AN ALGORITHM FOR 3D RECONSTRUCTION OF DEFORMABLE SHAPE SEQUENCES

Amit K. Roy-Chowdhury

Department of Electrical Engineering
University of California, Riverside, CA 92521

ABSTRACT

In this paper we present an algorithm for estimating the 3D model of a deformable shape from a video sequence. Our method assumes that a deformable shape sequence can be represented by a linear combination of basis shapes, where the weights assigned to each basis shape change with time. While there is existing work on estimating the basis shapes and their combination coefficients, they lack the crucial information about the number of basis shapes that are required for the model. This is usually determined through heuristics about the physics of the underlying structure. We show that it is possible to estimate the number of basis shapes from the tracked points obtained from the video sequence, using a scaled orthographic camera projection model. This estimate is then used to compute the 3D structure of each of the basis shapes. We present experimental results in recreating the structure of the human body during various activities from a video sequence.

1. INTRODUCTION

Estimating the 3D structure of an object is a classical problem in video analysis. Some of the recent efforts in this area has concentrated on estimating the shape of a deformable object [1]. One of the methods that have been proposed for this problem is to represent the deforming shape as a linear combination of certain basis shapes, where the weights assigned to each basis shape changes with time, thus resulting in the sequence of deforming shapes. However, the crucial information that is lacking in this approach is an estimate of the number of basis shapes that can effectively represent the deforming sequence. This has often been decided based on heuristics about the physics of the underlying structure. In this paper, we show that it is possible to estimate the number of basis shapes from the set of feature points on the object tracked over all the frames. This is followed by reconstruction of the basis shapes. We assume a scaled orthographic projection model for the camera. Experiments are shown on reconstructing the structure of the human body in different activities.

Some of the commonly used representation of shape are Fourier descriptors, extended Gaussian images, splines and deformable snakes, all of which model the shape of continuous curves. Active shape models [2] and Kendall's statistical shape theory [3] have considered the shape of a discrete set of points. Methods for deformation of one shape into another and for comparing the similarity of two shapes have been proposed in [3, 4]. However, there has been very little work on shape sequence processing. Some recent work in this area involves shape-dynamical models for activity [5] and human motion analysis [6] and for image synthesis [7]. In the domain of 3D shape estimation from 2D images, the factorization theorem is one well-known approach, though it is usually applied under the assumption of a scaled orthographic camera projection model [8]. Its extension to modeling deformable shapes was proposed recently in [1], by approximating a non-rigid object by a composition of basis shapes, thus limiting the rank of the measurement matrix of the entire image sequence. In this paper, we propose a non-iterative method for estimating the number of basis shapes that can effectively model a deforming shape and then use this estimate to reconstruct the shape itself.

The basic input to our estimation procedure is the set of tracked feature points (i.e. the trajectories) of the object over all the frames. These trajectories are transformed to a 3D shape space using the ideas of deformable shape modeling in [1]. Using statistical models to separate out the "true" deformations from those induced by noise in the trajectories, the dimension of this shape space is estimated using tools from linear algebra. The number of basis shapes required is determined by the dimension of this shape space. A complete algorithm for modeling of deformable shapes is presented.

2. MODELING ALGORITHM

We outline the theory for modeling of deformable shapes from video sequences. For ease of explanation, we proceed in the reverse order. We first explain how to estimate the basis shapes (based on the work in [1]), assuming the number of basis shapes, K is known. We then show how to estimate K. Finally, we present our reconstruction algorithm.

2.1. Estimating 3D Basis Shapes

We hypothesize that each shape sequence can be represented by a linear combination of 3D basis shapes. Mathematically, if we consider the trajectories of P points representing the shape (e.g. landmark points), then the overall configuration of the P points is represented as a linear combination of the basis shapes as

$$S = \sum_{i=1}^{K} l_i S_i, \qquad S, S_i \in \Re^{3\times P}, l_i \in \Re. \tag{1}$$

The choice of K determines the deformability of the shape sequence and is the focus of the derivation that follows. We will assume a scaled orthographic projection model for the camera.

A number of methods exist in the computer vision literature for estimating the basis shapes. In [8], the authors considered P points tracked across F frames in order to obtain two $F \times P$ matrices $\mathbf{U}$ and $\mathbf{V}$. Each row of $\mathbf{U}$ contains the x-displacements of all the P points for a specific time frame, and each row of $\mathbf{V}$ contains the corresponding y-displacements. It was shown in [8], that for 3D rigid motion under orthographic camera model, the rank, r, of $\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ has an upper bound of 3. The rank constraint is derived from the fact that $\begin{bmatrix}\mathbf{U}\\ \mathbf{V}\end{bmatrix}$ can be factored into two matrices $\mathbf{M}_{2F\times r}$ and $\mathbf{S}_{r\times P}$, corresponding to the pose and 3D structure of the scene, respectively. In [1], it was shown that for non-rigid motion, the above method could be extended to obtain a similar rank constraint, but one that is higher than the bound for the rigid case. We will adopt the last mentioned method for computing the basis shapes. We will outline the basic steps of their approach in order to clarify the notation for the remainder of the paper.

Given F frames of a video sequence with P moving points, we can obtain the trajectories of all these points over the entire video sequence. These P points can be represented in a measurement matrix as

$$\mathbf{W}_{2F\times P} = \begin{bmatrix} u_{1,1} & \cdots & u_{1,P} \\ v_{1,1} & \cdots & v_{1,P} \\ \vdots & \vdots & \vdots \\ u_{F,1} & \cdots & u_{F,P} \\ v_{F,1} & \cdots & v_{F,P} \end{bmatrix}, \tag{2}$$

where $u_{f,p}$ represents the x-position of the p^{th} point in the f^{th} frame and $v_{m,p}$ represents the y-position of the same point. Under weak perspective projection, the P points of a configuration in a frame f, are projected onto 2D image points $(u_{f,i}, v_{f,i})$ as

$$\begin{bmatrix} u_{f,1} & \cdots & u_{f,P} \\ v_{f,1} & \cdots & v_{f,P} \end{bmatrix} = \mathbf{R}_f \left(\sum_{i=1}^{K} l_{f,i} S_i \right) + \mathbf{T}_f, \tag{3}$$

where,

$$\mathbf{R}_f = \begin{bmatrix} r_{f,1} & r_{f,2} & r_{f,3} \\ r_{f,4} & r_{f,5} & r_{f,6} \end{bmatrix} \triangleq \begin{bmatrix} \mathbf{R}_f^{(1)} \\ \mathbf{R}_f^{(2)} \end{bmatrix}. \tag{4}$$

$\mathbf{R}_f$ represents the first two rows of the full 3D camera rotation matrix and $\mathbf{T}_f$ is the camera translation. The translation component can be eliminated by subtracting out the mean of all the 2D points (assuming that they are seen in all views), as in [8]. We now form the measurement matrix $\mathbf{W}$, which was represented in (2), with the means of each of the rows subtracted. The weak perspective scaling factor is implicitly coded in the configuration weights, $\{l_{f,i}\}$.

Using (2) and (3), it is easy to show that

$$\begin{aligned} \mathbf{W} &= \begin{bmatrix} l_{1,1}\mathbf{R}_1 & \cdots & l_{1,K}\mathbf{R}_1 \\ l_{2,1}\mathbf{R}_2 & \cdots & l_{2,K}\mathbf{R}_2 \\ \vdots & \vdots & \vdots \\ l_{F,1}\mathbf{R}_F & \cdots & l_{F,K}\mathbf{R}_F \end{bmatrix} \begin{bmatrix} S_1 \\ S_2 \\ \vdots \\ S_K \end{bmatrix} \\ &= \mathbf{Q}_{2F\times 3K}.\mathbf{B}_{3K\times P}, \end{aligned} \tag{5}$$

which is of rank $3K$. The matrix $\mathbf{Q}$ contains the pose for each frame of the video sequence and the weights $l_1, ..., l_K$. The matrix $\mathbf{B}$ contains the basis shapes corresponding to each of the activities. In [1], it was shown that $\mathbf{Q}$ and $\mathbf{B}$ can be obtained using singular value decomposition (SVD), and retaining the top $3K$ singular values, as $\mathbf{W}_{2M\times P} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ and $\mathbf{Q} = \mathbf{U}\mathbf{D}^{\frac{1}{2}}$ and $\mathbf{B} = \mathbf{D}^{\frac{1}{2}}\mathbf{V}^T$.

2.2. Estimating the Number of Basis Shapes

The above mentioned rank constraint requires knowledge of K in order to estimate the shape and motion parameters. This is usually determined heuristically from the physics of the object whose structure is being estimated. We now provide a theoretical method for estimating K by reinterpreting the above equations in stochastic framework.

Consider the set of coordinates representing the shape of the deformable object in a particular frame of a video sequence to be the realization of a random process. The sequence of frames depicts the deformation of the shape, along with the effects of the 3D translation and rotation. Represent the x and y coordinates of the sampled points in a single frame as a vector $\mathbf{y} = [u_1, ..., u_P, v_1, ..., v_P]^T$. Then, from (5), it is easy to show that for K basis shapes (K is

unknown)

$$\mathbf{y}^T = \left[l_1\mathbf{R}^{(1)}, ..., l_K\mathbf{R}^{(1)}, l_1\mathbf{R}^{(2)}, ..., l_K\mathbf{R}^{(2)}\right] * \begin{bmatrix} S_1 & \\ \vdots & 0 \\ S_k & \\ & S_1 \\ 0 & \vdots \\ & S_k \end{bmatrix} + \eta^T,$$

$$i.e., \quad \mathbf{y} = (\mathbf{q}_{1\times 6K}\mathbf{b}_{6K\times 2P})^T + \eta = \mathbf{b}^T\mathbf{q}^T + \eta, \tag{6}$$

where η represents the noise in the sequence of tracked points and is assumed to be a zero-mean random process. The vector $\mathbf{q}$ is obtained by juxtaposing two consecutive rows of $\mathbf{Q}$, corresponding to the same image frame, in equation (5). The matrix $\mathbf{b}$, which is constant across all the frames, is obtained by duplicating $\mathbf{B}$ in equation (5), as shown in equation (6).

Assuming that the coordinates of the points representing the shape in all the F frames can be considered to be realizations of the same random process, with possibly different noise variances, we can compute the correlation matrix of $\mathbf{y}$. Let $\mathbf{R_y} = E[\mathbf{y}\mathbf{y}^T]$ be the correlation matrix of $\mathbf{y}$ and $\mathbf{C}_\eta$ the covariance matrix of η. Hence,

$$\mathbf{R_y} = \mathbf{b}^T E[\mathbf{q}^T\mathbf{q}]\mathbf{b} + \mathbf{C}_\eta. \tag{7}$$

The correlation matrix, $\mathbf{R_y}$, is of size $2P \times 2P$ and can be estimated from the sequence of points representing the shapes as $\mathbf{R_y} = \frac{1}{F}\sum_{f=1}^{F}\mathbf{y}_f\mathbf{y}_f^T$, where $\mathbf{y}_f$ is the vector $\mathbf{y}$ (defined above) in the frame f. The expectation on the right hand side of equation (7) can be computed similarly as $E[\mathbf{q}^T\mathbf{q}] = \frac{1}{F}\sum_{f=1}^{F}\mathbf{q}_f^T\mathbf{q}_f$, where $\mathbf{q}_f$ is the vector $\mathbf{q}$ (defined above) for frame f and is obtained from the matrix $\mathbf{Q}$ in equation (5).

The noise covariance matrix, C_η, represents the accuracy with which the feature points are tracked and needs to be estimated from the image frames. Since η need not be an IID noise process, C_η will not necessarily have a diagonal structure (but it is symmetric and positive semi-definite). Now, consider the diagonalization of $\mathbf{C}_\eta = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^T$, where $\mathbf{\Lambda} = \text{diag}[\mathbf{\Lambda}_s, 0]$ and $\mathbf{\Lambda}_s$ is an $L \times L$ matrix of non-zero singular values of $\mathbf{\Lambda}$. Let $\mathbf{P}_s$ denote the orthonormal columns of $\mathbf{P}$ corresponding to the non-zero singular values. Therefore,

$$\mathbf{C}_\eta = \mathbf{P}_s\mathbf{\Lambda}_s\mathbf{P}_s^T. \tag{8}$$

Premultiplying equation (6) by $(\mathbf{P}_s\mathbf{\Lambda}_s^{\frac{1}{2}})^{-1}$, we see that (6) becomes

$$\tilde{\mathbf{y}} = \tilde{\mathbf{b}}^T\mathbf{q}^T + \tilde{\eta}, \tag{9}$$

where $\tilde{\mathbf{y}} = \mathbf{\Lambda}_s^{-\frac{1}{2}}\mathbf{P}_s^T\mathbf{y}$ is a $L \times 1$ vector, $\tilde{\mathbf{b}}^T = \mathbf{\Lambda}_s^{-\frac{1}{2}}\mathbf{P}_s^T\mathbf{b}^T$ is a $L \times 6K$ matrix and $\tilde{\eta} = \mathbf{\Lambda}_s^{-\frac{1}{2}}\mathbf{P}_s^T\eta$. It can be easily verified that the covariance of $\tilde{\eta}$ is an identity matrix $\mathbf{I}_{L\times L}$. This is known as the process of "whitening", whereby the noise process is transformed to be IID.

Representing by $\mathbf{R}_{\tilde{\mathbf{y}}}$ the correlation matrix of $\tilde{\mathbf{y}}$, it can be seen that

$$\mathbf{R}_{\tilde{\mathbf{y}}} = \tilde{\mathbf{b}}^T E[\mathbf{q}^T\mathbf{q}]\tilde{\mathbf{b}} + \mathbf{I} = \mathbf{\Delta} + \mathbf{I}, \tag{10}$$

where, for simplicity, $\mathbf{\Delta} \triangleq \tilde{\mathbf{b}}^T E[\mathbf{q}^T\mathbf{q}]\tilde{\mathbf{b}}$. Now, $\mathbf{R}_{\tilde{\mathbf{y}}}$ is of dimension $L \times L$, $\tilde{\mathbf{b}}^T$ is of rank $L \times 6K$ and $E[\mathbf{q}^T\mathbf{q}]$ is of rank $6K \times 6K$. Thus, $\mathbf{\Delta}$ has maximum rank $6K$, where K is the number of basis shapes. For a general 3D scene undergoing translation and rotation, the rank will be $6K$. This is based on the fact that if $\mathbf{A}_{m\times n} = \mathbf{F}_{m\times r}\mathbf{G}_{r\times n}$, then the Rank$(\mathbf{A}) \leq r$. Representing by $\mu_i(\mathbf{A})$ the i^{th} eigenvalue of the matrix $\mathbf{A}$, we see that

$$\begin{aligned} \mu_i(\mathbf{R}_{\tilde{\mathbf{y}}}) &= \mu_i(\mathbf{\Delta}) + 1, \quad \text{for } i = 1, ..., 6K, \text{ and} \\ \mu_i(\mathbf{R}_{\tilde{\mathbf{y}}}) &= 1, \quad \text{for } i = 6K+1, ..., L. \end{aligned} \tag{11}$$

Hence, there are $6K$ eigenvalues above 1. By counting the number of eigenvalues that are greater than 1 and dividing it by 6, we can obtain an estimate of K, which is the number of basis shapes required to represent the sequence of deforming points.

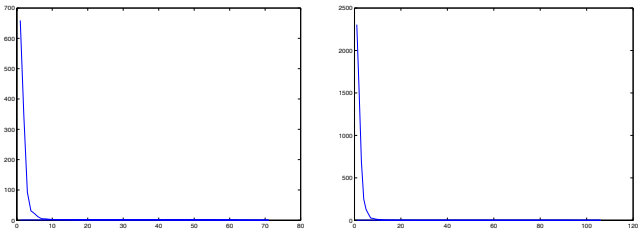
2.3. 3D Reconstruction Algorithm for Deformable Shapes

The complete 3D modeling algorithm for deformable shapes is as follows.

1. Compute the measurement matrix, $\mathbf{W}$, as in equation 5.
2. Estimate the number of basis shapes, K, according to the theory in Section 2.2.
3. Using the estimate of K, compute the 3D structure, combination coefficients and motion parameters of the basis shapes, according to Section 2.1.
4. Using equation (1), compute the sequence of deformable 3D shapes.

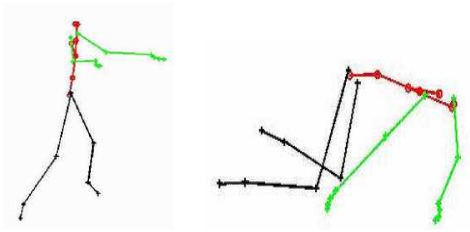
3. EXPERIMENTAL EVALUATION

We now present results on 3D reconstruction of deformable shapes using the algorithm presented in Section 2.3. We used the motion-capture data available from Credo Interactive Inc. and Carnegie Mellon University in the BioVision Hierarchy and Acclaim formats. It has a number of examples of different activities and is thus a rich dataset



(a) (b)

Fig. 1. Plot of the eigenvalues, in decreasing order of magnitude, for the (a) walk sequence and (b) crawl sequence.



(a) (b)

Fig. 2. Plots of the first basis shape, S_1, for walk and crawl sequences, respectively.

for studying shape sequences [1]. The combined dataset included a number of subjects performing various activities, like walking, jogging, sitting, crawling, brooming, etc. For each of these activities, we had multiple video sequences. Also, many of the activities contained video from different viewpoints. We will show here the results of the 3D reconstruction on two activities in this dataset, walking and crawling.

Using the video sequences and the theory outlined in Section 2.2, we estimated the number of basis shapes for each of the sequence. For the walk sequence, the number of eigenvalues greater than one was 35, which resulted in a value of K (= number of eigenvalues greater than one/6) to be 5.8. A plot of the eigenvalues, in decreasing order of magnitude, for the walk sequence is shown in Figure 1(a). We used six basis shapes for reconstructing the 3D model. For the crawl sequence, the number of eigenvalues over one was 48, and the value of K was 8. A plot of the eigenvalues, in decreasing order of magnitude, for the crawl sequence is shown in Figure 1(b).

The 3D model, combination coefficients and motion parameters were computed using the method described in Section 2.1. The first basis shapes are shown in Figure 2. Most of the information was contained in the first basis shape and it was the only one which was physically similar to the original shap. The remaining basis shapes, however, contributed to reducing the overall error in the reconstruction. We resynthesized the original sequences using the basis shapes and combination coefficients obtained from equation (5). Equation (3) was used for the synthesis. In both the cases, the error at none of the feature points was more than 1 pixel. If only the first basis shape was used for the synthesis, the error at some of the points in some of the frames was as high as five pixels. Using the basis shapes for computing similarity between various activities is an interesting problem in its own right that we have addressed in a separate paper.

[1] While there are a number of standard datasets for shapes, we could not find any large database for the study of shape sequences.

4. CONCLUSIONS

In this paper, we have presented an algorithm for 3D modeling of deformable shapes from video sequences. Our method assumes that a deformable shape can be represented as a linear combination of certain basis shapes, where the combination coefficients change with time. While there exist methods that had addressed this issue, they lacked the crucial input of the number of basis shapes required to model the deformable shape, which was usually chosen heuristically. We show that it is possible to estimate the number of basis shapes from the video sequence, followed by 3D modeling of these basis shapes. The procedure assumes a scaled orthographic camera projection model. We present results of 3D modeling on a few human activities, including details of the estimation of the number of basis shapes.

5. REFERENCES

[1] L. Torresani and C. Bregler, "Space-time tracking," in *European Conference on Computer Vision*, 2002.

[2] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active shape models: Their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, January 1995.

[3] I. Dryden and K. Mardia, *Statistical Shape Analysis*, John Wiley and Sons, 1998.

[4] W. Mio and A. Srivastava, "Elastic-string models for representation and analysis of planar shapes," in *Computer Vision and Pattern Recognition*, 2004.

[5] N. Vaswani, A. Roy-Chowdhury, and R. Chellappa, "Activity recognition using the dynamics of the confi guration of interacting objects," in *Computer Vision and Pattern Recognition*, 2003.

[6] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa, "Role of shape and kinematics in human movement analysis," in *Computer Vision and Pattern Recognition*, 2004.

[7] C.B. Liu and N. Ahuja, "A model for dynamic shape and its applications," in *Computer Vision and Pattern Recognition*, 2004, pp. II: 129–134.

[8] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: A factorization method," *International Journal of Computer Vision*, vol. 9, pp. 137–154, November 1992.