Stereo 3D Lip Tracking

Gareth Loy, Roland Goecke, Sebastien Rougeaux and Alexander Zelinsky

Research School of Information Sciences and Engineering
Australian National University, Canberra 0200, Australia
{gareth, roland, rougeaux, alex}@syseng.anu.edu.au

Abstract

A system is presented that tracks in 3D a person's unadorned lips, and outputs the 3D locations of the mouth corners and ten points describing the outer lip contour. This output is suitable for audio visual speech processing, 3D animation, or expression recognition. A stereo head tracker is used to track the subject's head, allowing for robust performance whilst the subject's head is moving and turning with respect to the cameras. The head pose is used in conjunction with the novel *adaptable templates* to generate a robust estimate of the deforming mouth corner locations. A 3D geometric model is used to generate search paths for key points on the outer lip contour which are subsequently located using adaptable templates and geometric constraints. The system is demonstrated robustly tracking the head pose and 3D mouth shape on a person speaking while moving his head.

1 Introduction

Tracking the lips has a broad scope of applications across the field of human-computer interaction (HCI), from animation, to expression recognition [11], to audio visual speech processing [1], [4]. A number of techniques have been reported to extract mouth features from facial images. Active contour models [5] have been used to detect lip contours [1], [2] and Li *et al.* [7] applied the eigensequence approach that is often used in facial expression recognition.

There are two limitations of these systems. Firstly, they require the subject to be directly facing the camera, and do not allow any head movement that would distort the lip shape in the captured images. Secondly, they only track the mouth in 2D, and are unable to describe the full 3D mouth shape.

The mouth is a 3D feature which deforms in all dimensions. In order to fully describe the mouth shape it is necessary to track it in 3D. Providing such a description of the mouth shape is essential for accurate 3D character animation, and also provides significantly more information for audio-visual speech processing and HCI other applications.

As people talk, their heads naturally move about as they gesture and follow conversation cues. It is necessary for a lip tracking system to be robust with respect to this behaviour; to be able to detect, monitor and account for movement of a speaker's head.

We have developed a 3D lip tracking system that allows the speaker's head to move naturally. The basis of our system is a real-time stereo face tracker which robustly tracks the subject's head [9]. Using the head pose information from the face tracker to correct for head movement, our lip tracker tracks the 3D shape of the subject's mouth as it deforms through speech and other motion.

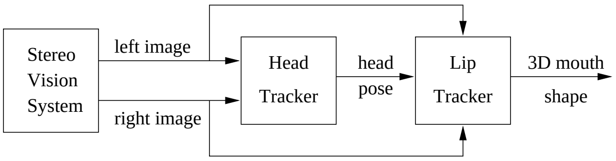


Figure 1: Overview of the system.

Figure 1 shows the key components of the system presented in this paper. Section 2 describes the stereo vision system used to capture images of the subject, Section 3 describes the head tracking software, and Section 4 covers the lip tracking system. Section 5 presents some results from the system, and finally Section 6 concludes with some discussion of the future direction of this work.

2 Stereo Vision System

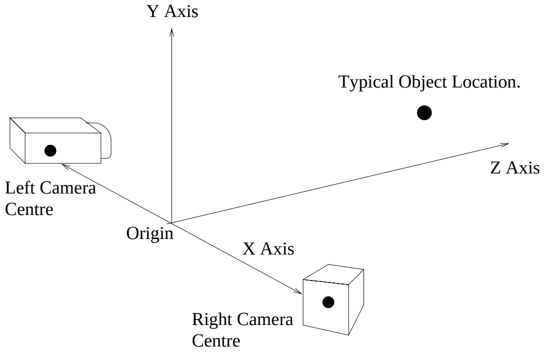


Figure 2: The stereo camera arrangement.

The layout of the stereo vision system is shown in figure 2. The two cameras are positioned equidistant from the origin and are verged (angled towards the origin in the horizontal plane) at about 5^o. This is designed to offer the best measurements of an object the size of a human head placed approximately 600mm in front of the cameras.

Both cameras are standard, colour analog NTSC video cameras whose outputs are multiplexed into a single channel before being acquired by a Hitachi IP5005 video card. The result is a 512 × 480 colour image, captured every 33ms, where the top half contains the right hand image and lower half the left hand image.

3 Head Tracking

The head tracking software consists of the following parts, each described separately below:

1. 3D Facial Model Acquisition
2. Face Acquisition
3. 3D Face Tracking

3.1 Face Model Acquisition

The *pose* of a rigid body is defined as the rotation and translation that maps a set of 3D *model* points to their observed 3D locations. Although identifying such model points is ideally an automatic process, the best results are obtained by identifying a set of features in a stereo image manually.

The face *model* consists of up to 32 features ($T_i, i = 0, 1, 2, ...$) corresponding to a set of 3D points in the head reference frame. The head frame is placed between the eyes and oriented as shown in Figure 3.

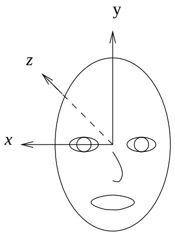


Figure 3: The head reference frame.

3.2 Face Acquisition

The system starts in this mode where it attempts to find an initial lock on the face in the image stream. During this phase a template constructed from the edge map of the entire central region of the face is searched for. This template is automatically extracted during the model acquisition phase where the position of the face in the image is known. Normalised correlation matching is used both here and during tracking to make this process insensitive to changes in lighting conditions.

When a match is found with a correlation above a preset value, the approximate positions of the features T_i are easily identified based on their known offsets from the centre of the face (again calculated during model acquisition).

3.3 Face Tracking System

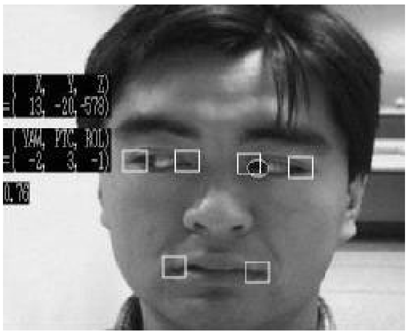


Figure 4: 3D face tracking.

Tracking is performed using the templates T_i obtained during model acquisition. These are correlated with the current stereo view in the input stream and their $3D$ positions are calculated.

The optimal pose (rotation and translation) is calculated that best maps the model to these 3D positions. The strength of each template correlation becomes a weight in a least squares minimisation procedure. Thus the solution is biased towards points which track well, making the results robust to occlusions, noise and perspective distortions. The model points are transformed by the optimal pose estimate and back projected onto the image plane where they are used to locate the search areas where each template is searched for in the subsequent frame.

The number of templates tracked can be less than the total number. This allows the system to continue tracking when some templates suffer severe perspective distortion or are occluded altogether. The best templates to track can be determined from the estimated head pose as those that are visible and will appear most fronto-parallel to the image plane.

Figure 4 shows the system in operation. The system tracks in real-time (NTSC video frame rate).

4 Lip Tracking

The mouth is a deformable feature, however, there are a number of restrictions on its behaviour: it is firmly attached to the head, is only able to deform in an elastic manner, and its shape tends to remain close to symmetric (especially during speech). Our lip tracking system uses these properties in conjunction with visual cues and a specific search structure to to track the mouth in 3D.

The lip tracker is initialised along with the head tracking templates (section 3.1). Initialisation is performed on a frame where the subject is facing close to front-on to the cameras and the mouth is in a neutral (closed and relaxed) position. The mouth corners and centre of the upper outer lip contour are identified (as with the head tracking templates this is currently done manually), these are matched in the other stereo image to determine the 3D neutral mouth locations. In addition to this, the initial locations of ten tracking points on the outer lip contour are identified on the lip contour search lines which are described below in section 4.1.

The overall sequential operation of the lip tracker is shown in Figure 5a, and the process for tracking each set of features, namely the mouth corners, and the upper and lower lip contours, is illustrated in Figure 5b. Initially search regions are defined. Visual cues are then determined across the search region, giving a number of possible feature locations. The 3D location of each possible feature is determined via stereo matching, and a cost function is determined for every combination of possible feature locations. The final set of estimated feature locations is chosen to minimise the cost function.

4.1 Search Structure

The appearance of the mouth is constantly changing. In order to effectively track the mouth contour we adopt a search structure that allows us to reliably locate specific mouth features. This approach is a further development of our previous work in 2D lip tracking [8].

The search procedure consists of the following steps:

1. search areas for the mouth corners are defined from the projected positions of the neutral mouth corners (determined from the head pose),
2. the current mouth corner locations are identified,
3. search lines are defined for locating the outer lip contour, and
4. the outer lip contour is identified.

The corners of the mouth are the most suitable points for tracking due to their distinctive (albeit drastically changeable) appearance. Their location is constrained to a small region of the face. Each mouth corner will always be in a region centred at the neutral mouth corner location. Because this region is quite small (typically 50 × 50mm) it is feasible to search the whole region in every frame.

From this search several potential mouth corner locations are identified, and the best pair chosen by minimising a cost function as discussed in section 4.3.

In addition to the corners, ten points are used to characterise the outer lip contour, five each for the upper and lower lip. These points are located on search lines parallel to the vertical head axis. Figure 6 shows the upper and lower search lines, the mouth corner tracking points and the lip contour tracking points.

The search lines are initially defined in 3D. Two search lines are placed on each side of the mouth, spaced equidistant between the mouth corner and the centre of the neutral mouth, and a fifth search line is placed at the centre of the mouth. These 3D lines are projected onto one of the stereo images to provide the set of 2D search lines on which to locate the outer contour in the image.

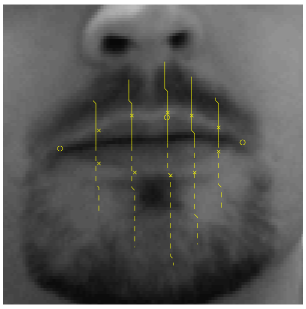


Figure 6: Search lines for the upper and lower contours, drawn solid and dashed respectively. Contour features shown with crosses, mouth corners and neutral mouth centre as circles.

For the upper lip contour the search lines start at the level of the lip corners (in the head reference frame) and extend upwards, for the lower lip contour the search lines start at the level of the lip corners and extend downwards. The length of the search lines is proportional to how central the point is, and is determined in 3D, before the lines are projected onto the image plane. For our experiments the 3D search lines were chosen to be 14, 20 and 24mm for the upper contour and 21, 30 and 36mm for the lower contour.

Locating the mouth edges on these lines avoids the potential problem of the templates drifting along the top and bottom mouth edges, and the computational requirement for searching is drastically reduced by only examining a line of points, rather than a 2D region.

Up to three potential contour points are identified on each search line using image cues, and the optimal set of contour points is determined by minimising a cost function, as discussed in section 4.3 below.

4.2 Visual Cues

After identifying the search region for a particular feature, be it a mouth corner or contour point, visual cues are used to determine a set of potential feature locations within this space. Processing time is the only factor restricting the number and type of cues used.

For this paper we have implemented a single visual cue; we use correlation with *adaptable templates* — a new form of template especially developed for tracking elastically deformable features — these are discussed in detail below in section 4.2.

Up to four possible locations are determined for each i^{th} mouth feature by correlating an adaptable template M_i across the search region or search line. The normalised correlation is used, and the correlation coefficient at each location in the search region is stored in

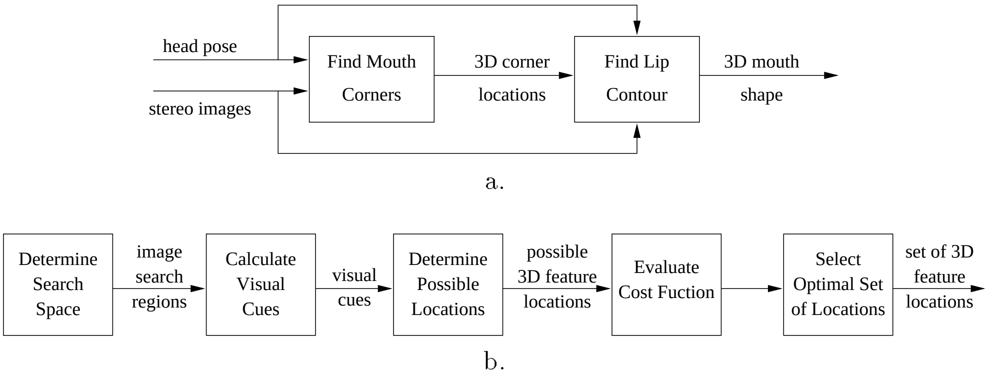


Figure 5: a. Lip tracking system. b. Process for identifying a lip feature.

a matrix $C(M_i)$. We determine a set a possible feature locations $P(M_i)$ as the local maximums of $C(M_i)$. These local maximums are located by calculating the *rank transform* [12] of $C(M_i)$ with radius 1. The rank transform is defined as the number of pixels in a local region with intensity less than the centre pixel. A point in a 2D matrix with a rank transform of 8 is a local maximum, while a local maximum on a line will have a rank transform of 2. If there are no elements with a rank transform of the maximum value, $P(M_i)$ is defined as the set of elements with the highest rank transform.

In order to limit the computational load we restrict ourselves to considering up to four possible locations for each mouth corner and three possible locations for each contour point. If $P(M_i)$ has more than the desired number of elements we choose those with the highest correlation coefficients.

Adaptable Templates

The changing appearance of the mouth features — especially the corners — makes fixed templates ineffective. To address this issue *adaptable templates* have been developed for the tracking of elastically deformable features. They make use of the initial object appearance in a similar vain to prototype-based deformable templates [13].

In order for templates to maintain adequate tracking performance whilst tracking deformable features it is necessary to dynamically adapt the templates to keep them up-to-date with the changing appearance of the target. An obvious approach is to update the templates each tracking cycle to equal the new feature appearance, however, this raises several problems. If there is ever an error in the template matching and the new template is chosen off target there is no way for the error to be recovered. The chance of this can be minimised by only updating templates when the match is very good. However, even if there is never any incorrect matching the templates will do a random walk about the image over time. This is due to the quantisation error (of up to half a pixel) present in each template match.

A solution to this problem is to ground each new template with a portion of the initial template, which is assumed to have been chosen correctly, and thus contain the desired target. This is the basis of adaptable templates. For the k^{th} frame, once the best match is found (and provided the correlation is above a certain threshold) the template $M_i[k]$ is updated to become the weighted average of the initial template $M_i[0]$ and the image region $R_i[k]$ in the new frame that best matches the current template, that is

$$M_i[k+1] = \alpha M_i[0] + (1-\alpha) R_i[k],$$

where the constant $\alpha \in (0...1)$ is the *grounding factor* which determines the contribution of the initial template to the new template. $\alpha = 0$ is the case of fully updated templates and $\alpha = 1$ gives standard templates.

Updating the templates is crucial as the appearance of the mouth corners changes drastically with the mouth shape. Figure 7 shows how the appearance of a mouth corner changes over time. The initial template was taken with the mouth in a neutral position, by frame 223 the mouth is wide open and the initial template is no longer suitable for locating the feature. The adaptable template formed from these two images is also shown.

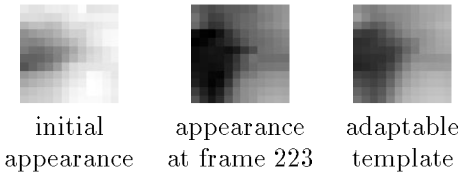


Figure 7: Example of an adaptable template for the corner of an open mouth, with $\alpha = 0.33$.

4.3 Incorporating Physical Mouth Constraints

As with the visual cues, computational load is the only limitation to the number and complexity of the physical-model-based constraints placed on the tracked points.

We have already introduced several implicit assumptions about the form of the mouth in the way the search spaces were defined in section 4.1. Additional constraints are imposed to identify the optimal — or most 'mouth-like' — pair of mouth corner locations, and set of locations for the upper and lower lip contour tracking points.

A cost function is evaluated for every combination of possible feature locations and the optimal estimated feature locations are chosen to minimise this function. This technique is applied first to the mouth corners, independently of the lip contour. Once the mouth corner locations have been fixed the upper and lower contour points are each determined separately.

The cost function consists of two components: one based on visual cues (in our case the negative of the normalised correlation coefficients for the adaptable templates) and one based on physical mouth constraints.

For the mouth corners the physical constraint is the asymmetry q of the displacement between the candidate corner locations and the neutral mouth corners, about the vertical head plane, as shown in Figure 8.

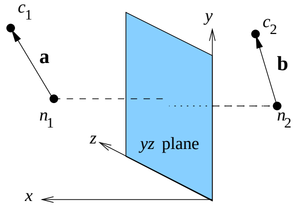


Figure 8: Determining the asymmetry of the mouth corners c_i with respect to the neutral mouth locations n_i and the yz plane in the head reference frame.

$$q = \sqrt{(\mathbf{a}_1 + \mathbf{b}_1)^2 + (\mathbf{a}_2 - \mathbf{b}_2)^2 + (\mathbf{a}_3 + \mathbf{b}_3)^2}$$

For each lip contour the physical constraint is the sum of the distances between adjacent points on the contour (including the corners) in the image plane. A similar constraint is used in the internal energy constraint in Kass *et al.*'s active contour models [6].

The optimal sets of feature locations for the mouth corners, and upper and lower contours are determined by minimising the appropriate cost function.

5 Experimentation

A significant amount of experimentation has already been carried out to analyse the performance of the head tracker using a mannequin head mounted on a pan-tilt device [10]. The head tracker has been shown to accurately recover the head pose within 10 degrees. It can accommodate head velocities of up to 100 degrees per second and head rotation up to 45 degrees away from the cameras.

It has not yet been feasible to formally quantify the performance of the lip tracker, as it is very difficult to determine a ground truth, however, the system has been seen to accurately track the mouth throughout a short sequence of footage of a subject moving his head and mouth in 3D.

Figure 9 shows several snap shots of the system in action. Both left and right stereo images are shown with the lip tracking points indicated by circles. The 3D mouth shape is also shown. A full video sequence of the tracker in action will be presented at the conference, and is available online at

`http://www.syseng.anu.edu.au/rsl`

6 Discussion and Further Work

This paper has presented a technique to track the 3D shape of a deforming mouth whilst the subject's head is moving in 3D. The mouth corners are tracked along with ten points on the outer lip contour, and the 3D locations each of these points determined via stereo correspondence. The lip tracking results in this paper were generated off-line, however, the technique is efficient enough for real-time implementation. To this end the computation has been limited where ever possible.

At present the lip tracking system relies only on grayscale intensity information and some simple geometric constraints generated from the 3D head model. We wish to extend the system to utilise a number of other cues which will be merged together to increase the robustness and versatility of the system. Colour image information is one such cue which shows great promise for lip tracking and analysis of the mouth region [3].

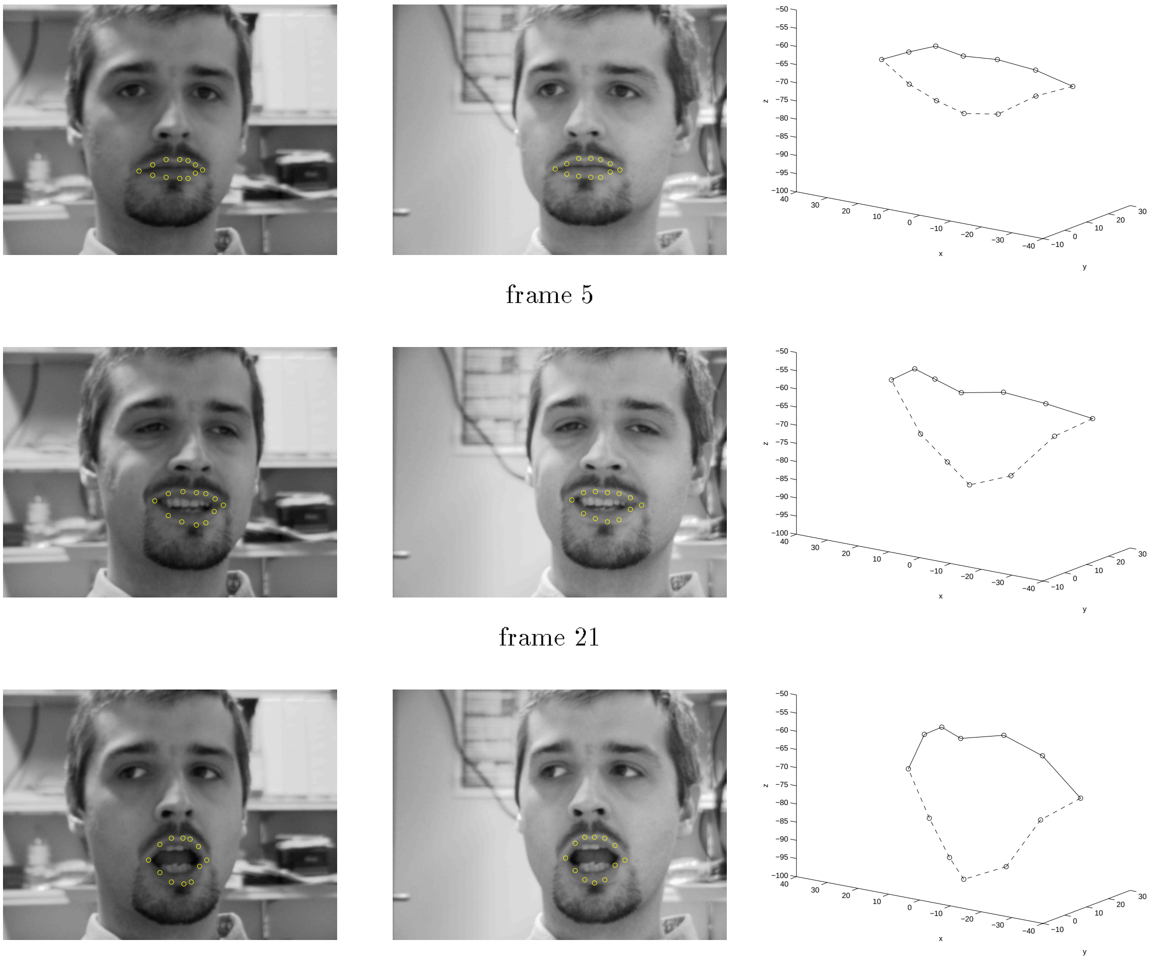
The template matching-based approach adopted in this paper restricts the tracking to the outer lip contour. Whilst this is arguably the more important contour for animation and visualisation purposes, the inner contour is much more useful for audio-visual speech processing, since it is the inner contour which defines which defines the airflow in and out of the mouth [3].

Ultimately it is desired for a lip tracker to track both inner and outer contours in 3D. However, the inner contour is an elusive target, and is difficult to define in 3D. Unlike the outer contour which is a distinctive line on a surface, the inner contour is the boundary between the lip and the oral cavity space and has no definite 3D location. It is expected that any attempt to determine the inner contour in 3D will rely heavily on knowledge of the outer contour.

References

[1] C. Bregler and S. M. Omohundro. Nonlinear manifold learning for visual speech recognition. In *Proc. of 5th International conference on computer vision*, pages 494–499, 1995.

[2] G. I. Chiou and J. N. Hwang. Image sequence classification using a neural network based active contour model



frame 5

frame 21

frame 30

Figure 9: The system in operation. The first two columns show the left and right stereo images and lip tracking points, the third column shows the 3D mouth shape in the head reference frame (dimensions in mm).

and a hidden markov model. In *Proc. of International Conference on Image Processing*, pages 926–930, 1994.

[3] Roland Goecke, J Bruce Miller, Alexander Zelinsky, and Jordi Robert-Ribes. Automatic extraction of lip feature points. In *Australian Conference on Robotics and Automation*, pages 31–36, 2000.

[4] R. Harvey, I. Matthews, J. A. Bangham, and S. Cox. Lip reading from scale-space measurements. In *IEEE*, pages 582–587, 1997.

[5] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. of IEEE First International Conference on Computer Vision*, pages 259–269, 1987.

[6] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. In *Proc. of IEEE* 1^st *International Conf on Computer Vision*, pages 259–269, 1987.

[7] N. Li, S. Dettmer, and M. Shah. Lipreading using eigensequences. In *Proc. of Workshop on Automatic Face and Gesture Recognition*, pages 30–34, 1995.

[8] Gareth Loy, Eunjung Holden, and Robyn Owens. A 3d head tracker for an automatic lipreading system. In *Australian Conference on Robotics and Automation*, pages 37–42, 2000.

[9] Yoshio Matsumoto and Alexander Zelinsky. An algorithm for real-time stereo vision implementation of head pose and gaze direction estimation. In *Proc. of the Fourth International Conference on Automatic Face and Gesture Recognition*, pages 499–504, 2000.

[10] Rhys Newman, Yoshio Matsumoto, Sebastien Rougeaux, and Alexander Zelinsky. Real-time stereo tracking for head pose and gaze estimation. In *Proc. of the Fourth International Conference on Face and Gesture Recognition*, pages 122–128, 2000.

[11] M. Pantic and L.J.M Rothkrantz. Expert system for automatic analysis expressions. *Image and Vision Computing*, 18(11):881–905, August 2000.

[12] R. Zabih and J. Woodfill. Non-parametric local transforms for computing visual correspondence. In *3rd European Conf. Computer Vision, Stockholm*, 1994.

[13] Yu Zhong, ANil K. Jain, and M.-P. Dubuisson-Jolly. Object tracking using deformable templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(5):544–549, May 2000.