both MA and TMA will be noted as the *Visual Intersection Of Attention* (VIOA). The proposed method detects an MA event with its attributes from a single image (simple MA) or from video. Those attributes include its VIOA, the observers and their spatiotemporal locations.

Various applications could be assisted by detecting an MA event and its attributes. One domain for applications is surveillance systems. A surveillance system may be assisted by passersby to enhance its abilities by using the detection of an unusual MA event to indicate that an important event was detected by the observers. In addition, it is often that an antisocial behavior of an individual includes exceptional attention patterns. In particular, an exception to the MA should be checked. For example, while waiting at a crosswalk for the light to change, a pickpocket will look at bags and wallets and not at the traffic light as the others. In addition, the intention of a single observer may be estimated while he is moving and fixating on a single point.

Another domain is robotics. When a social robot enters a new place it can obtain valuable information just by detecting the VIOA points where human attention is aggregating at.

Support-systems applications may include the online analysis of a team's attention such as: a choir, orchestra members, a basketball team, actors, dancers, etc. Moreover, an analysis of an audience's attention can assist quality assurance. For example, in a lecture, it may help to address non-attentive parts of the audience. In a theater or a show it may help rate the performance of the actors/dancers.

Other application domains that could be assisted by the proposed method, include: content management, advertisement, architecture and interior design, psychology and sociology.

The analysis of the MA event using a single camera depends on three types of information: (1) the 3D structure of the scene (the positions of observers, camera and the VIOA point), (2) the camera's internal calibration matrix and (3) the detection of observers and their gaze direction from the images.

The research so far used gaze estimation to infer the visual focus of attention (VFOA) of a single independent observer while assuming a known environment and a calibrated camera. The known environment generally includes a specific setup of observers (sitting persons), a fixed camera position and a predefined set of possible interest points. For example a meeting room with two fixed cameras, four sitting persons and a set of six possibly attended targets including a screen, a table, and the persons [4, 5].

The setup addressed by previous research is suitable in meeting rooms or other controlled environments but not for the general case when the observable targets can't be determined in advance and the observers can change their positions. This work will close the gap by handling a broader set of setups when the position of the observers, the camera and the attended targets are not known in advance.

This work shows how an uncalibrated camera can be used to detect the VIOA position in a general environment. The only assumption is that the visual

attention is aggregating at a point from several directions. In addition to the detection of a mutual awareness event the suggested method is able to recover the calibration of the camera, the 3D locations of the observers and the 3D location of the VIOA. The only exception to this is when the VIOA is the camera, in this case the structure is recovered up to an unknown scale factor.

Finally, the suggested method demonstrates how and at what extent head pose estimation algorithms can be used in a high level application.

The following section reviews related work. Next, Section 2 describes the proposed method, including its main ideas. The associated geometry is presented in Section 3. Section 4 addresses the challenges in real-world scenarios. Simulation and real-data experiments are presented in Section 5, including the outcome of the method on many images obtained from the Internet. Finally, the last section, 6, is devoted to conclusions.

1.1 Related Work

Related work on the recovery of VFOA will be presented first. Next, related work on head pose estimation and face detection will be addressed.

Detection of VFOA of a Single Observer To the best of our knowledge there is no work that addresses the VIOA of a group as a whole in the general case. However, for a single observer, the recovery of the VFOA using a single camera was addressed when scene structure and camera calibration are known [6, 7, 4, 8, 5]. An attention monitoring system for air-traffic controllers [6], is used to build the attention distribution of attended objects. They use a single camera to track head pose and a complete environment model including 3D structure and objects regions. Meeting rooms are addressed by [7, 4, 5, 9]. The structure of the meeting room is known in detail, including the sitting positions of the observers, the set of possible VFOAs and the location of the calibrated cameras. A training set is used to associate a set of specific head-poses with the set of image-VFOAs (the 2D projection on the image plane of a VFOA in 3D space). Moving observers that are attending to a single VFOA are addressed by [8]. The camera and the single VFOA are fixed and predefined, while detecting events of attending to the VFOA and counting how many observers are attending to the VFOA. Meeting context is used by [9] to cope with the ambiguity in a meeting room with moving observers and several VFOAs.

Face Detection and Head Pose Estimation Even the most accurate gaze estimation methods do not find a single point, but return a cone (centered around a ray) in the 3 dimensional world. Thus, additional cues are inherently required for the recovery of the VFOA or the VIOA. Nevertheless, an accurate estimation of gaze direction of a person may significantly improve the accuracy of the recovery. A basic finding states that the VFOA can be reasonably approximated by head-pose in many cases [10]. 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 accuracy of current solutions is lacking. Specifically for a head image in moderate resolution the pitch angle is usually estimated with large errors. A recent survey [11], covers about 100 methods for pose estimation, including [12] that is used in our implementation. This method is mentioned in the survey as one of the most accurate methods. Still, head pose estimation remains an open research problem, e.g. [13].

The detection of faces in an image is the first step in VFOA recovery. An efficient method that is based on a boosting algorithm is the popular method by Viola and Jones [14]. There are several extensions of it including [15, 16] and a popular implementation [17].

2 The Suggested Method

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

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

An hypothesis of a mutual awareness event includes a group of observers and their VIOA. Obtaining a small set of MA hypotheses is essential for an efficient solution. The geometry of the VIOA is used to constrain the hypothesis search. When there is no noise in the measurements 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.

The problem's setup consists of n observers, the intersection point, Q, the camera (including its calibration matrix K), the head pose algorithm and finally the measurements vector $U = \left(U_1\, U_2 \ldots U_n\right)^T$, where $U_i = \left(p_i\, r_i\, \alpha_i\, \beta_i\, \gamma_i\right)^T$. This parameters are the head position in the image, its size and its pose angles, which are the results of the detection algorithms.

The following algorithm handles the ideal situation when all observers have the same VFOA, K is known and the attributes for all observers are known and accurate:

---

procedure IS-MUTUAL-AWARENESS($image/s, K$)

$U \leftarrow$ DETECT-OBSERVERS($image/s$)

$Q \leftarrow$ ESTIMATE-VIOA(U, K)

output (Q, U)

First it detects the observers and their associated measurements and then applies the geometric constraint and finds the VIOA. The estimation of the VIOA is explained in Section 3.

In real life, the VIOA is shared by a subset of the observers in the images having $U' \subseteq U$ as their set of measurements. In addition, the measurements of an observer, $U_i \in U'$, are noisy and K might be unknown. The full algorithm estimates an unknown K while obtaining an initial U' and Q. Then, it refines K and Q to find the largest U' with the highest probability.

procedure IS-MUTUAL-AWARENESS($image/s, K_0, \sigma, T_{\chi^2}$)
$U \leftarrow$ DETECT-OBSERVERS($image/s$)
$(K, Q) \leftarrow$ ESTIMATE-CALIB-MAT($U, K_0, \sigma, T_{\chi^2}$)
$(K, Q, U', score) \leftarrow$ FINE-TUNE($U, K, Q, \sigma, T_{\chi^2}$)
output $(K, Q, U', score)$

The details of this algorithm are given in Section 4.

3 The Geometry of Mutual Awareness

Estimating the position of the VIOA is inherently coupled with the detection of an MA event as it is its main attribute. The analysis of the geometry of an MA event is a major part in this work. It enables the estimation of 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 $(\alpha_i, \beta_i, \gamma_i)$ that respectively are the yaw, pitch and roll angles. The angles are given with respect to a head facing the camera (frontal viewed head). In this situation $\alpha_i = \beta_i = \gamma_i = 0$. The attention ray of a head is a ray that is perpendicular to the face plane and is directed forward. Note that the roll rotation does not change the direction of the attention ray. For a frontal 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 toward the camera (frontal viewed) is rotated by the rotation matrix to the direction of the VIOA, Q (current pose). This can be written as

$$(Q - P_i) \times (R_i \cdot P_i) = 0. \tag{1}$$

3.1 VIOA Estimation via Geometric Constraints

For every head two planes can be considered. The first plane, Π_i, is spanned by P_i and $R_i \cdot P_i$. This plane contains the origin. The second plane, Π_i^β, is the i'th local YZ plane. Both planes contain the intersection point, Q.

The normals to the planes Π_i and Π_i^β will be denoted N_i and M_i respectively. The two plane types complement each other, when all the planes of one type coincide, the planes of the other type do not. Plane Π_i constrains the intersection point by $N_i^T \cdot Q = 0$. Plane Π_i^β constrains the intersection point by $M_i^T \cdot (Q - P_i) = 0$. Arranging the normals in a matrix yields for the first type normals

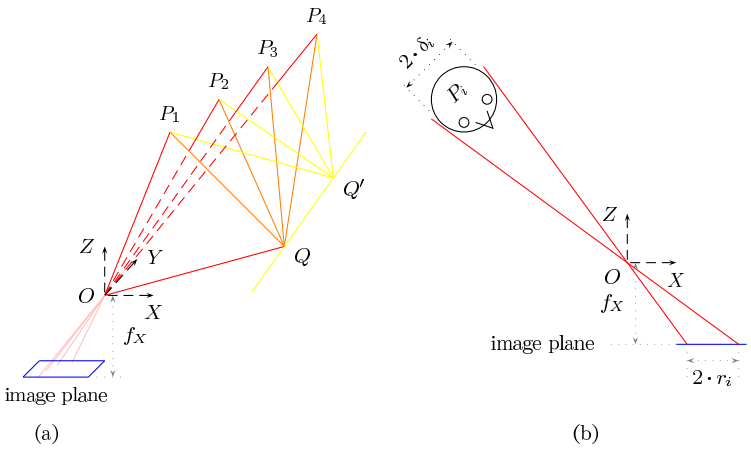


Fig. 1. (a) Four observers $\{P_i\}_{i\in\{1,2,3,4\}}$, their projection on the image plane and their associated planes. The attention rays of the four observers intersect at Q. The first type planes, Π_i, contain the red triangles $\triangle_{P_i,O,Q}$. The second type planes, Π_i^β, contain the yellow triangles $\triangle_{P_i,Q',Q}$. The VIOA, Q, sits on the intersection line of the first type planes. (b) Scene top view, the XZ-plane. Depth estimation of the i'th observer is computed from the diameter of head's projection.

the matrix $N = \begin{pmatrix} N_1 & N_2 & \dots & N_n \end{pmatrix}$ and similarly for the second type normals the matrix M. The above constraints are written as

$$Q = \begin{pmatrix} M^T \\ N' \end{pmatrix}^\dagger \cdot \begin{pmatrix} \mathrm{mp} \\ 0 \end{pmatrix} , \tag{2}$$

where $\mathrm{mp}_i = M_i^T \cdot P_i$, $N_i = \frac{P_i \times (R_i \cdot P_i)}{||P_i \times (R_i \cdot P_i)||_2}$ and $M_i = R_i \cdot \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}^T$. It will be shown in Section 3.1 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 relative size of a head corresponds to its relative distance from the camera.

Note that the normals N_i determines 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, i.e. $P_i = Z_i \cdot K^{-1} \cdot p_i$, and then writing the above as

$$N_i = \frac{(K^{-1} \cdot p_i \cdot Z_i) \times (R_i \cdot K^{-1} \cdot p_i \cdot Z_i)}{||(K^{-1} \cdot p_i \cdot Z_i) \times (R_i \cdot K^{-1} \cdot p_i \cdot Z_i)||_2} = \frac{(K^{-1} \cdot p_i) \times (R_i \cdot K^{-1} \cdot p_i)}{||(K^{-1} \cdot p_i) \times (R_i \cdot K^{-1} \cdot p_i)||_2}.$$

Thus, another possible method to solve for Q is to find the direction of Q by applying SVD on the homogeneous equations of the first constraint $N_i^T \cdot Q = 0$. This option is more stable as the estimated direction of Q is not affected by the errors in depth (face radius) measurements. The magnitude of Q can then be determined by using $(Q - P_i) \times (R_i \cdot P_i) = 0$. The last equation yields a set of linear equations in the magnitude of Q. This step requires the depth information.

Depth Estimation The suggested method requires a depth estimation for each head, which is computed from the 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. Let δ_i be the 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 there is an average face size, δ, and that $\delta_i \approx \delta$. The projection of the head and the camera on the XZ plane is illustrated in Figure 1(b). From this figure it can easily be verified that $Z_i = \frac{\delta_i}{r_i} \cdot \sqrt{f_X^2 + (x_i - u_0)^2}$. Similarly, the projection on the YZ plane yields $Z_i = \frac{\delta_i}{r_i} \cdot \sqrt{f_Y^2 + (y_i - v_0)^2}$. Since the data is noisy and K is only approximated, the depth is estimated as the average of the two expressions above.

Planar Approximation The depth can be corrected when heads lie approximately on a plane in $\mathbb{R}^3$. Assuming the plane is represented by a 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 $P = Z \cdot K^{-1} \cdot p$, it 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:

$$(K^{-1} \cdot \mathcal{P})^T \cdot L = \begin{pmatrix} \frac{1}{Z_1}\ \frac{1}{Z_2} \cdots \frac{1}{Z_n} \end{pmatrix}. \tag{3}$$

By applying the pseudo-inverse, $\left((K^{-1} \cdot \mathcal{P})^T\right)^\dagger$, on (3) 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.

Estimating the Yaw and Pitch from Q Given an intersection point Q and an estimated head position, P_i, 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, $\begin{pmatrix} \bar{X}_i^O\ \bar{Y}_i^O\ \bar{Z}_i^O \end{pmatrix}^T$. They are computed as,

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

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

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

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

$$\hat{\alpha}_i = \tan^{-1}\left(\frac{\left\langle \bar{Z}_i^Q, \bar{X}_i^O \right\rangle}{\left\langle \bar{Z}_i^Q, \bar{Z}_i^O \right\rangle}\right) \quad \text{and} \quad \hat{\beta}_i = \tan^{-1}\left(\frac{\left\langle \bar{Z}_i^Q, \bar{Y}_i^O \right\rangle}{\left|\left|\left(\left\langle \bar{Z}_i^Q, \bar{Z}_i^O \right\rangle \left\langle \bar{Z}_i^Q, \bar{X}_i^O \right\rangle\right)\right|\right|_2}\right).$$

A Special Case: "Say Cheese" — the Camera is the VIOA When all the observers look directly at the camera, e.g. when taking a group photo, 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, (1) becomes $\mathbf{0} = P_i \times P_i$, for $1 \leq i \leq n$. Thus, the geometric constraint on attention rays can not be used to recover K. In this case our method will detect the MA event but the 3D reconstruction will be known up to K.

4 Mutual Awareness in the Wild

The set of measurements, U, that was obtained by face detection and head-pose estimation algorithms might be noisy. In general, nothing is assumed on how those algorithms work internally. Thus, a reasonable approximation for the noise in measurements is that they are i.i.d normal variables. Therefore, the probability for an MA event given the measurements is proportional to the probability of the noise in the measurements given that MA event (Bayes rule).

Following from the above assumption: $U_i \sim N(\mu_i, \Sigma_i)$ and μ_i is estimated by $\hat{U}_i = \left(\hat{p}_i\ \hat{r}_i\ \hat{\alpha}_i\ \hat{\beta}_i\ \hat{\gamma}_i\right)^T$ that satisfies constraint (1). The noise in p_i is assumed to be very small compared to the other measurements and therefore is ignored. The roll angle, γ_i, is eliminated as it does not affect the direction of the attention ray. Thus, given a VIOA point, Q, and a camera's calibration matrix, K, the probability that the i'th observer is indeed looking at Q satisfies:

$$\log(\Pr\left(U_i \in \text{MA}|Q,K\right)) \propto \Big(\frac{(\alpha_i - \hat{\alpha}_i)^2}{\sigma_{\hat{\alpha}_i}^2} + \frac{(\beta_i - \hat{\beta}_i)^2}{\sigma_{\hat{\beta}_i}^2} + \frac{(r_i - \hat{r}_i)^2}{\sigma_{\hat{r}_i}^2}\Big). \tag{4}$$

The standard deviations, $\sigma = \{\sigma_{\alpha_i}, \sigma_{\beta_i}, \sigma_{r_i}\}_{i=1}^n$, can be estimated empirically using the selected algorithms for face detection and pose estimation. The above sum has a χ^2 distribution. The maximum likelihood $\hat{U}_i$ can be estimated given Q and K. When $\hat{r}_i$ is given, e.g. in the case of planer approximation, there is a closed form solution for $\hat{\alpha}_i$ and $\hat{\beta}_i$ (as shown above). Thus, in the general case $\hat{U}_i$ can be obtained using a single parameter optimization on $\hat{r}_i$.

In the general case only a subset of the observers participate in the MA event. The probability of the subset, U', of the observers is a product of the observers' probabilities. Taking its log results in a sum of χ^2 random variables which is also has χ^2 distribution. Thus, in order to accept or reject an MA hypothesis the χ^2-test is used with respect to U'. The MA event is likely occurring if the χ^2 probability is above a desired confidence level T_{χ^2}. First the entire set U is tested and if the test fails then a subset without the least probable observer is re-tested. The elimination of observers stops when the confidence level has been reached. This procedure is denoted by χ^2-FILTER$(U, \hat{U}_Q, \sigma, T_{\chi^2})$ and its result can be used as a score in the process of estimating the MA event's parameters. The MA event, in the special case when the camera is the VIOA, is detected or rejected by simply calling χ^2-FILTER$(U, \hat{U}_{\mathbf{0}}, \sigma, T_{\chi^2})$.

A general maximization algorithm can be used for this purpose. It is expected to converge to the global maximum if its initial guess is close enough. The maximization is done with respect to a score which is taken to be the number of

observers in the subset and their χ^2 probability, i.e.

$$\text{score}(U', Q, K) = |U'| + \Pr(U'|Q, K). \quad (5)$$

If the internal calibration of the camera, K, is unknown it is estimated using a general maximization algorithm as above. In each iteration a value of K is being tested and the best Q needs to be found. To find such Q the RANSAC algorithm is used to obtain many Q's. Each such Q is obtained geometrically from a small random subset, U', of the observers by the following procedure:

procedure GEOMETRIC-MODELING$(U, U', K, \sigma, T_{\chi^2})$
  $score' \leftarrow 0$
  repeat
    $score = score'$
    $Q \leftarrow$ ESTIMATE-VIOA(U', K)
    $\hat{U}_Q \leftarrow$ ENHANCE-MEASUREMENTS(U, K, Q, σ)
    $(U'', prob) \leftarrow \chi^2$-FILTER$(U, \hat{U}_Q, \sigma, T_{\chi^2})$
    $score' \leftarrow$ NUM-OF-OBSERVERS$(U'') + prob$
    if $score' > score$
      then $U' = U''$
  until $score' <= score$
  return $(U', Q, score)$

Starting from U', this procedure tries to find Q and a subset that have the largest possible score. It iterates to find the largest subset U' that is most likely. It finds the VIOA (Q) of the subset U' and then for every observer it maximizes the likelihood using a single parameter maximization over the depth which results with optimal depth and head pose according to the estimated VIOA, $\hat{U}_Q$. Then it checks all the members of $\hat{U}_Q$ to find the largest subset $U'' \subseteq U$ agreeing with Q and which has probability above T_{χ^2}.

When a reasonable (K, Q) pair has been obtained, e.g. detection of an MA event of 3 observers, it is taken as the initial guess in a fine tuning algorithm. The fine tuning here is another general maximization algorithm that searches for the (K, Q) pair with the largest score.

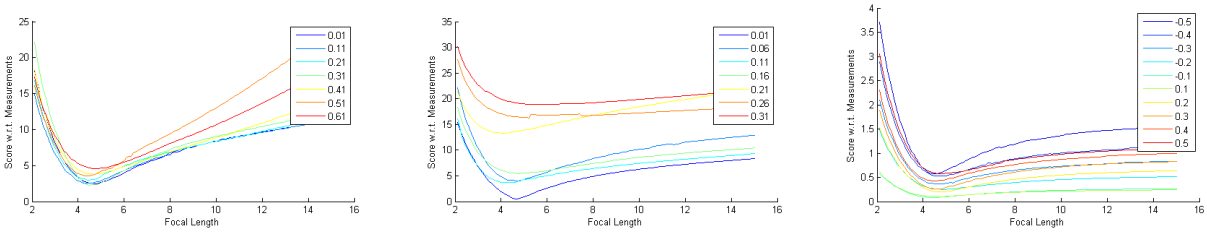
5 Experiments

To validate our method three types of experiments were conducted. Simulations were performed to understand the behavior of the minimization w.r.t. the score (5). In the experiments the method was tested on real images obtained from the Internet of scenes with VIOA. Common assumptions on the perspective model were made, i.e. no skew, the principal point is in the center of the image and that the ratio between f_x and f_y is known (in our case 1).

5.1 Simulations

The method strongly depends on the selected face detection and pose estimation algorithms. Each instance of those algorithms hes its own noise to the measurements. Following is a simulation that demonstrates the expected performance of

the method regardless of the selection of specific algorithms. In Figure 2(a)-(b) the method was applied in common situations with varying levels of noise in the measurements. Each curve represents the score as a function of the focal length, where the true focal length is at 4.6. As the noise increases, the ability of the algorithm to converge to the correct result decreases. When the VIOA is the camera, the method can not recover K. Figure 2(c) demonstrates the performance as a function of the X coordinate of the VIOA. As X moves further from zero the ability of the algorithm to find the right focal length improves.



(a) (b) (c)

Fig. 2. Simulation of a search over K (f) and its associated score (eq-5) for: (a) different levels of noise in the angle's measurements; (b) different levels of noise in the depth's measurements; (c) different locations of the VIOA near the camera.

5.2 Real-Data

The method was implemented in C++ using the OpenCV library and MATLAB. First the head pose algorithm of [12] was applied to the image. In addition the OpenCV implementation of the Viola-Jones algorithm was applied to the image to refine the estimate of the size of each head. It was noticed that there are false detections of heads or gross errors in head pose estimation. To address this the algorithm in Section 4 was used, where the size of a subset drawn by the RANSAC algorithm is 3.

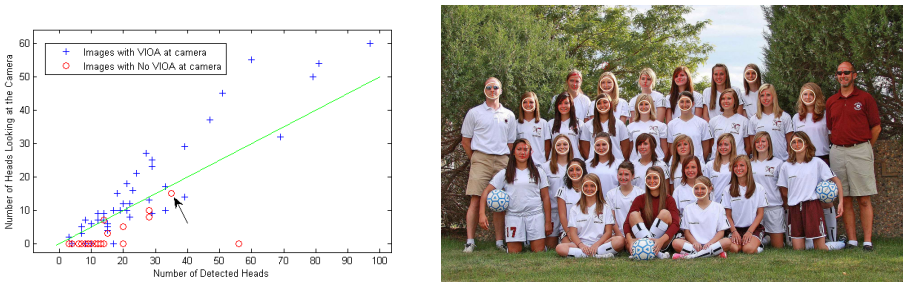
Pose Estimation and Face Detection Algorithm The face detector and pose estimator that we have used[3] uses a convolutional network [12]. This method estimates only two angles: the yaw angle, α, and a second angle, ρ. The second angle corresponds to the roll angle while in frontal view but corresponds to the pitch angle for profile view. Specifically, the pitch is only partially estimated by $\beta = \sin(\alpha) \cdot \rho + (1 - \sin(\alpha)) \cdot \beta_0$ and the roll is only partially estimated by $\gamma = \cos(\alpha) \cdot \rho + (1 - \cos(\alpha)) \cdot \gamma_0$, where β_0 and γ_0 are unknown. Thus, when the yaw angle is $\pm\pi/2$ the roll angle is completely unknown and the pitch angle is estimated with no corrections. When the yaw angle is 0 the pitch angle is completely unknown and the roll angle is estimated with no corrections.

[3] with the kind permission of the authors.

Therefore, the roll and pitch angles are estimated as closer to zero than they actually are. This causes their standard deviation to become larger. In addition, the method is more accurate (has a smaller variance) for small yaw angles and noisier for large ones. Those characteristics of the standard deviations are modeled by the algorithm.

As a consequence of the large uncertainty in the pitch angle the estimated vertical position of Q, Q_Y, is less accurate.

Detection of an MA Event The detection of an MA event is a major capability of the method. In the following experiment a simple MA event was tested, the one where the VIOA is the camera. When the VIOA is the camera the yaw angle is expected to be zero. The squared deviation of the measured yaw from zero is the score of a face. The total score of an image was obtained by the sum of the scores of all detected faces divided by the variance ($\sigma = 9°$). The maximal subset of faces from each image was selected such that the χ^2-test of the their score was above 0.95. In the experiment, a set of 68 images were collected from the Internet (included in the supplementary material [18]) and manually classified w.r.t this MA event. As a result, a subset of 43 the images was classified as positive. The two subsets are shown in Figure 3(a). Each image is represented by the number of detected heads and the number of heads that were classified as looking towards the camera. As can be seen, there is a good separation between the two classes. A negative example is shown in Figure 3(b) in which the location of the VIOA is close to the camera. On such an example an automatic separation is expected to be inconclusive.



(a) (b)

Fig. 3. (a) Detection's quality of the MA event when the VIOA is the camera. (b) An extreme example of an image with a VIOA close to the camera (pointed by an arrow)

The General Results The method was tested on images obtained from the Internet of scenes with VIOA. The results are shown in Figures 4, 5 and in the supplementary material [18]. In each image, the face of each detected person is

circled with a red circle. The radius of the circle indicates the size of the face which is proportional to its distance from the camera. Within the circle there is a triangle that indicates the yaw, pitch and roll angles. A face is circled with an additional white circle if it belongs to the group that is attending to the VIOA. In each image a white box has been inserted to indicate the structure of the reconstructed scene. For clarity reasons, for few images only the projection on the XZ plane is shown. In each box an observer is represented by a black circle, the blue lines are the attention rays, the VIOA is the intersection of the red lines, which connect the VIOA with the nearest point on each attention ray. The red circle at the XZ origin (0,0) is the position of the camera. Since it is quite common that observers are captured in a vertical body posture the rotations w.r.t. the yaw and pitch angles are aligned with the x and y axes of the camera coordinate system. Note that because the pitch angle is not estimated well by the head pose estimation algorithm that was used, the detected VIOA has an inaccurate Y position.

In Figure 4 a group photo was taken by two cameras. The observers were looking towards one camera, while the algorithm was applied to the other one. The algorithm was applied to this image twice: the first time while assuming that the heads are on a plane and the second time when such an assumption is not made. The image's projection of a grid on the detected plane is drawn. The left box is the result when the planar assumption is not made, while the right box is the result when the planar assumption is made. By comparing the two boxes it is clear 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 which is not seen in the image. The algorithm estimated the FOV of the camera as 36°.

Four other setups can be seen in Figure 5 including VIOA within the image, temporal aggregation of attention in a video and the detection of an MA event without the calibration matrix in the case the camera is the VIOA.

6 Discussion

In this work we have defined the problem of mutual awareness (MA). We have demonstrated experimentally that it can be solved using off the shelf software. An important byproduct of the algorithm is the recovery of sparse scene structure and the internal calibration matrix of the camera. The algorithm was tested on images from the Internet of scenes exhibiting VIOA. This method can be applied in a general environment. It therefore frees its users from defining a strict setup when observing human behavior.

In addition, the geometric constraint, which is presented in this work (VIOA), can significantly reduce the size of the training set, used by previous works [4, 5, 8, 9]. The training data-set, which enables the detection of the most likely VFOA of a new observer, must densely cover the space of position pairs, i.e. observer position and VFOA position. However, using the geometric constraint the training data-set should cover densely only the VFOA positions. By including, for every

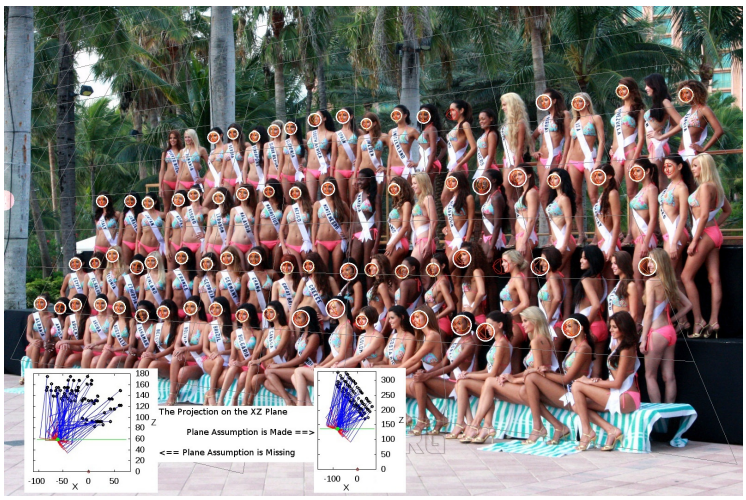


Fig. 4. A large group with a VIOA outside the field of view of the camera (should be viewed in color). See text for details.

VFOA, only few examples in sparse positions the VFOA of a novel observer is the VIOA of those training examples and the novel observer.

References

1. Baron-Cohen, S.: 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. Emery, N.: The eyes have it: the neuroethology, function and evolution of social gaze. Neuroscience and Biobehavioral Reviews **24** (2000) 581–604
3. Frank, M., Vul, E., Johnson, S.: Development of infants' attention to faces during the first year. Cognition **110** (2009) 160–170
4. Ba, S.O., Odobez, J.M.: A study on visual focus of attention recognition from head pose in a meeting room. MLMI (2006) 75–87
5. Ba, S.O., Odobez, J.M.: Multi-party focus of attention recognition in meetings from head pose and multimodal contextual cues. ICASSP (2008)
6. Brolly, X.L.C., Stratelos, C., Mulligan, J.B.: Model-based head pose estimation for air-traffic controllers. ICIP **2** (2003) 113–116
7. Otsuka, K., Takemae, Y., Yamato, J., Murase, H.: Probabilistic inference of multiparty-conversation structure based on markov-switching models of gaze patterns, head directions, and utterances. ICMI (2005)
8. Smith, K., Ba, S.O., Odobez, J.M., Gatica-Perez, D.: Tracking the visual focus of attention for a varying number of wandering people. PAMI **30** (2008) 1212–1229
9. Ba, S.O., Odobez, J.M.: Multi-person visual focus of attention from head pose and meeting contextual cues. Accepted for publication in PAMI (2010)
10. Stiefelhagen, R., Finke, M., Yang, J., Waibel, A.: From gaze to focus of attention. VIIS (1999) 761–768

(a) Demo (b) Fixating Person

(c) "Say Cheese" (d) Who Will Catch the Bouquet?

Fig. 5. Four examples: (a) the VIOA is the computer screen. (b) 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. (c) the VIOA is the camera, the MA event is detected but not its attributes. (d) the VIOA is the bouquet in the air its position was estimated to be between the observers and the bride. Since the bride is not looking at the bouquet she is not circled with a white circle.

11. Murphy-Chutorian, E., Trivedi, M.: Head pose estimation in computer vision: A survey. PAMI **31** (2009) 607–626
12. Osadchy, M., LeCun, Y., Miller, M.: Synergistic face detection and pose estimation with energy-based models. JMLR **8** (2007) 1197–1215
13. Kaminski, J.Y., Knaan, D., Shavit, A.: Single image face orientation and gaze detection. MVA **21** (2009) 85–98
14. Viola, P., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. CVPR **1** (2001) 511–518
15. Lienhart, R., Maydt, J.: An extended set of haar-like features for rapid object detection. ICIP **1** (2002) 900–903
16. Wang, Y., Liu, Y., Tao, L., Xu, G.: Real-time multi-view face detection and pose estimation in video stream. In: ICPR. (2006) 4:357–360
17. OpenCV: `http://www.sourceforge.net/projects/opencvlibrary` (2009)
18. We: `http://mis.haifa.ac.il/~mis/userfiles/file/MutualAwarenessSupp.pdf` (2010)