

Collaborative Face Orientation Detection in Wireless Image Sensor Networks

Chung-Ching Chang
Wireless Sensor Networks Lab
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
bobbyc@stanford.edu

Hamid Aghajan
Wireless Sensor Networks Lab
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
aghajan@stanford.edu

Abstract

Most face recognition and tracking techniques employed in surveillance and human-computer interaction (HCI) systems rely on the assumption of a frontal view of the human face. In alternative approaches, knowledge of the orientation angle of the face in captured images can improve the performance of techniques based on non-frontal face views. In this paper, we propose a collaborative technique for face analysis in smart camera networks with a dual objective of detecting the camera view closest to a frontal view of the subject, and estimating the face orientation angles in all the camera views based on additional fusion of local angle estimates. Soft information indicating the probabilities of face and eye candidates in each image is exchanged between the cameras, and epipolar geometry mapping is employed to assess correspondence between candidates in different views. Once the camera with the closest view to the frontal face view is identified, further exchange of the face orientation angles estimated by all cameras allows for a collaborative refinement of the estimates according to their associated confidence levels. The proposed collaborative detection and estimation schemes employ low-complexity algorithms and do not require image transfer between the cameras. Hence, these schemes are applicable to networks of image sensors with in-node processing and narrowband wireless communication.

Keywords

Face analysis, Epipolar geometry, Camera networks

1 Introduction

Advent of image sensor and embedded processing technologies has enabled novel approaches to the design of security and surveillance networks while facilitating creation of new application classes such as smart environments. Many applications of distributed image sensors that involve monitoring of humans make use of visual information obtained by the cameras from the face of the subject. When multiple image sensors view a freely moving (i.e. non-cooperative) person, only a few snapshots captured during the observation period may provide an adequate view of the person's face for further analysis. Detection and recording of those frames would hence be the key to enabling effective facial analysis techniques while allowing the network to operate under an efficient data management regime. In surveillance applications, capturing the frontal face view of the intruder is often

of paramount importance. While the current methods based on off-line sifting through the multitude of saved frames can yield snapshots of the intruder's face captured by different cameras, real-time selection and distribution of a facial view enables other applications such as tracking to be activated by the network.

Most face recognition and analysis algorithms such as principle component analysis (PCA) [14], linear discriminant analysis (LDA) [9], and hidden markov model (HMM) techniques [7], require face images with approximately frontal views to operate efficiently. In other words, in order to be robust, the PCA and LDA techniques require a large number of training samples in different face orientation angles. Other approaches such as robust recognition by stereo vision [15] require large amount of computation in 3D reconstruction of the face. The 3D morphable model algorithm [1] highly reduces the computational complexity in reconstructing a 3D model; however, it requires a frontal view image of the face in the training stage.

Many existing face orientation detection algorithms are based on analyzing a frame captured by a single camera in which both eyes and/or the mouth are observable [4], [6]. In other words, they assume the image is taken from a generally frontal view. Many of these techniques require a first step of eye and mouth detection, and then calculate the geometric relationship between these features such as the deviation of the eye and mouth centroids from the center of the face ellipse. On the other hand, symmetry also provides much information about the orientation of a face without feature detection. For example, discrete Fourier transform (DFT) techniques applied to edge-detected image and methods based on the principal and secondary texture orientations in the frequency domain [13] have been used in the past. The work presented by Wiles *et al.* [16] uses corners and hyper-patches to detect the pose, based on the assumption that typically human head has 20 stable corners, of which 15 may lie on roughly frontal plane of the face. Detection of corners and making correspondences between them in the temporal domain add overhead computational load to this scheme. In addition, all these techniques fail when the image is not in frontal view.

The approach proposed in this paper is based on a networked camera setting, in which the nodes collaborate on detecting the best frontal view snapshot by exchanging soft information each one extracts from its view to the person's face. Distributed image sensing allows accessing different

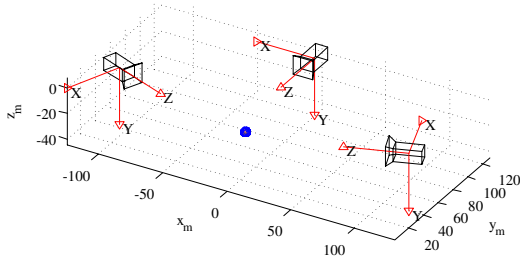


Figure 1. Multi-camera network geometry

views of the face of the person under observation. The network can in fact exploit both the spatial as well as temporal spans in observing the person to opportunistically acquire and save the snapshots that are close to the frontal face views. Besides providing surveillance and tracking applications with a selection of frontal views of the face, the proposed approach allows other face analysis applications to employ low-complexity facial analysis schemes that assume a frontal face view, hence relaxing the demands for adoption of complex invariant features. In addition to solving the detection problem of determining the best snapshot of the frontal face view, the proposed approach also formulates an estimation problem in which the orientation angles of all the face images are found by further exchange of local estimates between the camera nodes.

2 Local Processing Methods

The geometry of the network consisting of distributed cameras for a typical smart environment or surveillance application is shown in Fig. 1. Local processing algorithms performed in each camera node consist of methods to detect a set of face and eye candidates in each frame. These techniques are developed to be of low computational complexity nature, allowing them to be adopted for in-node processing implementation. Due to their rather simple design, each individual scheme running at each camera node may not produce a unique candidate for the face and eye features sought in the body mask. However, the fact that each camera employs multiple methods to detect the feature candidates, and that the camera nodes exchange their soft information with each other, allows the network to enhance its detection accuracy as well as confidence level and produce accurate results describing the orientation of each facial view. The architecture of the developed algorithms for in-node local processing is shown in Fig. 2. Each camera finds candidates for the eyes and the face with primitive image processing. This includes a simple segmentation scheme, a skin color detection method with geometric constraints, and using a function of chrominance and weighted positioning on the face candidate to detect eye candidates. The probability of those candidates being correct is measured by a goodness figure. Location and probability of those candidates are exchanged with other network nodes. With the additional information received from other nodes, each camera can produce a better estimation of the face orientation in its captured view.

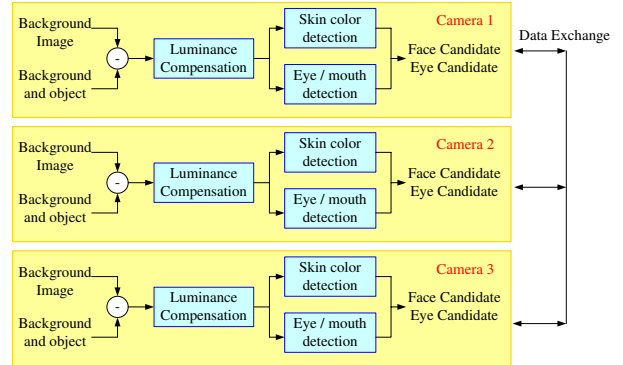


Figure 2. In-node processing architecture

2.1 Skin Detection and Face Candidates

Hsu *et al.* [5] proposed that skin tone color can change due to different lighting conditions. Therefore, lighting compensation to a reference white is necessary before skin-color detection. Modeling skin color requires choosing an appropriate color space and identifying a cluster associated with skin color in this space. Majoor *et al.* [11] proposed that in the normalized rg space, a parametric ellipse is used as a region indicating skin color. In our experiments, we adopt the ellipse definition proposed in [10]. After determining skin color region in the image, we apply some primitive constraints such as minimum area and maximum eccentricity for rejecting false alarms. We then calculate the mean and covariance matrix of the points in the enclosure of the face candidate region, and fit an ellipse to the region using the eigenvectors of the covariance matrix.

2.2 Eye Detection and Eye Candidates

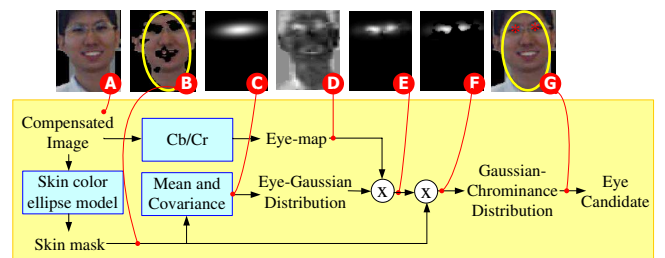


Figure 3. Flowchart and example of the eye detection scheme

Hsu *et al.* [5] also proposed that dilation and erosion by a hemispheric structuring element in luminance can be designed to emphasize brighter and darker pixels in eyes. In addition, an analysis of the chrominance components shows that high Cb and low Cr values are found around the eyes. In our experiments, we notice that the value of Cb/Cr is actually higher in the eye regions than in the other parts of the face; however, it is not discriminative enough. Combined with the prior information of the face ellipse, we can emphasize the region that most probably includes the eyes by a 2 dimensional Gaussian distribution with its center and covariance matrix adaptive to the major and minor axes and the orientation of the face ellipse. After multiplying the Gaus-

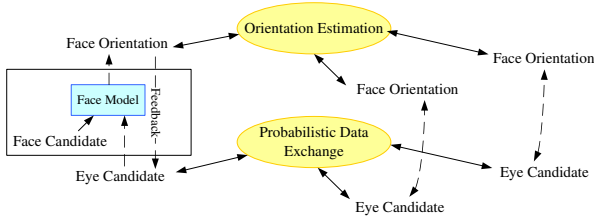


Figure 4. Two layers of data exchange

sian function with the complement of the skin mask, indicating that only non-skin color regions may include eyes, we pick pairs of regions that are orthogonal to the orientation of the ellipse as eye candidates. The procedure is illustrated in Fig. 1. The benefit of using the Gaussian-Chrominance method is that it can successfully detect the eyes located on the edge of the skin region.

3 Data Exchange Mechanisms

Collaboration between cameras is achieved by data exchange. Due to employing low-complexity schemes to detect the face and eye candidates, there could be deviations in their positions from the true locations, and outliers may also be detected as such candidates. By sharing information with other cameras, each camera can obtain additional information from other perspectives and make a better decision. The data exchange mechanism is designed as a two-layer approach shown in Fig. 4. In the lower layer, locations of eye candidates are exchanged between the nodes using a probabilistic transformation described in section 3.3. The face and eye candidates are input to a face model depicted in Fig. 5, and a face orientation estimate is produced. This estimate is given to the upper layer, which determines the camera with the closest view to the frontal view. A goodness figure described in section 3.4 is provided as feedback to the lower layer to refine the eye candidate position estimates. All calculations are distributed, i.e., there is no central server involved. Correspondence between camera views plays an important role in the data exchange mechanism, and the key to the correspondence is camera localization and epipolar geometry.

3.1 Camera Localization

The purpose of localizing the cameras is to provide a basis for relating and combining detection results for the face and eye candidates obtained by the different camera nodes. Relative location and orientation angles of the cameras can be detected in prior by a vision-based localization technique such as the method proposed by Lee and Aghajan [8]. Knowing this information, we can calculate the fundamental matrix between each pair of the cameras. Alternatively, Hartley *et al.* [3] proposed several ways to calculate fundamental matrix by manually finding corresponding points in two camera frames.

3.2 Epipolar Geometry

Correspondence between images is found via epipolar geometry [12]. Consider a feature point in a camera's field of view (FOV). The point corresponds to a ray in the 3D space, and the actual observed feature point could be anywhere on that ray. Projection of the ray from the 3D space onto another camera's FOV is a line. The relation between a point

in one FOV and the corresponding line in the other FOV is fully determined by the relative positions (x, y, z) and orientations $(\theta_x, \theta_y, \theta_z)$ of the cameras, which can be modeled as a 3-by-3 fundamental matrix F . A point in one frame and any point on the corresponding epipolar line in the other frame have the relation:

$$x_1^T F x_2 = 0 \quad (1)$$

where x_1 and x_2 are vectors containing the homogeneous coordinates.

3.3 Probabilistic Data Exchange Model

The local feature detector, image processing in each camera, makes two types of errors - classification errors and measurement errors. Measurement errors generally follow a normal distribution. Classification errors, however, are gross errors, having a significantly larger effect than measurement errors, and do not average out. RANSAC [2] methods are often used to select feature points in this case. The method starts from a small subset of candidates and finds a subset with a model that fits to the biggest set. We can combine this idea with a probability model as follow:

$$\begin{bmatrix} P_{x_1} \\ P_{x_2} \\ \vdots \\ P_{x_n} \end{bmatrix} = \begin{bmatrix} P(x_1|x_{1,true}) & P(x_1|x_{2,true}) & \dots & P(x_1|x_{n,true}) \\ P(x_2|x_{1,true}) & P(x_2|x_{2,true}) & \dots & P(x_2|x_{n,true}) \\ \vdots & \vdots & \ddots & \vdots \\ P(x_n|x_{1,true}) & P(x_n|x_{2,true}) & \dots & P(x_n|x_{n,true}) \end{bmatrix} \begin{bmatrix} P_{x_{1,true}} \\ P_{x_{2,true}} \\ \vdots \\ P_{x_{n,true}} \end{bmatrix} \quad (2)$$

where $P(x_n)$ represents the probability of measurement being in location x_n , and $P(x_n|x_{n,true})$ represents the probability of x_n being the true location. $P(x_m|x_{n,true})$ is the probability of measurement being in x_m given the true location is in x_n . We assume that $P(x_n)$ in the current step becomes $P(x_{n,true})$ in the following step. We can assume that the distribution is jointly normal caused by measurement errors. We consider each eye candidate as a measurement. If we suppose candidate x_n is the true location, the probability of the measurement at location x_m in the same frame can be calculated directly by probability density function (PDF) of the Gaussian distribution. On the other hand, for a candidate in the other frame, which corresponds to an epipolar line in the current frame, we have to take a line integration along the epipolar line. The calculation can be simplified since the marginal of a Gaussian random vector is a Gaussian random variable.

Suppose there are three camera frames A, B, C, each one with N_A , N_B , N_C eye candidates, respectively. Then $P \in \mathcal{R}^{N_A+N_B+N_C}$, and $P_{true} \in \mathcal{R}^{(N_A+N_B+N_C) \times (N_A+N_B+N_C)}$. The initial condition of P is given by the goodness of the candidate, which is given by the Gaussian-Chrominance model.

For each time instance, we calculate the Markov model matrix A in Eq. 2, and regard the converged value of the vector P as the fusion information. Note that the entries of the Markov matrix A are positive, and the matrix itself is regular (A^k has all positive entries for certain positive k). Therefore, there is an asymptotic rate of convergence, and

$$(\lambda_{pf}^{-1} \cdot A)^t \xrightarrow{t \rightarrow \infty} wv^T, \quad (3)$$

where λ_{pf} is the Perron-Frobenius eigenvalue, and w, v are

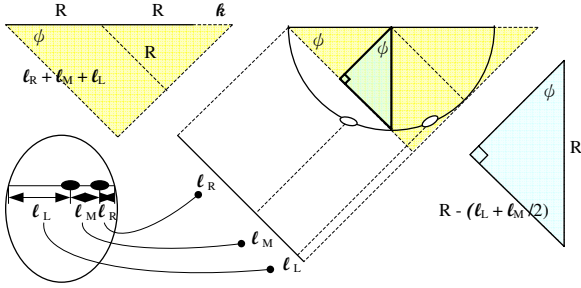


Figure 5. Face orientation geometry

the corresponding left and right eigenvectors. In short,

$$P_{inf} = wv^T P(0), \quad (4)$$

where wv^T represents the information in geometrical relation and $P(0)$ represents the information given by the Gaussian-Chrominance method. Furthermore, we notice that the largest entry in P_{inf} corresponds to the epipolar lines in frames that collect the most probability from the nearby candidates, hence rendering the idea similar to RANSAC.

After finding the most probable eye candidate, we regard the candidate in each frame with the highest probability as the decision of the eye location in that frame. In our experiments, left and right eye candidate information are exchanged separately.

3.4 Face Orientation Detection

Given the right and left eye candidates and the face ellipse, we can estimate the orientation of the face with the following method. Consider the line that passes through both eyes and intersect the face ellipse at two points. This line can be divided into three regions: a segment between the left face edge to the left eye (l_L), a segment between the two eyes (l_M), and a segment between the right eye and the right face edge (l_R). The line is actually corresponding to a surface in the 3D space, which intersects the head in the semi-circular region shown in Fig. 5, in which R represents the radius of the semi-circle, and k is the line segment to the right of the diameter segment. According to the properties of trigonometry and similar triangles, we have:

$$\begin{cases} \frac{l_L + l_M + l_R}{2R + k} = \frac{R}{R + k} = \cos \varphi \\ \sin \varphi = \frac{R - (l_L + l_M/2)}{R} \end{cases} \quad (5)$$

from which we obtain

$$\varphi = 2 \arctan \left(1 - \sqrt{\frac{2(l_L + l_M/2)}{l_L + l_M + l_R}} \right). \quad (6)$$

Similarly, if there is only one eye candidate, considering the line passing through the eye and perpendicular to the major axis of the face ellipse, there are only two segments, l_L and l_R . Then the estimation of face orientation would be approximately the orientation estimated with the method above (with $l_M=0$) plus a constant angle, representing the angle between the eye orientation and the face orientation. Note that we do not take regions covered by hair into consideration. Besides, we use the radius of the face ellipse instead of the

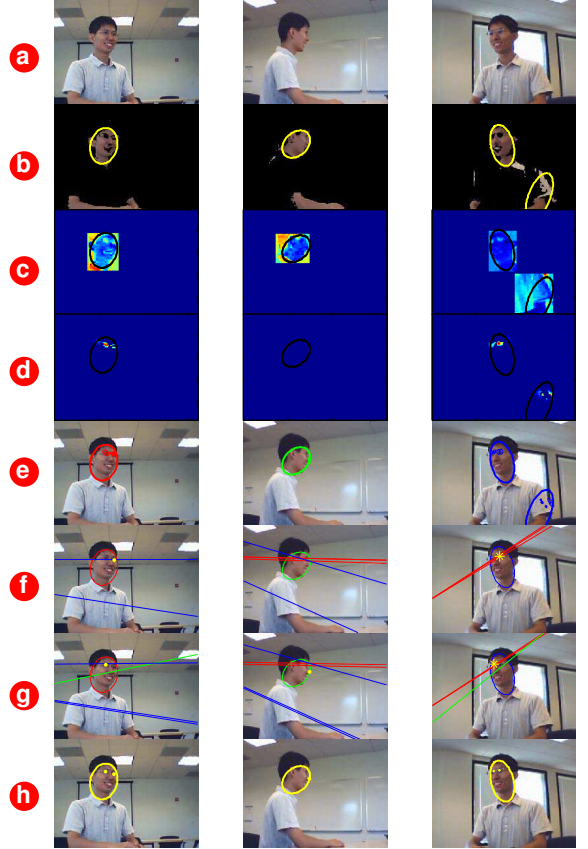


Figure 6. Signal Processing in three cameras (a) original images. Results of (b) skin detection (with face candidates), (c) the Cb/Cr operation, (d) the skin mask complement applied to the Gaussian-Chrominance method, (e) the face/eye candidates, (f) left eye data exchange, (g) right eye data exchange, (h) selected eyes with their correspondence in other frames and final decision.

head ellipse; therefore, the angle estimated would generally appear smaller.

Given the individual face orientation estimates, we calculate the mean of those estimates and find the camera whose angle is closest to this orientation estimation mean. This camera is selected as the one having the best frontal view to the face. The confidence of the selection is measured by the ratio of the second smallest angular separation from the face orientation among other cameras to the angular separation of the selected camera. After determining the camera with the best frontal view, we increase the goodness of the candidates in that frame and reduce the variance of the frame. At this point, we can re-run the probability data exchange and orientation estimation schemes iteratively until the accumulated product of confidence levels goes above a certain threshold. Finally, we calculate the estimated face orientation as the weighted sum of orientation estimates by all the cameras using their goodness measures as the weight.

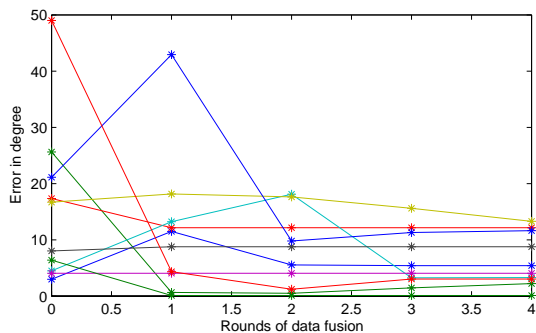


Figure 7. Error reduction behavior versus rounds of data fusion between cameras

4 Experiments

The setting of our experiments is shown in Fig. 1. Three cameras are placed approximately on the same horizon, with one camera placed in frontal direction to the seat, and the two other cameras with about $+70^\circ$ and -70° deviations from the frontal direction, respectively. The experiment is conducted with a person sitting in 13 different sets of directions and poses, such as head tilted, head turned, and hand raised, with each set composed of three images from the cameras.

The results of each step of signal processing in cameras are illustrated in Fig. 6. Each column represents results from one camera. Yellow stars in the second and third rows from the bottom show the final decisions for the right and left eyes, with yellow dots representing the locations of the eyes in the other frames corresponding to the final decision. The experiment shows that face orientation can be detected correctly with high confidence.

The reduction in the face orientation error versus rounds of data exchange between the camera nodes is illustrated in Fig. 7 for 10 experiments. Before data exchange, each frame may choose eye candidates with large deviation from true locations, which in turn causes large errors in the orientation estimate. Data feedback would adjust the goodness of each eye candidate and improve the weighted sum of the face orientation estimate. The errors are largely reduced after several iterations of feedback loop consisting of the proposed probabilistic data exchange and orientation. The error of the estimated face orientation are reduced to ± 15 degrees in angle, and the detection of the camera with the best shot is made with very low probability of error.

5 Conclusions

This work addressed the face orientation estimation problem in a novel way. In addition to local processing in each camera node consisting of primitive image processing operations, collaboration between cameras through a data exchange and feedback mechanism is employed to achieve higher accuracy in orientation estimation. Epipolar geometry is used as a basis for correspondence between camera views, and two layers of data exchange and feedback are applied.

The system works well if both eyes and the face are available in at least one of the camera views, even with

considerable head tilt and head turn. Further work consists of incorporating analysis over both temporal and spatial domains such that smart camera networks can learn over time.

6 References

- [1] X.-M. Bai, B.-C. Yin, Q. Shi, and Y.-F. Sun. Face recognition using extended fisherface with 3d morphable model. In *Proc. of the fourth ICMLC*, volume 7, pages 4481–4486, 2005.
- [2] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395, 1981.
- [3] R. Hartley and N. Dano. Reconstruction from six-point sequences. In *Proc. of IEEE Conf. on CVPR*, 2000.
- [4] H. Hongo, A. Murata, and K. Yamamoto. Consumer products user interface using face and eye orientation. In *Proc. of ISCE*, 1997.
- [5] R. L. Hsu, M. Abdel-Mottaleb, and A. Jain. Face detection in color images. *IEEE Trans. on PAMI*, 24:5:696:706, May 2002.
- [6] A. Kapoor and R. Picard. Real-time, fully automatic upper facial feature tracking. In *Proc. of Fifth IEEE Conf. on AFGR*, pages 8–13, 2002.
- [7] D. Kurata, Y. Nankaku, K. Tokuda, T. Kitamura, and Z. Ghahramani. Face recognition based on separable lattice hmms. In *Proc. of ICASSP*, 2006.
- [8] H. Lee and H. Aghajan. Collaborative self localization techniques for wireless image sensor networks. In *Asilomar Conf. on SSC*, 2005.
- [9] C. Liu and H. Wechsler. Enhanced fisher linear discriminant models for face recognition. In *Proc. of ICPR*, volume 2, pages 1368–1372, 1998.
- [10] F. Liu, Q. Liu, and H. Lu. Robust color-based tracking. In *Proc. of Third Conf. on Image and Graphics*, 2004.
- [11] T. Majoor. Face detection using color based region of interest selection. Technical report, University of Amsterdam, 2000.
- [12] G. Mariottini and D. Prattichizzo. Egt: a toolbox for multiple view geometry and visual servoing. *IEEE Robotics and Automation Magazine*, 3(12), Dec 2005.
- [13] G. P. Shi Han and Z. Wu. Human face orientation detection using power spectrum-based measurements. In *Proc. of the sixth IEEE Conf. on AFGR*, 2004.
- [14] M. Turk and A. Portland. Eigenfaces for recognition. *J. Cognition Nueralscience*, 3(1):71–86, 1991.
- [15] N. Uchida, T. Shibahara, and T. Aoki. Face recognition using passive stereo vision. In *Proc. of ICIP*, 2005.
- [16] C. Wiles, A. Maki, and N. Matsuda. Hyperpatches for 3d model acquisition and tracking. *IEEE Trans. on PAMI*, 23:1391–1403, 2001.