A FACTORIZATION METHOD IN STEREO MOTION FOR NON-RIGID OBJECTS

Yu Huang*, Jilin Tu**, Thomas S Huang**

*Thomson Corporate Research at Princeton, **University of Illinois at Urbana-Champaign
E-mail: yu.huang2@thomson.net, jilintu@ifp.uiuc.edu, huang@ifp.uiuc.edu

ABSTRACT

In this paper we propose a framework of factorization-based non-rigid shape modeling and tracking in stereo-motion. We construct a measurement matrix with the stereo-motion data captured from a stereo-rig. Organized in a particular way this matrix could be decomposed by Singular Value Decomposition (SVD) into the 3D basis shapes, their configuration weights, rigid motion and camera geometry. Accordingly, the stereo correspondences can be inferred from motion correspondences only requiring that a minimum of $3K$ point stereo correspondences (where K is the dimension of shape basis space) are created in advance. Basically this framework still keeps the property of rank constraints, meanwhile it owns other advantages such as simpler correspondence and accurate reconstruction even with short image sequences. Results with real data are given to demonstrate its performance.

Index Terms— Visual Tracking, Stereo Vision

1. INTRODUCTION

Tracking the object and recovering its 3D shape from sequences of images are fundamental problems in computer vision community. They have various applications such as scene modeling, robot navigation, object recognition and virtualized reality [1, 2, 6, 9]. Traditionally there exist two vision-based methods for 3-D reconstruction: visual *motion* and *stereo* vision. Both methods depend on how to solve the notorious *correspondence* problem. Basically this problem is relatively easy to handle in visual motion [12] because the extracted features have strong temporal association even without any prior knowledge of the dynamic model. Comparatively, stereo vision undergoes a much easier reconstruction task by triangulation, but the stereo correspondence task is severely ill-posed though we have the epipolar constraints.

In visual motion, Tomasi and Kanade [12] proposed one of the most influential approaches as the *factorization* method for *rigid* objects and orthographic projection. The key idea is decomposition of a measurement matrix into its shape and motion components. Various extensions have been put forward [7-8, 14]. Stemming from the rigid factorization method, a non-rigid factorization method was first proposed by Bregler et. al [13]. In the case of non-rigid factorization, the 3D shape is represented by a linear combination of basic modes of deformation. Brand proposed a flexible factorization approach which minimizes the deformations relative to the mean shape by introducing an optimal correction matrix [1]. Recently Xiao et.al proposed a new set of constraints on the shape basis in [15] and gave a close-form solution to non-rigid structure from motion.

Researchers have tackled this topic of augmenting "structure from motion" with stereo information. Some works are *feature*-based [4, 6], while others are called the "*direct*" method using the spatial and temporal image gradient information [10]. The notable problem is how to fully utilize the redundant information in the stereo-motion analysis, but practically the more important issue would be how to make the two basic cues complement with each other. Recently there are some stereo-motion papers taking into account *non-rigid* motion [2, 9, 3]. A basic primitive called dynamic-surfel which encodes the instantaneous local shape, reflectance and motion of a small region in the scene, is proposed in [2] to build the scene's structure in space-time from multiple views. Likewise, the object is modeled by a time-varying multi-resolution subdivision surface in [9], which is fitted to the image data from multiple views. It can be figured both methods above have to solve really complicated optimization problems. Only Del Bue et. al addressed non-rigid stereo motion by a factorization method [3], nevertheless stereo correspondence is assumed to be created and its focus was on shape recovery only.

In this paper, we will discuss 3D *non-rigid* shape recovery and tracking based on *factorization*. Our motivations come from the work in [5, 13]. Performing singular value decomposition (SVD) on the well-organized stereo-motion measurement matrix, we could factorize it into 3D basis shapes, their configuration weights, stereo geometry and rigid motion parameters. Moreover, we infer stereo correspondences from motion correspondences only requiring that at least $3K$ point stereo correspondences (where K is the dimension of shape basis space) are created initially. Basically this framework still owns the property of rank constraints [13]. It is an extension of [5]'s work to non-rigid objects, so such advantages as simpler correspondence and accurate reconstruction even with short sequences are preserved.

Sect. 2.1 reviews the factorization work for the non-rigid motion model in [13]. Our work as an extension to stereo-motion is described in Sect. 2.2. In Sect. 2.3 we discuss how to infer stereo correspondences. Sect. 3 provides our experiment results of real sequences.

2. STEREO-MOTION FACTORIZATION

2.1 Non-rigid Motion Model

The shape of the non-rigid object is described [13] as a key-frame basis set S_1, S_2, …, S_K. Each key-frame S_i is a 3xP matrix describing P points. The shape of a specific configuration S^t at the time frame t is a linear combination of the basis set:

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

Assume a weak-perspective model (scaled orthographic model) for the camera projection process. The 2D image points

$(u_{t,i}, v_{t,i})$ are related to 3D points of a configuration S^t at a specific time frame *t* by

$$\begin{bmatrix} u_{t,1} & \dots & u_{t,P} \\ v_{t,P} & \dots & v_{t,P} \end{bmatrix} = R'_t \left(\sum_{i=1}^{K} l_{t,i} S_i \right) + T'_t , \qquad (2)$$

$$R'_t = \begin{bmatrix} r_1 & r_2 & r_3 \\ r_4 & r_5 & r_6 \end{bmatrix}. \qquad (3)$$

where R'_t (2x3) contains the first two rows of the full 3-D rigid rotation matrix R_t, and T'_t is the 2-D rigid translational vector (it consists of the first two components of the 3-D translation vector T_t). The weak perspective scaling has been implicitly coded in $l'_{t,1}, \dots l'_{t,K}$. Actually we can eliminate T'_t by subtracting the mean of all 2D image points, and then can assume that S^t is centered at the origin. We can rewrite the linear combination in (2) as a matrix multiplication:

$$\begin{bmatrix} u_{t,1} & \dots & u_{t,P} \\ v_{t,P} & \dots & v_{t,P} \end{bmatrix} = \begin{bmatrix} l'_{t,1} R'_t & \dots & l'_{t,K} R'_t \end{bmatrix} \cdot \begin{bmatrix} S_1 \\ S_2 \\ \dots \\ S_K \end{bmatrix}, \qquad (4)$$

Stacking all point tracks over the whole sequence into a large measurement matrix W, we can write

$$W = \underbrace{\begin{bmatrix} l'_{1,1} R'_1 & \dots & l'_{1,K} R'_1 \\ l'_{2,1} R'_2 & \dots & l'_{2,K} R'_2 \\ \dots & \dots & \dots \\ l'_{N,1} R'_N & \dots & l'_{N,K} R'_N \end{bmatrix}}_{Q'} \cdot \underbrace{\begin{bmatrix} S_1 \\ S_2 \\ \dots \\ S_K \end{bmatrix}}_{B}, \qquad (5)$$

Here the 2*N*x3*K* matrix *Q'* contains for each time frame *t* the pose R'_t and configuration weights $l'_{t,1}, \dots l'_{t,K}$, and the 3*K*x*P* matrix *B* codes the *K* key-frame basis shapes S_i. In the noise free case, rank of *W* is $r \le 3K$. This factorization can be realized using SVD, i. e. $W = U \sum V^T = \hat{Q}' \cdot \hat{B}$, only considering the first *r* singular values and singular vectors.

The next step is to extract the pose R'_t and shape basis weights $l'_{t,1}, \dots l'_{t,K}$ from the matrix $\hat{Q}'$. For each $\hat{Q}'_t$ in $\hat{Q}'$, it can be written as (for convenience, the time index is dropped) [13]

$$\hat{Q}'_t = \begin{bmatrix} l'_{t,1} R'_t & \dots & l'_{t,K} R'_t \end{bmatrix} = \begin{bmatrix} l'_1 r'_1 & l'_1 r'_2 & l'_1 r'_3 & \dots & l'_K r'_1 & l'_K r'_2 & l'_K r'_3 \\ l'_1 r'_4 & l'_1 r'_5 & l'_1 r'_6 & \dots & l'_K r'_4 & l'_K r'_5 & l'_K r'_6 \end{bmatrix}.$$

The elements of $\hat{Q}'_t$ can be reordered into a new matrix:

$$\overline{Q}'_t = \begin{bmatrix} l'_1 r'_1 & l'_1 r'_2 & l'_1 r'_3 & l'_1 r'_4 & l'_1 r'_5 & l'_1 r'_6 \\ l'_2 r'_1 & l'_2 r'_2 & l'_2 r'_3 & l'_2 r'_4 & l'_2 r'_5 & l'_2 r'_6 \\ & & \dots & \dots & & \\ l'_K r'_1 & l'_K r'_2 & l'_K r'_3 & l'_K r'_4 & l'_K r'_5 & l'_K r'_6 \end{bmatrix}$$

$$= \begin{bmatrix} l'_1 \\ l'_2 \\ \dots \\ l'_K \end{bmatrix} \cdot \begin{bmatrix} r_1 & r_2 & r_3 & r_4 & r_5 & r_6 \end{bmatrix}. \qquad (6)$$

which shows that $\hat{Q}'_t$ is *rank of 1* and also can be factored by SVD. Because this factorization is not unique, there exists one invertible matrix *G* that ortho-normalizes all of the sub-blocks $\hat{Q}'_t$. Thus it leads to an alternative factorization:

$$Q' = \hat{Q}' \cdot G , \quad B = G^{-1} \cdot \hat{B} . \qquad (7)$$

Irani exploited rank constraints in [7] for optic flow estimation in the case of rigid motion. Building on this technique, a framework of robust tracking could be set up (details are in [13]).

2.2 Stereo-Motion Model

Below we also utilize the rank constraints to help stereo correspondence. Let $(\overline{R}, \overline{T})$ be the rotational and translational relationships between the stereo cameras. Under a scaled orthographic camera model we can also assume the shape has been centered at the origin. Therefore the translation $\overline{T}$ could be subtracted from the shape relationship, since a translation part in depth has only effect on the scale factor and a translation part in the image plane is eliminated. So the 3D coordinates of any point with respect to the two camera coordinate frames, S_l and S_r, and the corresponding shape basis, $S_{l,i}$ and $S_{r,i}$ ($i = 1, 2, \dots, K$), are related by

$$S_r = \overline{R} \cdot S_l , \; S_{r,i} = \overline{R} \cdot S_{l,i} , \; i = 1, 2, \dots, K. \qquad (8)$$

Now we rewrite (4) as

$$W = \underbrace{\begin{bmatrix} \dots & & \\ & F_t & \\ & & \dots \end{bmatrix}}_{F} \underbrace{\begin{bmatrix} l_{1,1} R_1 & \dots & l_{1,K} R_1 \\ l_{2,1} R_2 & \dots & l_{2,K} R_2 \\ \dots & \dots & \dots \\ l_{N,1} R_N & \dots & l_{N,K} R_N \end{bmatrix}}_{Q} \cdot \underbrace{\begin{bmatrix} S_1 \\ S_2 \\ \dots \\ S_K \end{bmatrix}}_{B}, \qquad (9)$$

where F_t is the 2x3 scaled orthographic projection matrix given by

$$F_t = s_t \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad (10)$$

with s_t as the scale factor at time frame *t*.

Applying the non-rigid motion model to the two cameras separately, one obtains two image measurement matrices respectively as

$$W_l = F_l Q_l B_l , \; W_r = F_r Q_r B_r . \qquad (11)$$

Because the shape is centered at the origin, we can omit the translation component in the relationship between two rigid motion representations for two camera coordinate frames and only consider the relationship of rotation components as $R_{r,t} \overline{R} = \overline{R} R_{l,t}$ (Some derivations are given in our technical report*) . Consequently, we write

* http://www.ifp.uiuc.edu/~yuhuang/Factorization03.pdf

$$\underbrace{\begin{bmatrix} W_l \\ W_r \end{bmatrix}}_{\Psi} = \begin{bmatrix} F_l Q_l B_l \\ F'_r \widetilde{E} Q_l B_l \end{bmatrix} = \underbrace{\begin{bmatrix} F_l & \\ & F'_r \end{bmatrix}}_{H} \underbrace{\begin{bmatrix} I_{3N} \\ \widetilde{E} \end{bmatrix}}_{E} Q_l B_l \qquad (12)$$

where F'_r actually has coded the scaling change of F_r due to translation $\overline{T}$, and the 3*N*x3*N* matrix $\widetilde{E}$ is given as

$$\widetilde{E} = \begin{bmatrix} \cdots & & \\ & \overline{R} & \\ & & \cdots \end{bmatrix}. \qquad (13)$$

Equation (12) represents the matrix decomposition of the stereo-motion correspondences into 3D structure B_l, the rigid motion and shape basis weights Q_l, the stereo geometry E and the camera parameters H. It is obvious, like W_l and W_r, Ψ is of rank at most 3*K*: *rank* (Ψ) $\leq$ 3*K*. Below based on this rank property, we can infer stereo matching from motion correspondences.

2.3 Stereo Matching Inference

Assume distinct feature points are extracted from the stereo image sequences, and in each sequence they are tracked separately using the motion correspondence method. Now the stereo correspondences are not established yet while the estimated dense motion correspondences are assumed to be mostly correct. With such motion correspondences, the measurement matrixes W_l^* and W_r^* can be constructed, here different from W_l and W_r, their columns have not been properly ordered. As Ψ is of rank at most 3*K*, a basis of the 3*K*-dimensional subspace could be set up as long as a minimum of 3*K* linearly independent columns of Ψ are available. Then all the other columns of Ψ are inferred from the set of basis.

Suppose *k* matches are obtained by some stereo correspondence technique with epipolar constraints (To simplify 1D searching on the epipolar line, the technique of image *rectification* could be done prior to stereo matching), where $k \geq 3K$. The corresponding columns of W_l^* and W_r^* can be stacked into a 4*N*x*k* sub-matrix Ψ_k. SVD of Ψ_k is $\Psi_k = U_k \sum_k V_k^T$. Actually the first 3*K'* columns of U_k construct the optimal basis of 3*K'*-dimensional vector subspace (Note *K'* is the estimated number of shape basis, which maybe is not equal to the true number *K*.). Let $\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{3K'}$ be the extracted basis vectors of the column space of Ψ and let a 4*N*x3*K'* matrix $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{3K'}]$, so a column $\mathbf{v}$ of Ψ is only a linear combination of the columns of $\mathbf{A}$. Let two 2*N*x3*K'* matrixes $\mathbf{A}_l$ and $\mathbf{A}_r$ be the top-half and the bottom-half sub-matrixes of $\mathbf{A}$ respectively such that the columns of $\mathbf{A}_l$ belong to W_l^* and the columns of $\mathbf{A}_r$ to W_r^*. For a column $\mathbf{v}_l$ of W_l^* its stereo correspondence $\mathbf{v}_r$ in W_r^* can be predicted from $\mathbf{A}_l$ and $\mathbf{A}_r$ as:

$$\mathbf{v}_r = (\mathbf{A}_r \mathbf{A}_l^+) \mathbf{v}_l \qquad (14)$$

where $\mathbf{A}_l^+$ is the pseudo-inverse of $\mathbf{A}_l$ and is given by

$$\mathbf{A}_l^+ = (\mathbf{A}_l^T \mathbf{A}_l)^{-1} \mathbf{A}_l^T. \qquad (15)$$

But the predicted result may not be exact due to noise. A measure for feature matching could be count on: normally we calculate the least-mean-squares-error (LMSE) in all the positions over the entire image sequence with reference to the prediction results; However, even this measure is small enough, we can not guarantee it is a correct pair of stereo matching; An additional measure related to windowed template matching is probably taken into account, i.e. the average normalized correlation must be high enough [5]. If not, the image feature is ignored. Finally all the inferred stereo correspondences are grouped together to re-estimate the basis **A**, which is supposed to be more accurate. This process could be iterated till convergence.

However, we still reconstruct 3D deformable shape via triangulation from views of the calibrated stereo cameras once all the stereo correspondences are obtained [6]. Consequently we can calculate by factorization the 3D shape basis from the measurement matrix of 3D point positions, similar to (5) and (9), then extract the pose parameters and shape basis configuration weights by rank-1constraints. Different from (5), this time we can extract all nine components of the rotation matrix rather than only the top two rows. Recovering the pose R_t and original configuration weights $l_{t,1}, ... l_{t,K}$ actually has realized 3-D non-rigid tracking.

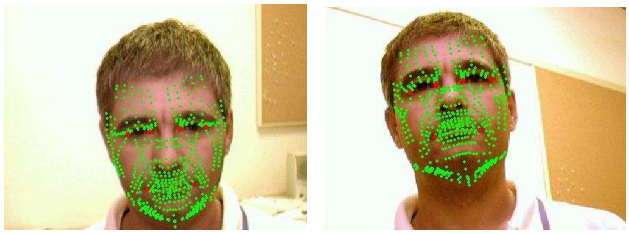
3. EXPERIMENTAL RESULTS

Because of limitation in space, only results with real data are given here. In the experimental setup the two digital video cameras are mounted vertically and connected to a PC through 1394 links. The human face recordings in the collected videos are captured with resolution 320x240 at 30 frames per second. They contain rigid head motions, and non-rigid eye/eyebrow/mouth facial motions.

It is difficult to estimate optical flow from facial motions using traditional gradient-based or template matching methods because the facial surface is smooth and its motion is non-rigid. We choose to use a Bazier Volume model-based face tracker to obtain the optical flow around the face area [11]. For each camera, we track the facial motion using independent face trackers with a dense 3D geometrical mesh model. The first experiment we did is to reconstruct the facial structure from rigid facial motions. In the videos, the human head moves up and backward within 30 frames. A pair of stereo images with depicted tracking points is shown in Fig. 1.

As the face trackers are applied independently to the video sequences of the two cameras. We don't know whether there is correspondence between the mesh points of the face models used by the two face trackers, except those points at the eye corners and mouth corners. We identify these points as distinct feature points (shown in red) and the correspondences of the rest points are inferred using the bases factorized from the optical flow vectors of these distinct feature points. In the rigid motion case, we take the number of bases K=3. Fig. 2 shows the found correspondences of optical flows estimated from the two face trackers. The red trajectories are the mapping of the optical flow of the mesh points from upper camera view to lower camera view using equation (14). The green trajectories show the found

correspondent trajectory of mesh points from video of lower camera. After the correspondence is established, the 3D face geometrical structure in each time instant can be reconstructed. Fig. 3 shows the reconstructed mesh points in the 3D space.



(a) Upper camera (b) Lower camera

Fig. 1. Tracking result for rigid motion

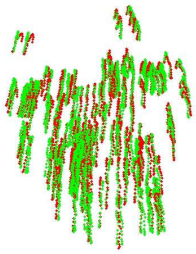
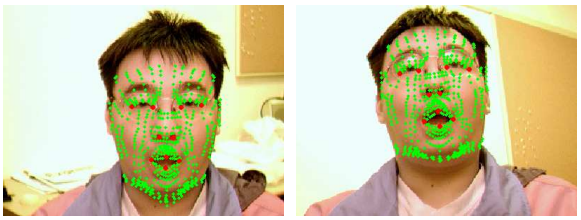


Fig. 2. Optical flow trajectories

Fig. 3. Reconstructed points

In order to verify our theories with non-rigid motion, we further identified a stereo video sequences in which the subject opens mouth within 8 frames. As shown in Fig. 4, the distinct facial features (depicted in red) are the eye corners, mouth corners, nostrils, and the center of the upper and lower lip. As the non-rigid motion only contains the opening mouth, we take K=6 in this case. The found correspondences of optical flow trajectories are shown in Fig. 5. It is shown that most of the found correspondences of the optical flow trajectories are caused by the opening mouth. The reconstructed 3D face geometric structure is shown in Fig. 6, where the purple dots are the reconstructed 3D points.



(a) Upper camera (b) Lower camera

Fig. 4. Tracking results for non-rigid motion.

Fig. 5. Optical flow trajectories

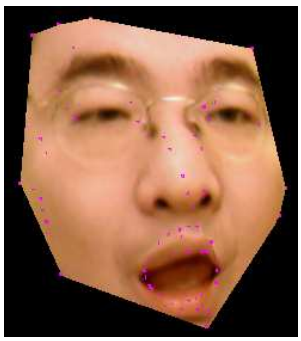


Fig. 6. Reconstructed face.

4. CONCLUSIONS AND FUTURE WORK

We have presented a framework for recovering 3D non-rigid shape and motion viewed from calibrated stereo cameras. This approach is a factorization-based method, so it naturally has the property of rank constraints. Meanwhile it gives a mechanism of inferring stereo correspondences from motion correspondences only requiring that a minimum of $3K$ point stereo correspondences are created initially. The combination of motion and stereo cues offers such advantages as simpler stereo correspondence and accurate reconstruction even with short sequences. Experimental results from real stereo sequences are also given to demonstrate the performance of the proposed method. Future work will address how to detect not a few outliers for "robust" factorization and how to realize 3D model-based tracking along with model refinement.

5. ACKNOWLEDGEMENT

Thank Dr. Zhengyou Zhang at Microsoft Research for allowing us to use the test stereo sequences.

6. REFERENCES

[1] M. E. Brand, "Morphable 3D models from video". IEEE CVPR'01, December 2001.
[2] R. L. Carceroni, K. N. Kutulakos, "Multi-View Scene Capture by Surfel Sampling: From Video Streams to Non-Rigid 3D Motion, Shape Reflectance", ICCV'01, June 2001.
[3] A. Del Bue, L. Agapito, "Non-rigid stereo factorization", IJCV, 66(2), 193-207, 2006.
[4] F. Dornaika and R. Chung, "Stereo Correspondence from Motion Correspondence", IEEE CVPR'99, pp 70-75, 1999.
[5] P K Ho and R Chung, "Stereo-Motion that Complements Stereo and Motion Analysis", IEEE CVPR97, pp213-218, 1997.
[6] Y. Huang, T. S. Huang, "Facial Tracking with Head Pose Estimation in Stereo Vision", IEEE ICIP'02, Sept., 2002.
[7] M. Irani, "Multi-Frame Optical Flow Estimation Using Subspace Constraints". IEEE ICCV'99, September 1999.
[8] M. Irani and P. Anandan, "Factorization with Uncertainty". ECCV'00, June 2000.
[9] J. Neumann and Y. Aloimonos, "Spatio-temporal stereo using multi-resolution subdivision surfaces". IJCV, 47(1): 181-193, 2002.
[10] G. Stein and A. Shashua, "Direct estimation of motion and extended scene structure for a moving stereo rig", IEEE CVPR'98, 1998.
[11] H. Tao and T. S. Huang, "Explanation-based facial motion tracking using a piecewise Bezier volume deformation model," IEEE CVPR'99, 1999.
[12] C. Tomasi and T. Kanade, ``Shape and Motion from Image Streams under Orthography: a Factorization Method", IJCV, vol.9, no.2, pp.137-154, 1992.
[13] L. Torresani, D. Yang, G. Alexander, C. Bregler, **"**Tracking and Modelling Non-Rigid Objects with Rank Constraints", IEEE CVPR'01, 2001.
[14] B. Triggs, ``Factorization Methods for Projective Structure and Motion", Proc. IEEE CVPR'96, pp.845--851, 1996.
[15] J. Xiao, J. Chai, T. Kanade, "A closed-form solution to non-rigid shape and motion recovery", ECCV'04.