PERGAMON

Pattern Recognition 34 (2001) 1565–1572

PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Fusion of perceptual cues for robust tracking of head pose and position☆

Jamie Sherrah*, Shaogang Gong

Department of Computer Science, Queen Mary and Westfield College, University of London, Mile End Road, London E1 4NS, UK

Received 15 May 2000; accepted 15 May 2000

Abstract

The paradigm of *perceptual fusion* provides robust solutions to computer vision problems. By combining the outputs of multiple vision modules, the assumptions and constraints of each module are factored out to result in a more robust system overall. The integration of different modules can be regarded as a form of *data fusion*. To this end, we propose a framework for fusing different information sources through estimation of covariance from observations. The framework is demonstrated in a face and 3D pose tracking system that fuses similarity-to-prototypes measures and skin colour to track head pose and face position. The use of data fusion through covariance introduces constraints that allow the tracker to robustly estimate head pose and track face position simultaneously. © 2001 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Data fusion; Pose estimation; Similarity representation; Face recognition

1. Introduction

The approach we have taken to computer vision, referred to as *perceptual fusion*, involves the integration of multiple sensory modules to arrive at a single perceptory output. The sensory modules all use the same physical sensor, the video camera, but compute different information. A data fusion approach is needed to integrate these different sources of perceptual information.

Data fusion is traditionally used to increase the accuracy of the measurement being performed, and to overcome unreliability in sensors or uncertainty in sensor outputs. There is another benefit of data fusion which is particularly useful for computer vision problems. Different sources undergoing fusion are usually based on different assumptions, some of which may be invalid at any given time. By performing data fusion, the assumptions are in a way "factored out" [1]. Hence fusion can reduce a system's dependence on invalid a priori assumptions and make the system more robust.

Given that data fusion is a beneficial approach, the primary issue is how to combine or fuse the outputs of systems that are possibly disparate. We propose the use of covariance estimation to fuse the outputs of perceptual sources. The covariance of the module outputs is estimated from training examples, and then used in the overall system to impose constraints. By imposing mutual constraints on the observed quantities, the covariance-estimation approach improves the robustness of the system, and has the advantage that the constraints are derived (learned) from practical measurement rather than heuristics. We demonstrate this approach in a case study on pose estimation which implicitly also requires accurate face position alignment and tracking over time. In Section 2, we describe head pose estimation based on similarities to prototypes. Tracking of pose and face position is performed using the CONDENSATION algorithm [2]. The covariance of the state quantities is learned from examples in order to estimate the state propagation density. The correlation between face and

☆Part of this work is funded by EPSRC ISCANIT Project GR/L89624.

* Corresponding author. Tel.: + 44-20-7882-5230; fax: + 44-20-8780-6553.

E-mail address: jamie@dcs.qmw.ac.uk (J. Sherrah).

PII: S0031-3203(00)00091-1

head positions is used to model the state-conditional observation likelihood function. Experiments are given in Section 3 and we draw conclusions in Section 4.

2. Fusion of head pose and position alignment

Automatic, robust pose estimation from a video sequence in real time is non-trivial. It implicitly requires pose-invariant face detection. We have previously developed a method for identity-invariant pose estimation based on similarities to faces in a prototype database [3]. Under suitable conditions, a temporal trajectory of tilt (elevation) and yaw (azimuth) angles can be computed from a video sequence. However, the similarity-based criterion is noisy and has many local optima, and the face position in the image must be determined independently. The existing method either relies on a Polhemus orientation and position sensor worn by the subject to obtain the face position or relies on the identity of the subject [3,4]. Here we describe a method to track both head pose and face position by fusing similarity measures and skin colour information. This is only made robust through the estimation of their covariance.

2.1. Pose estimation in similarity space

Using the similarity-based method, a novel face is represented by a vector of similarities to prototype faces. This concept is illustrated in Fig. 1. For pose estimation, the similarity vector of a novel face is computed for a hypothesised pose and compared with vectors at other poses. In the case that similarities are measured as Euclidean distances, one would expect the magnitude of the vector at the correct pose to be a minimum for the correct pose. However, this criterion is subject to many local minima. To further constrain the criterion to focus on the relevant optimum, one can assume that similarity vectors vary smoothly as the subject changes pose, and use a compound criterion including the distance between the current and previous similarity vectors:

$$S(t) = \alpha||\mathrm{s}(t)|| + (1-\alpha)\,\mathrm{d}(\mathrm{s}(t), \mathrm{s}(t-1)), \quad (1)$$

where $\mathrm{s}(t)$ is the similarity vector at time t, $\alpha \in [0, 1]$ is a real-valued mixing parameter.

A major problem in automatic real-time pose tracking is that one does not know where, in the current image, the face is located. Given an initial position and pose for the face, one could assign a search region in pose and image space and seek the minimum value of the criterion in Eq. (1). This method is impractical, however, since the criterion $S(t)$ is subject to the following sources of noise:

- local optima distract the search toward the wrong position, scale and pose,
- the input face may be poorly aligned with the database images,
- the illumination conditions may vary, and
- the database images themselves may be poorly aligned both in position and pose.

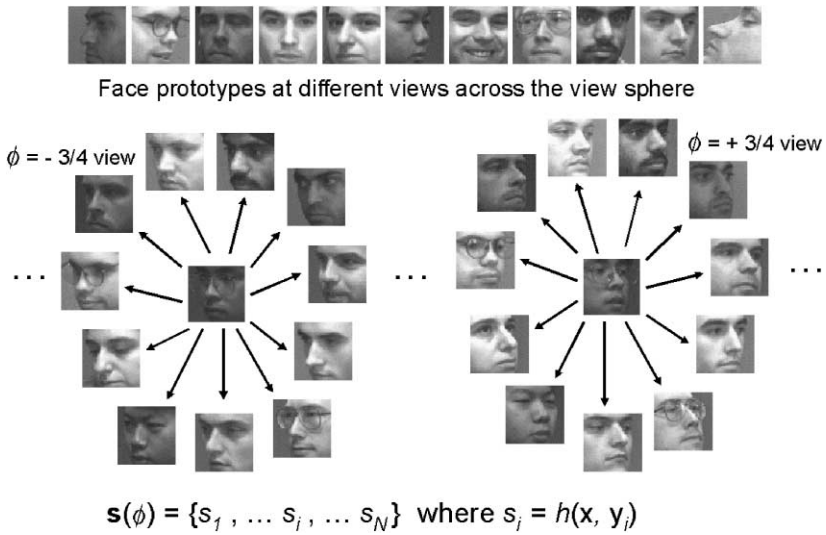


Fig. 1. Illustration of the similarity-to-prototypes representation. A database contains example faces (top) of N different people at different pose angles. For a given pose, a new face $\mathbf{x}$ (centre of circle) is compared with each prototype face at that pose $\mathbf{y}_i$ (perimeter of circle) using a similarity function $h(\cdot)$. The similarity measures s_i are concatenated into a similarity vector $\mathbf{s}$. In the example shown, eleven prototypes were used to represent a face therefore its similarity vector at any pose has 11 dimensions.

The noise in similarity measures can be compensated by incorporating other visual cues.

2.2. Tracking pose using CONDENSATION

Since the similarity measures are noisy, a tracking algorithm is required to simultaneously track pose, face position and scale. These different quantities can be fused together by the tracker. This approach is likely to fail if the quantities are assumed to be independent because the tracker state will then be under-constrained. Through covariance estimation of the tracked parameters, the tracker can be better constrained and become more robust. For our purposes, however, the tracker can still be easily misled by local optima in similarity measures and loses track of both face position and pose. The tracker can be made more robust by incorporating additional visual information such as skin colour to determine the approximate face position. Skin colour is an inexpensive but effective visual cue that can be easily computed in real-time at each frame [5]. Using skin colour, a separate head tracker can be used to supply a bounding box of the head position in the image. While the head box is generally larger than the face box, the displacement between head and face box position provides an additional constraint. In particular, correlations between head pose and the face-head displacement can be exploited.

For the tracking task, we adopt the CONDENSATION (CONditional DENSity propagATION) algorithm [2]. CONDENSATION is a particle filtering method which models an arbitrary state distribution by maintaining a population of state samples $\mathbf{x}_i(t)$ and their likelihoods. Compared to a Kalman filter (a single Gaussian density based model) commonly adopted for temporal tracking, CONDENSATION is more generic and flexible due to its propagation of arbitrary density models. In a sense the state samples can be considered as multiple hypotheses for the current state of the system, hence CONDENSATION is better suited to recover from distractions. In our case, such distractions generally manifest as local optima in the face similarity measure criterion. When a new measurement ($\mathbf{z}_t$) is obtained, the state samples and their probabilities are updated through two steps:

(1) Drift and diffusion: the state samples are modified through a deterministic component obtained from knowledge of the problem, and random perturbation based on the probability distribution of stage changes. The overall step equates to sampling from the distribution $p(\mathbf{x}_i(t)|\mathbf{x}_i(t-1))$.
(2) Measurement: a measurement is imposed on the state distribution by calculating the likelihood $p(\mathbf{z}_t|\mathbf{x}_i(t))$ for each state sample. This likelihood function again comes from knowledge of the problem.

Using CONDENSATION to track head pose and face position, the state is defined to consist of the object-centred face position (x, y) with respect to the body, the scale r of the face box, the head yaw (azimuth) ϕ and the head tilt (elevation) θ:

$$\mathbf{x} = [x, y, r, \phi, \theta]. \tag{2}$$

The scale r is the ratio of the face box size to the prototype image size in the similarity database. The measurements used by the tracker are the input image containing the face, and the head box position from an independent skin colour tracker.

To track the state using CONDENSATION, two distributions $p(\mathbf{x}_i(t)|\mathbf{x}_i(t-1))$ and $p(\mathbf{z}_t|\mathbf{x}_i(t))$ must be modelled. It is in the modelling of these distributions that tracked and measured quantities are fused through covariance estimation.

2.3. Fusing state quantities

For the state propagation distribution $p(\mathbf{x}_i(t)|\mathbf{x}_i(t-1))$, previous applications of CONDENSATION [6] have used a heuristic drift equation, and then arbitrarily added independent Gaussian noise to each element of the state. This approach has two problems. First, the noise parameters are not estimated from measurement, and could cause the tracker to lose lock. Second, the assumption of independence of state elements under-constrains the search space so that computational resources are wasted, and the tracker is distracted by local optima.

Our approach is to fuse the state elements by estimating their covariance. The rationale is that when a person turns their head, there is a correlated change in face box position. Let the state change between two frames be

$$\Delta\mathbf{x}(t) = \mathbf{x}(t) - \mathbf{x}(t-1). \tag{3}$$

A *state transition covariance matrix* was estimated from training video sequences of people varying their head pose freely in a number of scenes resulting in 454 sample frames. Head pose and face position were measured using a Polhemus sensor attached to the subject's head. The estimated state transition covariance matrix is

$$\Sigma_x = \begin{bmatrix} & \Delta x & \Delta y & \Delta r & \Delta\phi & \Delta\theta \\ \Delta x & 12.090 & 0.558 & 0.019 & 15.551 & 0.153 \\ \Delta y & 0.558 & 3.495 & 0.021 & 1.026 & 4.239 \\ \Delta r & 0.019 & 0.021 & 0.112 & 0.237 & 0.171 \\ \Delta\phi & 15.551 & 1.026 & 0.237 & 27.701 & 0.816 \\ \Delta\theta & 0.153 & 4.239 & 0.171 & 0.816 & 7.854 \end{bmatrix}. \tag{4}$$

All distances are measured in pixels and angles measured in degrees. For simplicity, changes in size (Δr) are

measured in pixels rather than as a change in ratio. There are clearly correlations between the state changes which are intuitively appealing. There is a strong correlation between change in x-position and yaw, and between changes in y-position and tilt. Changes in the horizontal quantities have a higher magnitude than changes in the vertical quantities.

It is precisely these constraints that will make the CONDENSATION state sampling more robust and efficient.

Assuming that these correlations are independent of absolute pose and position, the state update distribution is modelled as a fully-covariant Gaussian

$$p(\mathbf{x}(t)|\mathbf{x}(t-1)) = \frac{1}{\sqrt{2\pi}|\Sigma_x|^{1/2}} \exp\left(-\frac{1}{2}(\Delta\mathbf{x}(t))^{\mathrm{T}}\Sigma_x^{-1}(\Delta\mathbf{x}(t))\right). \quad (5)$$

2.4. Fusing measurements

The state-conditional distribution $p(\mathbf{z}_t|\mathbf{x}_i(t))$ is based on the similarity criterion and on the displacement between the face and head position. Let the signed x-difference

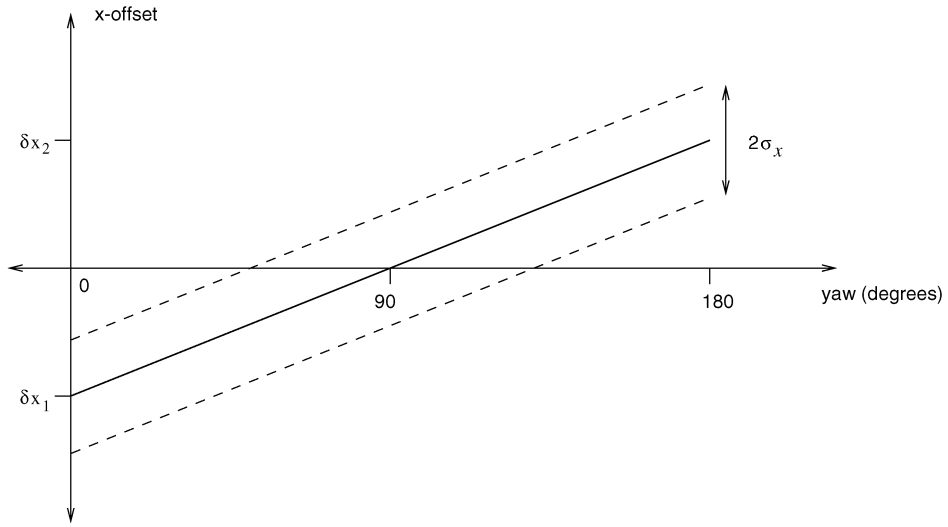


Fig. 2. An illustration of head and face box offset probability density in x-direction.

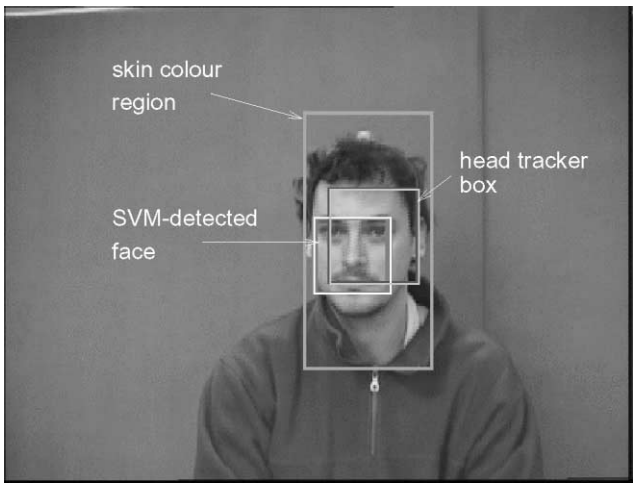


Fig. 3. Example of the information used to initialise the pose tracker.

between the centres of the face and head boxes be δx. The state-conditional distribution is then

$$p(\mathbf{z}_t|\mathbf{x}_i(t)) = p(S(t)|\mathbf{x}_i(t))\, p(\delta x|\phi), \tag{6}$$

where $p(S(t)|\mathbf{x}_i(t))$ is the similarity-based weighting function given the hypothesised state, and $p(\delta x|\phi)$ is a modelled density function expressing the dependence of face box x-displacements on yaw angle. The latter function incorporates observed correlations between absolute face position and pose. Displacements in the y-direction are too unreliable to be used due to varying neck lines, hair colour and illumination conditions. The two components of Eq. (6) work together to constrain the tracker to the correct pose and face position. The two constituent densities are defined as follows:

(1) The similarity-based weighting function gives high probabilities for low dis-similarity values, and vice versa:

$$p(S(t)|\mathbf{x}(t) = \exp\left(-\frac{(S(t) - S_{\min})^2}{2\sigma_s^2}\right), \tag{7}$$

where $S_{\min}$ is the minimum and σ_s is the standard deviation of S-values observed over a set of training sequences.

(2) The displacement density function is based on the observation that facial position displacements are correlated with pose. For example, as a subject turns his head to the left, the box surrounding the face moves left-of-centre of the body, while the box surrounding the head tends to stay central. Therefore covariance between absolute face position and head pose is exploited, whereas only relative face position is used in Eq. (5). The function also constrains the face position to lie close to the independently-tracked head position so that the tracker is not distracted by non-faces.

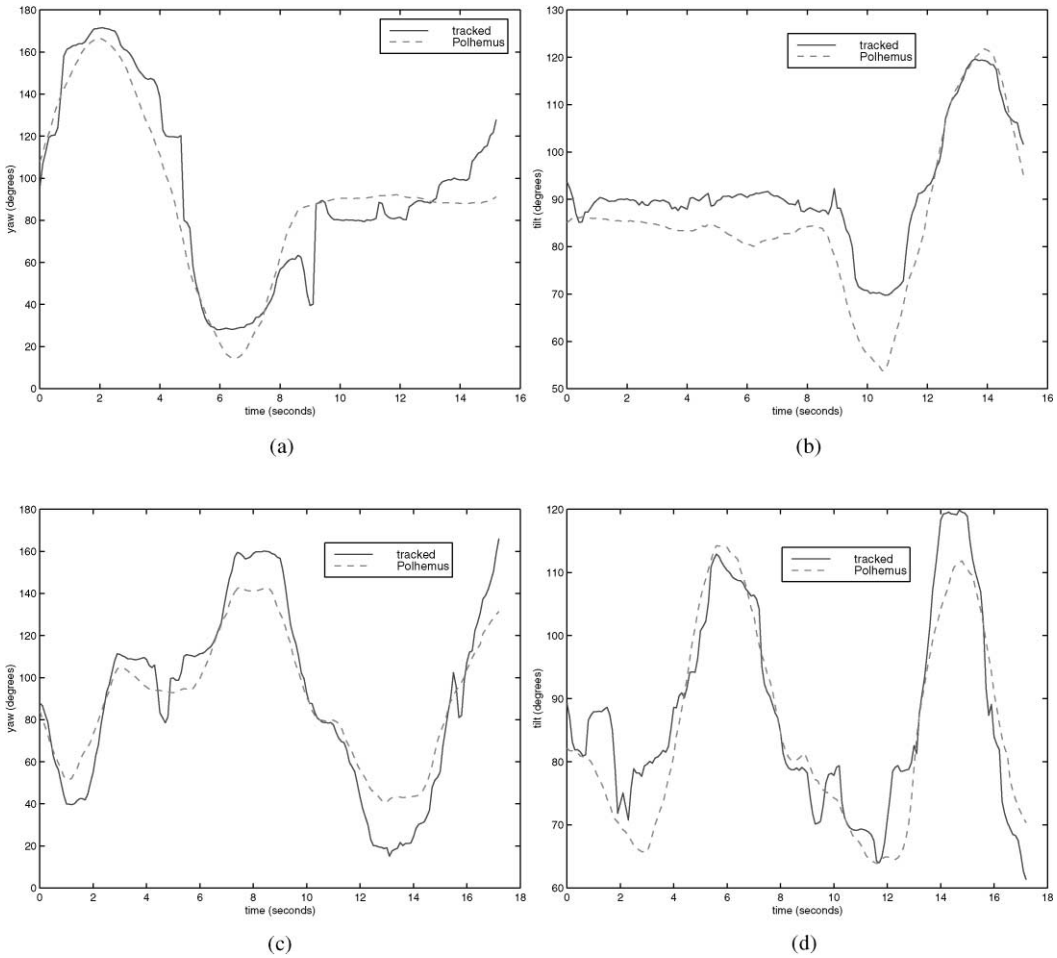


Fig. 4. Measured and tracked head pose angles for two example test sequences. (a) Yaw. (b) Tilt. (c) Yaw. (d) Tilt.

The displacement density model is shown schematically in Fig. 2. The density function is modelled as a Gaussian distribution with standard deviation σ_x and mean dependent on the currently hypothesised yaw angle. The solid line in the figure shows how the mean varies with yaw. The displacements δx_1 and δx_2 are

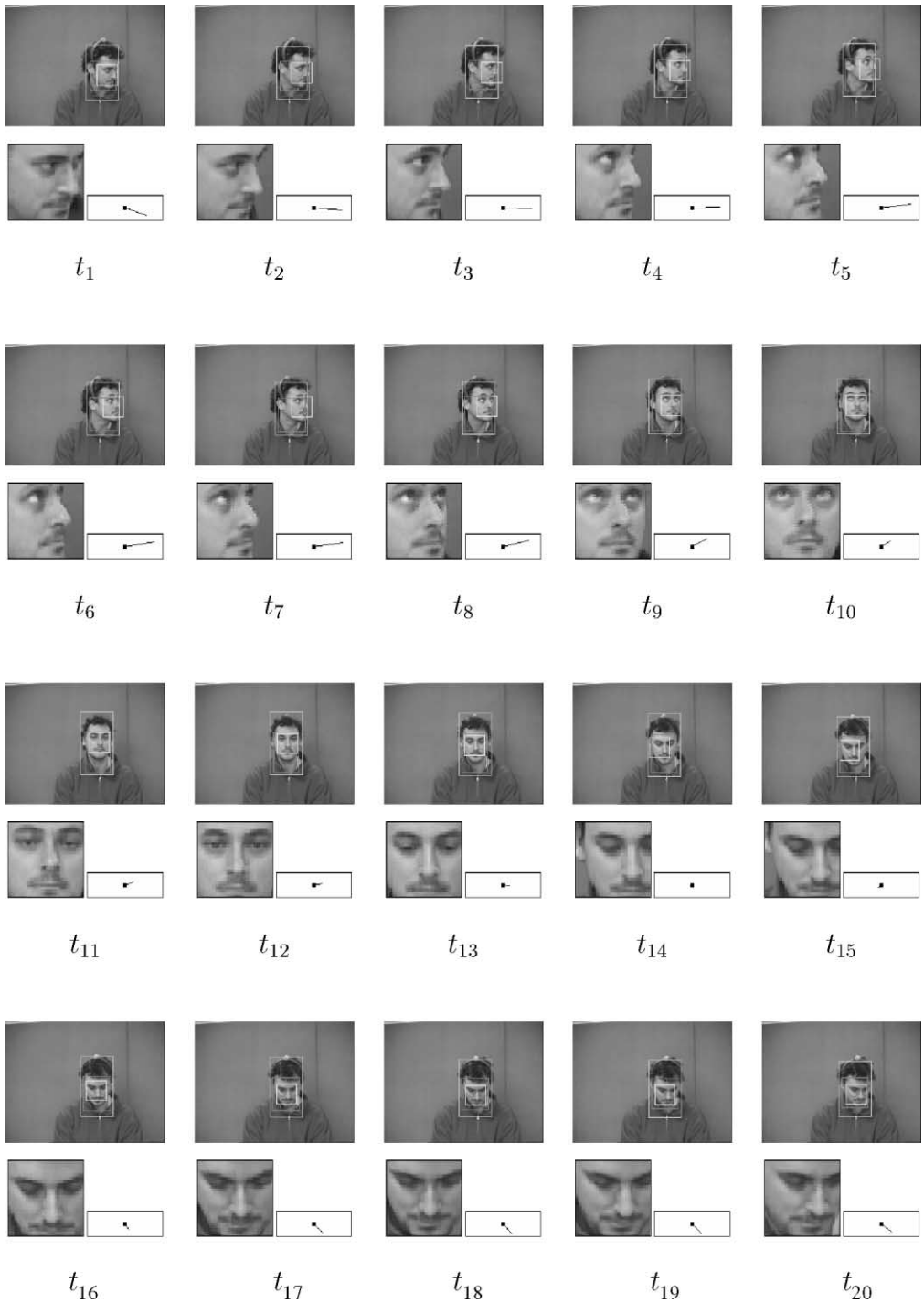


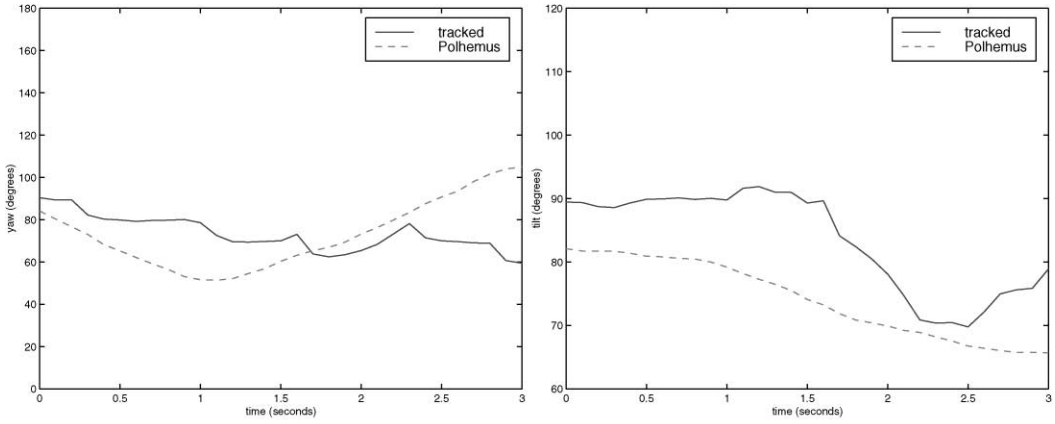
Fig. 5. An example of continuous face position alignment and pose tracking over time. Each frame shows the whole image (top), the cropped tracked face (lower-left), and the tracked head pose (lower-right).

measured from video sequences at the extremes of pose, 0° and 180° yaw. At frontal views high probabilities are given to small displacements, whereas at extremes of yaw the high probabilities go to larger signed displacements. At any yaw, a state with a face position that is far from the head tracker position is given a low probability.

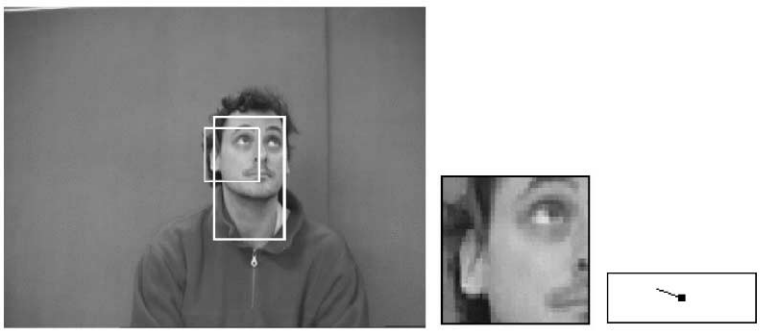
3. Experiments

A system based on the method described has been developed and tested on both recorded and live video sequences of subjects constantly varying their head pose. Two sequences are shown here for illustration. Each sequence is 200 frames long and the initial face position is detected using a support vector machine (SVM) based generic face model [7–9]. The SVM classifier is currently used to recognise faces from near-frontal views only, and can be too computationally expensive to use for every frame during real-time tracking. Fig. 3 shows an example of the initial face box detected by the SVM. The outermost box is obtained by spatially clustering skin-coloured pixels. The SVM searches within this box to find the face box. The head tracker then performs a localised search around the face box to obtain the head box. To simplify the process, a subject is initially assumed to face the camera giving $\theta = 90^\circ$, $\phi = 90^\circ$. The Polhemus sensor is worn by the subject to obtain the approximate ground-truth head pose angles for comparison. Our pose tracker uses only image data, and is independent of the Polhemus.

The yaw and tilt angles estimated and tracked over time by the fusion-based tracker are compared to the measurements of the Polhemus in Fig. 4. Five hundred state samples were used in the CONDENSATION tracker. In both cases, the tracker is able to approximately track the head tilt and yaw angles. Examples of the continuous visual output from the tracker are shown in Fig. 5. Each frame shows the whole image (top), the cropped tracked face (lower-left), and the tracked head



(a)



(b)

Fig. 6. Results of pose tracker on second sequence without the use of covariance. (a) Yaw (left) and tilt (right) angles for first 31 frames of a test sequence. (b) The camera image, tracked face and pose dial at frame 31.

pose (lower-right) using an intuitive dial. In this example, the tracker accurately follows the face and head pose until time t_{14} when the tracker momentarily loses lock on face position and starts to move away from the face. It regains lock again at t_{16}. At t_{15} the system is starting to lose pose but recovers gradually over time. The ability of the tracker to recover from momentary loss of lock demonstrates the importance of fusing the face position and pose. Without this function, the tracker would have wandered away to incorrect poses or non-faces.

To demonstrate the role of covariance estimation in tracker robustness, we remove the covariance information from the tracker. This step requires the off-diagonal elements of the state covariance matrix (Eq. (4)) to be set to zero, and the removal of $p(\delta x|\phi)$ from the state-conditional distribution (Eq. (6)). The tracker fully loses lock without recovering after 31 frames. The results are shown in Fig. 6 for the first 31 frames (3 s). Even though the tracked pose angles have not deviated wildly, the face box is far from its goal. Since the similarity criterion is only locally optimal, the tracker is unable to regain lock.

4. Conclusion

The concept of data fusion through covariance estimation has been demonstrated in a face position alignment and pose tracking system. Face position and head pose were fused to form state update and measurement noise models for a pose tracker. A principled approach to fusion of different visual cues utilises additional constraints and improves robustness.

References

[1] J.J. Clark, A.L. Yuille, Data Fusion for Sensory Information Processing Systems, Kluwer Academic Publishers, Dordrecht, 1990.

[2] M. Isard, A. Blake, CONDENSATION — conditional density propagation for visual tracking, Int. J. Comput. Vision 29 (1) (1998) 5–28.

[3] S. Gong, E. Ong, S. McKenna, Learning to associate faces across views in vector space of similarities to prototypes, British Machine Vision Conference, vol. 1, Southampton, UK, September 1998, pp. 54–64.

[4] S. McKenna, S. Gong, Real time pose estimation, Int. J. Real Time Imaging 4 (1998) 333–347; Real-Time Visual Monitoring and Inspection (special issue).

[5] Y. Raja, S.J. McKenna, S. Gong, Tracking and segmenting people in varying lighting conditions using colour, Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition, Nara, Japan, 1998, pp. 228–233.

[6] M.J. Black, A.D. Jepson, A probabilistic framework for matching temporal trajectories: Condensation-based recognition of gestures and expressions, ECCV, Freiburg, Germany, 1998, pp. 909–924.

[7] J. Ng, S. Gong, Learning support vector machines for a multi-view face model, British Machine Vision Conference, Nottingham, UK, September 1999, pp. 503–512.

[8] J. Ng, S. Gong, Using support vector machines for real-time face detection and pose estimation from frontal to profile views, IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, Corfu, Greece, September 1999, pp. 14–21.

[9] Y. Li, S. Gong, H. Liddell, Support vector regression and classification based multi-view face detection and recognition, IEEE International Conference on Face and Gesture Recognition, Grenoble, France, March 2000.

About the Author—JAMIE SHERRAH is a post-doctoral research fellow at the Machine Vision group of Queen Mary and Westfield College, London. Jamie received a first-class honours degree in computer systems engineering from the University of Adelaide, Australia in 1995, after which he embarked on a project with the Image Analysis group of the CSIRO, Sydney investigating symmetry detection in 3D data sets. Jamie then returned to Adelaide University to undertake a Ph.D. in collaboration with the Co-operative Research Centre for Sensor, Signal and Information Processing, investigating the use of genetic programming for pattern recognition. Jamie received his Ph.D. degree in 1998. His research interests include pattern recognition, adaptive systems, human tracking and behaviour modelling.

About the Author—SHAOGANG GONG is Reader in Computational Vision and Learning at the University of London, England. He received his B.Sc. in information theory and measurement from the University of Electronic Sciences and Technology of China in 1985 and his D.Phil. in computer vision from Oxford University in 1989. He was a recipient of a Sino-Anglo Queen's Research Scientist Award in 1987, a Royal Society Research Fellow in 1987 and 1988, a GEC-Oxford Industrial Research Fellow in 1989, and a research fellow on the European ESPRIT II Project VIEWS between 1989–1993. His current research interests include dynamic vision, the modelling of human faces, gestures and behaviour, temporal prediction models, visual learning, Bayesian and statistical learning theories, visual surveillance and visually mediated interaction.