Machine Vision: Structure from Motion on Face Videos

Candidate: 020845

1. INTRODUCTION

This report describes the implementation of a simple yet robust structure from motion (SFM) method originally proposed by Tomasi and Kanade [1]. The data used for these experiments comprises of four sets of images taken from videos of people's faces. The challenge is to produce a 3D model and subsequently be able to project a 3D representation of each person's face as in the image, using the data provided. The SFM method described makes the assumption that the faces being modelled are rigid and does not take into account any deformation such as mouths opening and closing. On initial inspection of the images (figure 1) it is apparent that the camera is stationary, and any movement appears to come from the models. It is expected that this method will obtain different levels of accuracy between each set, due to each set having different movement and deformation characteristics. For example, the first set of images show a man's face, close to the camera, as he turns his head to allow multiple angles to be captured. It is expected that this will produce an accurate model as his head is consistently kept close to the centre of each image, which will reduce noise, he turns his head slowly and fully; which enables a decent amount of structural information to be obtained, and he does not open his mouth, deformation is kept to a minimum. In contrast, the model in the last set of images is relatively stationary and therefore minimal structural information may be obtained. Image set two shows a lady who appears to be giving a speech. She is relatively still but turns at some point during the speech, so each side of her face is captured. The third set of images appears to be of another lady giving a speech but, in this case, she is moving around quite a lot. Furthermore, the way she is speaking is somewhat expressive leading her to open her mouth widely at certain points, change her facial expression as well as close her eyes occasionally.

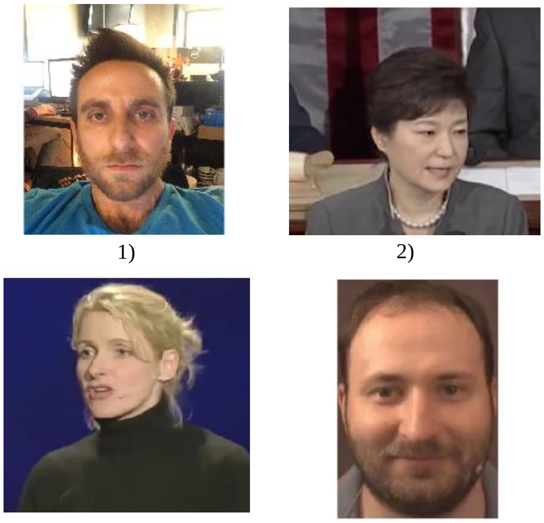
1) 2) 3) 4)

Figure 1. The first images from Image sets 1, 2, 3 and 4

2. METHOD

The method follows that outlined by Tomasi and Kanade [1]. The measurement matrix W is obtained by sorting the x and y coordinates of the landmark features, for each image in a set, into a single matrix:

$$W = \begin{matrix} x_{11} & x_{12} & \cdots & x_{1P} \\ y_{11} & y_{12} & \cdots & y_{1P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{F1} & x_{F2} & \cdots & x_{FP} \\ y_{F1} & y_{F2} & \cdots & y_{FP} \end{matrix}$$

W is centred by subtracting the row wise means. These will be used to construct a translation matrix later. W is decomposed by singular value decomposition (SVD):

$$W = U\Sigma V^T,$$

to obtain a rotation matrix $\hat{R}$ and a shape matrix $\hat{S}$:

$$\hat{R} = U\Sigma^{\frac{1}{2}}, \qquad \hat{S} = \Sigma^{\frac{1}{2}}V^T$$

A rank of 3 is enforced by making a 3×3 diagonal matrix Σ using the 3 largest singular values. The singular values of this matrix are square rooted. $\hat{R}$ is constructed by dotting the first 3 columns of the left singular vector matrix U with $\Sigma^{1/2}$. $\hat{S}$ is obtained by $\Sigma^{1/2} \cdot V^T$ where V^T is the first three rows of the matrix containing the right singular vectors of W. The next objective is to find a true solution for R and S and to achieve this we must find a Q such that:

$$R = \hat{R}Q, \quad S = Q^{-1}\hat{S} \tag{1}$$

This is achieved by linear least squares. R should contain orthogonal unit vectors $\hat{\imath}$ and $\hat{\jmath}$ so for each frame f we set up the metric constraints in the manner:

$$\hat{\imath}_f^T\hat{\imath}_f = 1 \Rightarrow \hat{\imath}_f^T QQ^T\hat{\imath}_f = 1$$

$$\hat{\imath}_f^T\hat{\imath}_f = 1 \Rightarrow \hat{\jmath}_f^T QQ^T\hat{\jmath}_f = 1$$

$$\hat{\imath}_f^T\hat{\jmath}_f = 0 \Rightarrow \hat{\imath}_f^T QQ^T\hat{\jmath}_f = 0$$

We construct a gram matrix A of dimensions 3F×6 as described by Morita and Kanade [2]. For each frame f we have a row-double in $\hat{R}$:

$$\hat{R}_f = \begin{matrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \end{matrix}$$

we use these to construct each triplet of rows in A. For this we use a function *compute_row* which takes 2 vectors of length 3 and outputs a 6 membered vector to produce a single row such that:

$$\begin{aligned} compute\_row(a, b) = [\, & a_1b_1, \quad a_1b_2 + a_2b_1, \quad a_1b_3 + a_3b_1, \\ & a_2b_2, \quad a_2b_3 + a_3b_2, \quad a_3b_3 \,] \end{aligned}$$

Three rows in each A_f will be constructed from the rows of $\hat{R}_f$ in the manner *compute_row(**x**, **x**)*, *compute_row(**y**, **y**)* and *compute_row(**x**, **y**)*. The constraints are formed into a column vector b of dimensions 3F×1:

$$b = [1, 1, 0, 1, 1, 0, \ldots 1, 1, 0]^T$$

we can then solve for g the system of linear equations: $Ag = b$.

The solution is a 6 membered vector $g = [g_1, g_2, \ldots g_6]^T$ which is then used to construct the 3×3 matrix G:

$$G = \begin{bmatrix} g_1 & g_2 & g_3 \\ g_2 & g_4 & g_5 \\ g_3 & g_5 & g_6 \end{bmatrix}$$

This matrix is factorized by eigen decomposition to yield the right eigen vector matrix D and 3×3 eigen value matrix Λ. Q is then derived by:

$$Q = D\Lambda^{\frac{1}{2}}$$

With Q, R and S are derived by equation 1. A third row must be derived for each R_f by taking the cross product of the first two rows, producing a final constructed R of dimensions 3F×3. A translation vector is constructed by taking the row-wise means of W, calculated earlier, which will be of dimension 2F×1, and padding each pair of values with 0 so that it is of dimension 3F×1. We have now obtained the structure S, the rotation matrix for each image R, and the translation vector T. The 3D projection X of an image f may be derived by:

$$X_f = R_f S + T_f$$

Sometimes there may be bas-relief ambiguity (BRA) in the 3D projections. This is detected by making the assumption that the Z coordinate of the nose tip should be lower than the median Z coordinate in an image. If this is not so, to solve this ambiguity we first multiply the Z coordinates in S, which are in the third row, by -1. We then iterate through R for each R_f and multiply the third column of the first two rows by -1 and recompute the third row as the cross product of the first two rows multiplied by -1.

3. RESULTS AND DISCUSSION

This structure from motion implementation had varying levels of success between the different image sets. Though, when successful an accurate 3D representation of the person's face was produced which was angled facing the same direction as in the original image along the Z axis (figure 2). Rotating the projection shows that the method has managed to capture a good amount of the structural information. Although, the model in image set 1 maintained the rigidity of their face throughout the image set and so this may have been an easy case. The model in image set 3 appeared to be giving a presentation or speech and so there are frames where their mouth is open, and their expressions change often. It can be seen in figure 3 that while the projections manage to match the orientation of the face and general structure, they do not match the exact structure in every image. The 3D projection does not match the women's mouth as in the original image. The matrix S is like an average of the structure taken over all input images so a projection will be of the average structure of the object projected to match the original orientation. The bottom row of figure 3 shows an instance where a small selection of 5 images from different angles of the model with similar open-mouthed expressions. As the new average structure between the selected images has an open mouth, the new projection also has an open mouth and better matches the original image.

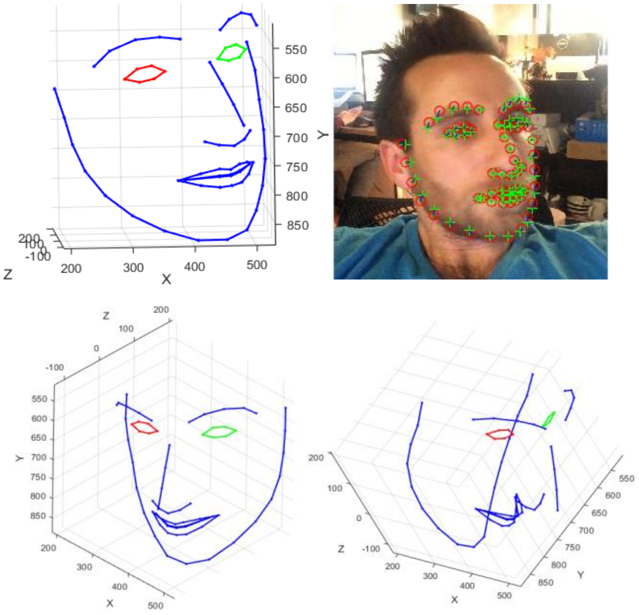

Figure 2. Visualizations for image 42 of set 1. **Top Right**: 2D landmarks (red circles) and projected points (green crosses). **Top Left**: front view of projected 3D points to match image view. **Bottom**: projected 3D points from different angles.

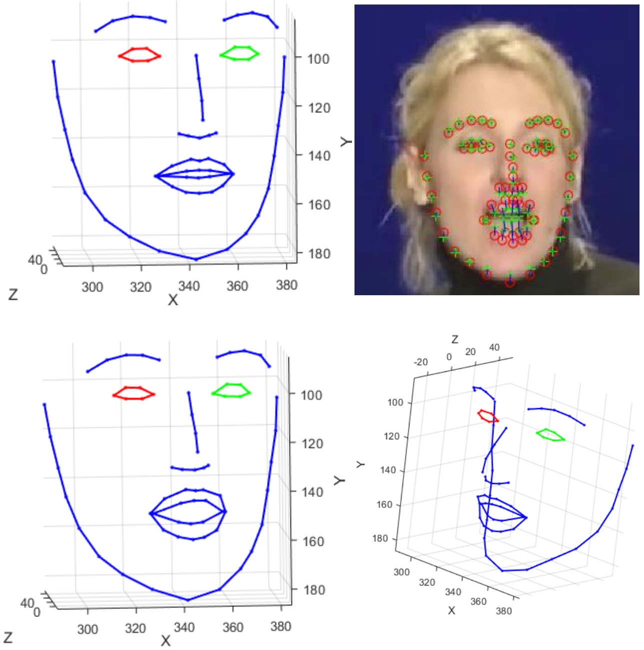

Figure 3. Visualizations for image 32 of set 3. **Top Right**: 2D landmarks (red circles) and projected points (green crosses) from using all image data. **Top Left**: 3D projection of front view using all image data. **Bottom Row**: 3D projections using a selection of image data.

Image set 2 suffered from BRA which was resolved effectively. It can be seen in figure 1 that the 3D projection, before processing, was orientated incorrectly and after processing was performed, the correct orientation was obtained. This may be related to the images of this set mostly showing only two views. Most of the input information is therefore heavily weighted from these two angles and there is not much in between. Applying the same method to some subsets of the images it was noticed that often there was no BRA occurring. Another point is that resolving the BRA should not affect the *x* or *y* coordinates, which is a good indicator that it has been implemented correctly. It can be seen in the top row of figure 4 that the pre-processed and post-possessed *x* and *y* coordinates between the two images match.

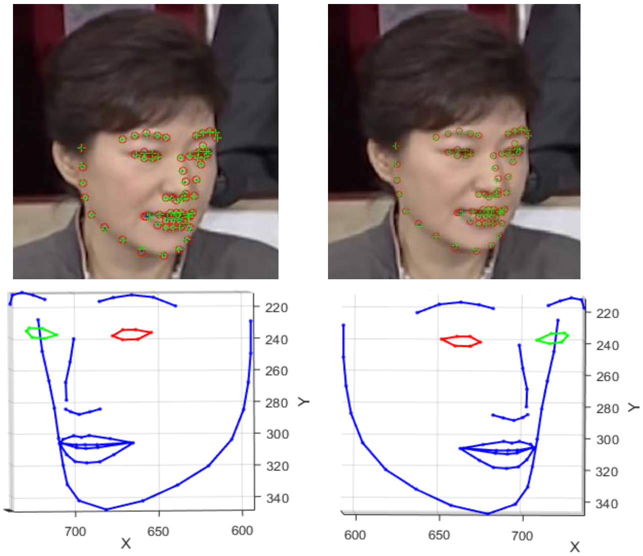

Figure 4. Visualizations for image 56 of set 2. **Top Row**: 2D landmarks (red circles) and projected points (green crosses). **Bottom row**: 3D projections. **Left column**: Pre-processed, showing bas-relief ambiguity. **Right column**: Post-processed, with ambiguity resolved.

Image set 4 produced the worst results. When using the entire image set, the 3D projections overestimate the magnitude along the Z axis. The face which is produced has too much depth and is barely recognizable as a face from some angles. Image set 4 does not have much motion and the face is not captured from any other angles other than head-on. However, the face does change shape as the model changes their facial expressions throughout the image set. The top row of figure 5 depicts the entirety of the movement in the set, which is just the transitions between 3 facial expressions. This SFM method sees movement but cannot distinguish this from rigid structural information and ends up modelling structure where there is no structure. Although, the woman in image set 3 also changes their facial expressions, and their face changes shape, the deformity effect is counteracted somewhat by their rotation. Enough true structural information is given by the images that significant noise from deformity is reduced. It is possible to filter out images which have this kind of noise, although, for image set 4 the most significant problem is that the model's face is not captured from more than one angle. This method is not appropriate for creating a 3D model from this image set.

4. CONCLUSION AND FURTHER WORK

This experiment has shown that it is possible to create a 3D projection of a single image, which appears to be a relatively accurate estimate, from a set of images from different angles. Although, we have not quantified any of the projection accuracies so any comments on accuracy are arbitrary. It would be necessary to quantify this in the future as to make comparisons of performance. Furthermore, it has been observed that this particular SFM method is vulnerable to noise from structural deformity, though this can be negated to an extent by providing enough rigid structural information i.e. from images of multiple angles. Sometimes it will be necessary to select only appropriate images if the object being modelled is not a solid structure. Another aspect to investigate would be how to detect outliers. The matrix *G* is determined by a linear least squares method which can be sensitive to outliers. Removing certain images based upon a calculated metric would validate their removal. The metric would have to be based upon the between-landmark distances as to detect deformities and not the movement of the camera or rotation of the object being captured.

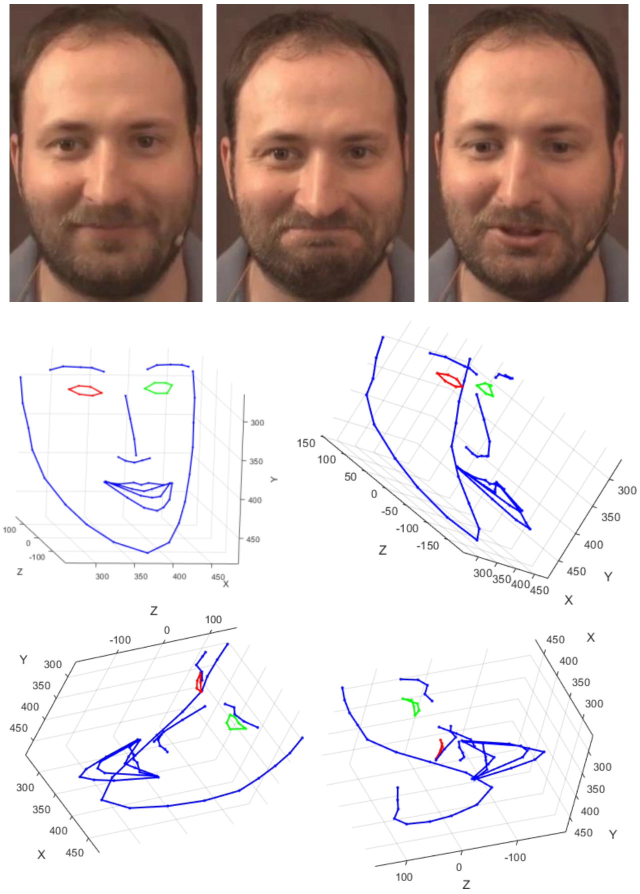

Figure 5. Images and projections of image set 4. **Top Row**: from left to right: image 1, 152, and 286. **Middle And Bottom Rows**: 3D projections of image 1.

5. REFERENCES

[1] Tomasi, C. and Kanade, T., 1992. Shape and Motion from Image Streams under Orthography: a Factorization Method. *International Journal of Computer Vision 9*, 2, 137-154.

[2] Morita, T. and Kanade, T., 1997. A Sequential Factorization Method for Recovering Shape and Motion from Image Streams. *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE 19*, 8, 858-867.