Enforcing a Shape Correspondence between Two Views of a 3D Non-rigid Object

M. Benjamin Dias and Bernard F. Buxton

Department of Computer Science, University College London,
Gower Street, London, WC1E 6BT, UK
B.Dias@cs.ucl.ac.uk, http://www.cs.ucl.ac.uk/staff/B.Dias/

Abstract. We have developed an algorithm capable of enforcing a shape correspondence between two views of the same object in different shape-states. This algorithm, together with several other significant updates, has helped improve the performance of the Integrated Shape and Pose Model (ISPM) described in [1] by a factor of 10. The ISPM utilizes two flexible basis views to integrate the linear combination of views technique with a coupled-view Flexible Shape Model (FSM) [2]. As a proof-of-principle we have evaluated the performance of the improved ISPM in comparison to that of its predecessor [1] and of the conventional FSM [3], via two different databases. The results show that, unlike the FSM, the current ISPM is view-invariant and that, on average, it out-performs the FSM. It also out-performs the initial ISPM described in [1].

1 Introduction and Background

Machine vision systems that utilize two or more two-dimensional (2D) images to represent three-dimensional (3D) objects have recently become quite popular because they are sufficient for many purposes, while computationally being relatively easy to build. In particular, not building an explicit 3D model means that we can avoid poorly conditioned 3D reconstruction steps and can, therefore for example, generate virtual images with less noise [4]. There is also some evidence to suggest that such view-based representations are used by the human visual system [5]. Ullman and Basri [6] developed the view-based approach, also known as the Linear Combination of Views (LCV) technique, though only for representing rigid objects. In the LCV technique, any image of a 3D object is represented as a linear combination of at least "$1\frac{1}{2}$" other images of the same object. Ullman and Basri [6] used line-drawings, whilst others have taken this concept further to the combination of real images [7,8] but using an over-complete approach so that the basis views may be teated symmetrically. We reformulated the over-complete LCV approach [1], via the Centred Affine Trifocal Tensor (CATT) [4], introducing the required constraints [9].

Thus far, however, the LCV technique has only been used to represent rigid 3D objects. We have taken this approach even further to model non-rigid 3D objects. For this we integrate an LCV model with a Coupled-View Flexible Shape Model (CVFSM) [2], via two flexible basis views, to form the Integrated

A. Sanfeliu and J. Ruiz-Shulcloper (Eds.): CIARP 2003, LNCS 2905, pp. 163–170, 2003.
© Springer-Verlag Berlin Heidelberg 2003

Shape and Pose Model (ISPM), which was first introduced in [1]. In order to generate such a model we need: (*i*) a technique for mapping the intrinsic shape from any given image, simultaneously to two (or more) preselected views and, (*ii*) a technique for ensuring the two mapped shapes correspond to the same 3D shape (as though they were images captured simultaneously from different views). Given two such techniques, we could train a CVFSM on almost any given set of images, by first transferring the intrinsic shape from each given image to two preselected views providing the required corresponding pairs of images [2]. Once we have two such flexible basis views, we can synthesize, via the CVFSM, an image of the object in any view by, for example, the LCV technique.

The first of the above mentioned techniques, (*i*), which we refer to as a 2D Pose Alignment, was first described in [1]. However, since we were missing technique (*ii*) for ensuring the two mapped shapes correspond to the same 3D shape, the success of the ISPM described in [1], though encouraging, was limited. We have now developed the second technique, (*ii*), which is explained in Sect. 3 of this paper. It has, along with several other new steps, described in Sect. 2.2, helped improve the performance of the ISPM by an order of magnitude.

2 An Implicit (2D) Pose Alignment via the CATT

Here we assume, for the moment, that we are given two images of the mean shape as seen from two preselected views (the basis views) i.e. we have two sets of corresponding landmark points, $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$, respectively from two images of the same shape (the mean shape) as seen from the two basis views. Let's also use two sets of corresponding landmark points $\boldsymbol{X}_i'$ & $\boldsymbol{X}_i''$ to represent the images, as seen from the two basis views, of the shape in a given image, i say, represented by the landmark points $\boldsymbol{X}_i$ (i.e. $\boldsymbol{X}_i$, $\boldsymbol{X}_i'$ & $\boldsymbol{X}_i''$ have the same shape). The aim of the 2D (implicit) Pose Alignment (2D-PA) process is then to recover $\boldsymbol{X}_i'$ & $\boldsymbol{X}_i''$ given $\boldsymbol{X}_i$, $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$.

2.1 The Subset of Stable Points

Before we begin the 2D-PA process, we first select a subset of at least 4 non-co-planar landmark points that can be considered as forming a rigid sub-object. For this we employ a RANSAC algorithm [10] to select a subset of p (≥ 4) landmark points that best conforms to the constraints of multi-view geometry for a rigid object by minimizing

$$e^2 = \sum_i e^2(i) \ , \quad \text{where for each } \boldsymbol{X}_i \quad e^2(i) = \|\boldsymbol{T}(i)\boldsymbol{Y}(i)\|^2 \ \ , \qquad (1)$$

$\boldsymbol{T}(i)$ is the CATT matrix [1] and $\boldsymbol{Y}(i) = (\boldsymbol{X}_i^T, \overline{\boldsymbol{X}}''^T, \overline{\boldsymbol{X}}'^T)^T$. Here, in each case, each image is represented in the $\boldsymbol{X}_i, \overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$, only by the subset of p landmark points being considered. In our experiments we used a subset of 6 stable points (i.e. $p = 6$) and manually checked that the selected points were not co-planar. If the selected subset were co-planar, then we continued to check the subset that provides the next smallest value for e^2 until we found a non-co-planar one.

2.2 The 2D Pose Alignment (2D-PA) Process

Given the stable points, we begin the 2D-PA process, as in [1], by computing the best approximation to the CATT [4] corresponding to $\boldsymbol{X}_i$, $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$ by computing the $\boldsymbol{T}(i)$ that minimizes $e^2(i)$. However, we now use only the subset of stable points for this, which provides a more accurate estimate of the CATT than reported in [1] and makes the process a lot faster. We then use the computed CATT and ***all*** the landmark points (not just the stable points) to generate the least squares estimate of $\overline{\boldsymbol{X}}_i$, the mean shape in the view of the given image $\boldsymbol{X}_i$. Next, we compute the *in-view* shape difference, $\Delta\boldsymbol{X}_i = \boldsymbol{X}_i - \overline{\boldsymbol{X}}_i$, which is then added to $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$, to generate our first estimates of $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$:

$$\boldsymbol{X}'_i(\text{temp}) = \overline{\boldsymbol{X}}' + \Delta\boldsymbol{X}_i \quad \& \quad \boldsymbol{X}''_i(\text{temp}) = \overline{\boldsymbol{X}}'' + \Delta\boldsymbol{X}_i \ . \tag{2}$$

Since $\Delta\boldsymbol{X}_i$ is a shape difference in the view $\boldsymbol{X}_i$, applying it to the basis views will not, in general, lead to a valid result, since $\boldsymbol{X}'_i(\text{temp})$ & $\boldsymbol{X}''_i(\text{temp})$ will not, in general, conform to the constraints of multi-view geometry. However, in each case it provides a better estimate of the landmarks of the pose aligned image than the means $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$. We continue by extracting, from the CATT, the two fundamental matrices that link $\boldsymbol{X}_i$ to each of the basis views via *Algorithm 14.1, on page 366* of [11]. We then use the fundamental matrices to compute the equations of the epipolar lines in each of the $\boldsymbol{X}'_i(\text{temp})$ & $\boldsymbol{X}''_i(\text{temp})$ corresponding to each landmark point in the given image $\boldsymbol{X}_i$. Next, we move each landmark point in each of $\boldsymbol{X}'_i(\text{temp})$ & $\boldsymbol{X}''_i(\text{temp})$ to the nearest point on the corresponding epipolar line to generate $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$, which are updated estimates of $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$. This step ensures that $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$ conform to the multi-view geometry. We then align $\widetilde{\boldsymbol{X}}'_i$ to $\overline{\boldsymbol{X}}'$ & $\widetilde{\boldsymbol{X}}''_i$ to $\overline{\boldsymbol{X}}''$, as in [1], via a further affine transformation applied to the landmark points of each image. This is done in order to determine all the degrees of freedom in the alignment process. However, now we do not stop at this point, but complete the alignment by enforcing a shape correspondence between the two sets of pose-aligned points $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$, as described in Sect. 3.

3 Enforcing a Shape Correspondence between Views

Suppose now that we are using the 2D-PA algorithm to align points in an image $\boldsymbol{X}_i$ to the points, $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$, in two given mean basis views. The 2D-PA algorithm would generate two sets of points $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$ as explained in Sect. 2.2 above, which are aligned as well as possible. Since $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$ may not have the same shape, in order to complete the alignment, we need to update the shapes represented by each of them until they can be considered as simultaneous images of the same

3D object (i.e. we need to enforce shape correspondence between $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$). Thus, our aim, here, is to recover $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$, given $\overline{\boldsymbol{X}}'$, $\overline{\boldsymbol{X}}''$, $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$.

We begin by setting $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$ as our first estimates of $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$, respectively. Next, we use *Algorithm 13.1, on page 340* of [11], to compute the maximum likelihood estimate of the (affine) fundamental matrix that maps $\boldsymbol{X}'_i$ to $\overline{\boldsymbol{X}}''$. Here again we use only the stable points in order to compute an estimate of the fundamental matrix, since a rigid-object is assumed in the algorithm used. We then use the fundamental matrix and ***all*** the landmark points (not just the stable points) to map the shape of $\boldsymbol{X}'_i$ to $\overline{\boldsymbol{X}}''$ and generate $\hat{\boldsymbol{X}}''_i$. This shape transfer is achieved, as explained in the 2D-PA process, by moving each landmark point to the nearest point along the corresponding epipolar line (see Sect. 2.2). $\hat{\boldsymbol{X}}''_i$ corresponds to an image of the shape represented by points $\boldsymbol{X}'_i$ as seen from the view of $\overline{\boldsymbol{X}}''$. We do the same for the pair of images $\boldsymbol{X}''_i$ & $\overline{\boldsymbol{X}}'$ to generate $\hat{\boldsymbol{X}}'_i$ and update our estimates of $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$ as follows:

$$\boldsymbol{X}'_i \to \frac{1}{2}(\boldsymbol{X}'_i + \hat{\boldsymbol{X}}'_i) \quad \& \quad \boldsymbol{X}''_i \to \frac{1}{2}(\boldsymbol{X}''_i + \hat{\boldsymbol{X}}''_i) \ . \tag{3}$$

We then iterate, using our current estimates of $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$, to re-compute $\hat{\boldsymbol{X}}''_i$ & $\hat{\boldsymbol{X}}'_i$ and using $\hat{\boldsymbol{X}}''_i$ & $\hat{\boldsymbol{X}}'_i$ to update our estimates of $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$ via (3). We continue iterating until the difference between consecutive estimates of $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$ is smaller than some tolerance.

3.1 The Initial Reference Images

At this point we recall that in order to begin the 2D-PA algorithm, we require landmark points, $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$, in the two mean basis views. Thus, initially, we generate all the distinct combinations of landmark points from image pairs in the training set. We consider each pair of images, enforce a shape correspondence between them (as explained next), and compute the error e^2 defined in (1) corresponding to the selected pair. We then select the two images that produce the minimum value for e^2. The two images (with the shape correspondence enforced) thus selected, are then used as the initial reference images $\boldsymbol{X}_{\mathrm{r1}}$ & $\boldsymbol{X}_{\mathrm{r2}}$ in the Extended Procrustes Alignment (EPA) algorithm [1] to compute the points in the mean basis views. The EPA algorithm begins by considering $\boldsymbol{X}_{\mathrm{r1}}$ & $\boldsymbol{X}_{\mathrm{r2}}$ as the first estimates of $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$. Then we iterate, aligning points in all the training images (via the 2D-PA process) to the current estimates of $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$ and re-computing $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$ from the aligned sets of images, until convergence. Thus, until $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$ are computed, we use their current estimates instead.

During the process of selecting the two initial reference images for the EPA algorithm (see Sect. 3.1), however, we only have two sets of points that correspond to $\widetilde{\boldsymbol{X}}'_i$ & $\widetilde{\boldsymbol{X}}''_i$, since $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$ have not yet been computed. Therefore, in order to enforce a shape correspondence between these two sets of points, in each iteration we use the current estimates of $\boldsymbol{X}'_i$ & $\boldsymbol{X}''_i$ in place of $\overline{\boldsymbol{X}}'$ & $\overline{\boldsymbol{X}}''$.

4 The Integrated Shape and Pose Model (ISPM)

The ISPM, first introduced in [1], utilizes landmark points representing two flexible basis views to integrate the LCV technique with a CVFSM [2]. To build an ISPM, from (almost) any given set of images, we first select two reference images that define the basis views (see Sect. 3.1) and use the EPA algorithm [1] to simultaneously compute the points representing the two mean basis views and align all the training images to them (via the 2D-PA algorithm). This results, for each training image, in two corresponding sets of landmark points that represent simultaneous images of the object of interest taken from the two selected basis views while the object changes only its shape. Thus, we can then build two FSMs to model the intrinsic shape variation present in these two sets of landmark points and use the correspondence to build a hierarchical CVFSM. The parameters of the CVFSM (the shape parameters) enable us simultaneously to change the shape of the object in the two basis view images in a corresponding manner. Given the points representing the two basis views of the object with a particular shape, we may use the reformulated LCV technique [1] to synthesize that shape as seen from any desired view point via an appropriate CATT [4]. The elements of the CATT are the pose parameters. To use the ISPM, given the landmark points in a new image of the object, we first align them, via the 2D-PA algorithm, to the points representing the mean basis views. This provides: (*i*) the CATT that defines the pose of the object in the image (i.e. the pose parameters) and, (*ii*) the input set of landmark points to each of the two individual FSMs. The parameters of the individual FSMs then provide the input to the CVFSM, from which we extract the shape parameters in the usual way [2,3].

5 Evaluation

As a proof-of-principle the updated version of the ISPM detailed in this paper was evaluated in comparison to its predecessor [1] and a conventional Flexible Shape Model (FSM) as built by Cootes et al. [3]. The evaluation was carried out on landmark points selected from both real and synthetic image data. For the real images, we used landmark points manually selected from the same data as in [1], with five expressions (Neutral, Angry, Happy, Sad & Surprised) sampled at 13 different poses ($\sim 5^\circ$ intervals from $\sim -30^\circ$ to $\sim +30^\circ$ where 0° corresponds to the frontal view) giving 65 sets of points in total. For the synthetic data, we utilized the 3D head model of Loizides et al. [12] to generate images of a face. A subset of the 3D model points was manually selected as landmarks. The error-free locations of these landmark points in the corresponding synthetic images, and the images themselves were computed via an affine projection matrix. Four expressions (Fear, Happiness, Sadness & Neutral) were sampled at 11 different poses (at 5° intervals from -25° to $+25^\circ$) to generate 44 sets of points in total. In both cases (real and synthetic), rotations were performed only about the vertical axis. Owing to the space limit here, we refer the reader to [13] for a complete description of the databases. Some example images and the landmark points used in each case are shown in Fig. 1.

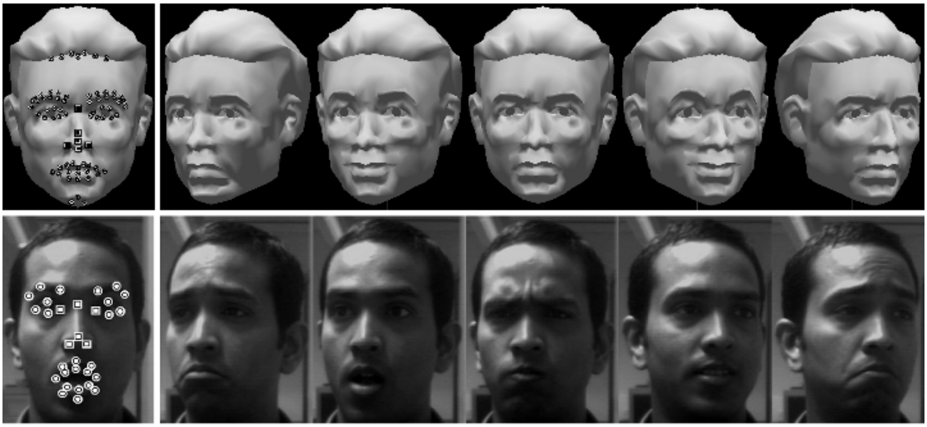
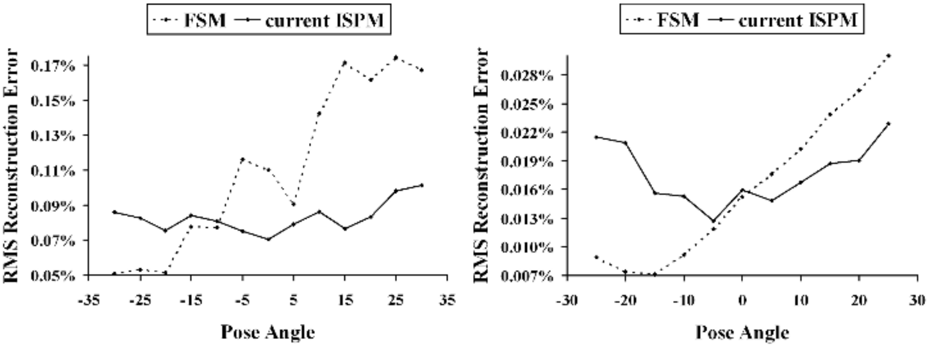


Fig. 1. Some examples of the real (bottom) and synthetic (top) images used in the evaluation. The subset of stable points (squares) and the other landmark points (circles) used in each case are shown in the leftmost images.

We evaluated the performance of each model (i.e. of the FSM, the initial ISPM [1] and the current ISPM described in this paper) by its ability to reconstruct the point configuration in a given image. For this we used each model to extract its own representation of a given image and use this representation to reconstruct the points representing the original image. The reconstruction error was then computed to be the root mean square error between the positions of the landmark points in the original image and the points reconstructed by the model. We represent this error as a percentage of the scale of the original image in order to make it scale invariant. The reconstruction errors were also averaged over expression, providing an error measure as a function of pose and independent of expression.

We performed cross-validation [14] and leave-one-out experiments on the two data sets in order to determine the accuracy of each model. The results of the leave-one-out experiments are shown in Fig. 2. The cross-validation experiments produced similar results. The graphs in Fig. 2 clearly show that the FSM is dependent on pose whereas the current ISPM isn't. Furthermore, in all experiments, on average the current ISPM out-performed the FSM. The minimum error of the initial ISPM was, however, much larger (always $> 1.0\%$) and is therefore not shown in the graphs in Fig. 2. Thus, although the initial ISPM was pose-invariant, it wasn't able to rival the FSM in terms of accuracy as the current ISPM does. Since similar results were generated on the real and synthetic databases, which were completely different in size, shape, number of landmark points and noise level, we are confident they reflect the performance of the models of interest and not some peculiarity of a particular database.

We also evaluated the performance of our algorithm that enforces a shape correspondence between two views (see Sect. 3) by computing the Pearson Cor-



(a)

(b)

Fig. 2. The reconstruction errors from the leave-one-out experiments on the (a) real & (b) synthetic data for the FSM (dotted line) and the current ISPM (solid line).

relation Coefficient (PCC) between the eigenvectors, eigenvalues and the scatter matrices of the two individual FSMs that were built from the pose-aligned images. We use the absolute value of the PCC which is, by definition, between 0 (no apparent correlation) and 1 (highly correlated). In all experiments the PCC values for the current ISPM were above 0.9, which shows that a shape correspondence was successfully enforced. Except for the PCC values for the eigenvalues (0.9) and of the first eigenvector (0.8), all the other PCC values for the initial ISPM were below 0.7. Since space is limited here, we refer the reader to [13] for more details on our results.

6 Conclusions and Future Work

We have developed an algorithm capable of enforcing a shape correspondence between two views of the same 3D non-rigid object in different shape-states. We have used this algorithm, along with many other significant updates, to improve the version of the Integrated Shape and Pose Model (ISPM) described in [1] by a factor of 10. As a proof-of-principle we have evaluated the performance of the improved ISPM in comparison to that of its predecessor [1] and of the conventional FSM [3], on two different databases via cross-validation and leave-one-out experiments. The results show that, unlike the FSM, the current ISPM is view-invariant since we separate the extrinsic (pose) variations from the intrinsic (shape) variations. Furthermore, on average the current ISPM described in this paper out-performs the FSM, while also completely out-classing the initial ISPM. The algorithm that enforces a shape correspondence between two views was also evaluated and shown to be successful. We anticipate that the ISPM will be useful in a variety of applications including calculation of head pose and view-invariant expression recognition. The approach may also be of relevance to theories of human vision.

Acknowledgements. We thank Andreas Loizides for providing the 3D model used in [12]. We also thank Matthew Trotter for his helpful comments and suggestions on evaluating our model. Benjamin Dias is funded by an *ORS award*[1], by financial support from *Unilever Research Colworth*[2] and by a *UCL Graduate Open Scholarship.*

References

1. Dias, M.B., Buxton, B.F.: Integrated shape and pose modelling. In: Proc. BMVC. Volume 2. (2002) 827–836
2. Cootes, T.F., Wheeler, G.V., Walker, K.N., Taylor, C.J.: Coupled-view active appearance models. In: Proc. BMVC. Volume 1. (2000) 52–61
3. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: A trainable method of parametric shape description. In: Image and Vision Computing. Volume 10. (1992) 289–294
4. Bretzner, L., Lindeberg, T.: Use your hands as a 3-d mouse, or, relative orientation from extended sequences of sparse point and line correspondences using the affine trifocal tensor. In: Proc. 5^{th} ECCV, LNCS. Volume 1406., Springer Verlag, Berlin (1998) 141–157
5. Tarr, M.J., Williams, P., Hayward, W.G., Gauthier, I.: Three-dimensional object recognition is viewpoint dependent. In: Nature Neuroscience. Volume 1. (1998) 275–277
6. Ullman, S., Basri, R.: Recognition by linear combinations of models. In: IEEE Transactions on PAMI. Volume 13. (1991) 992–1006
7. Pollard, S., Pilu, M., Hayes, S., Lorusso, A.: View synthesis by trinocular edge matching and transfer. In: Proc. BMVC. Volume 2. (1998) 770–779
8. Koufakis, I., Buxton, B.F.: Very low bit-rate face video compression using linear combination of 2d face views and principal component analysis. Image and Vision Computing **17** (1998) 1031–1051
9. Thórhallsson, T., Murray, D.W.: The tensors of three affine views. In: Proc. International Conference on Computer Vision and Pattern Recognition. Volume 1. (1999) 450–456
10. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In: Communications of the ACM. Volume 24. (1981) 381–395
11. Hartley, R., Zisserman, A.: Multiple View Geometry In Computer Vision. Cambridge University Press, ISBN 0-521-62304-9 (2000)
12. Loizides, A., Slater, M., Langdon, W.B.: Measuring facial emotional expressions using genetic programming. In: Soft Computing and Industry Recent Applications. (2001) 545–554
13. Dias, M.B., Buxton, B.F.: Advances in integrated shape and pose modelling. Technical Report RN/03/10, University College London (UCL), Department of Computer Science (2003)
14. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: Proc. IJCAI. (1995)

[1] Granted by the CVCP, Woburn House, London WC1H 9HQ, UK.

[2] Colworth House, Sharnbrook, Bedford MK44 1LQ, UK.