A Globally Optimal Approach for 3D Elastic Motion Estimation from Stereo Sequences

Qifan Wang, Linmi Tao, and Huijun Di

State Key Laboratory on Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Computer Science and Technology, Tsinghua University, China
{wqfcr618,tao.linmi,ajon98}@gmail.com

Abstract. Dense and markerless elastic 3D motion estimation based on stereo sequences is a challenge in computer vision. Solutions based on scene flow and 3D registration are mostly restricted to simple non-rigid motions, and suffer from the error accumulation. To address this problem, this paper proposes a globally optimal approach to non-rigid motion estimation which simultaneously recovers the 3D surface as well as its non-rigid motion over time. The instantaneous surface of the object is represented as a set of points which is reconstructed from the matched stereo images, meanwhile its deformation is captured by registering the points over time under spatio-temporal constraints. A global energy is defined on the constraints of stereo, spatial smoothness and temporal continuity, which is optimized via an iterative algorithm to approximate the minimum. Our extensive experiments on real video sequences including different facial expressions, cloth flapping, flag waves, etc. proved the robustness of our method and showed the method effectively handles complex nonrigid motions.

1 Introduction

3D elastic motion estimation is one of the long lasting challenges in computer vision. The most popular approach to motion capture is to attach distinctive markers to deformable objects, and tracked in image sequences acquired by two or more calibrated video cameras. The tracked markers are then used to reconstruct the corresponding 3D motion. The limitation to these distinctive marker based technologies is that the number of distinctive markers is relatively rather sparse comparing to the number of points of a reconstructed 3D surface in [1,3].

Markerless motion capture methods based on computer vision technology offer an attractive alternative. Two main streams of researches have been implemented in the past decades. On one hand, approaches based on scene flow [6,7,8,9,11] have been proposed to independently estimate local motions between adjacent and transfer into a long trajectory. Obviously, error accumulation is the main problem of these approaches. On the other hand, registration among reconstructed surfaces is also a way to get the trajectory by tracing the motion of vertices [14,15], but the error in 3D surface reconstruction will leads to the fail of the tracking, which means the errors in spatial surface reconstruction will be accumulated in temporal tracking. As a result, both the approaches are

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part IV, LNCS 6314, pp. 525–538, 2010.

limited to handle slow and simple non-rigid motion, in short time. In this paper, we propose a novel probabilistic framework to estimate markless non-rigid 3D motion from calibrated stereo image sequences.

1.1 Related Work

The most popular approach (such as [7,8]) on this topic obtain the scene flow to establish the 3D motion field. Scene flow was introduced by Vedula $et\ al.$ [10] as a 3D optical field which is naturally another form of 3D displacement field. These methods recover the scene flow by coupling the optical flow estimation in both cameras with dense stereo matching between the images. Work [9,11] first compute the optical flow in each image sequence independently, then couple the flow for the 3D motion. Others such as [6,13] directly estimate both 3D shape and its motion. A variational method is proposed in [12], this work proposed one formulation that does both reconstruction and scene flow estimation. Scene flow estimation is performed by alternatively optimizing the reconstruction and the 2D motion field. An efficient approach for scene flow estimation is proposed in [8], which decouple the position and velocity estimation steps, and to estimate dense velocities using a variational approach.

Although existing scene flow algorithms have achieved exciting results, these approach suffers from two limitations since they do not exploit the redundancy of spatio-temporal information. First, scene flow methods estimate the 3D motion based on local consistency (optical flow), which restrict the algorithms in handling little deformation. Second, local motions are independently calculated between adjacent frames and then concatenated into long trajectories, leading to error accumulation over time. Recently, work by Di $et\ al.$ [2,17] achieve groupwise shape registration on the whole image sequence which utilize a dynamic model to obtain the 2D elastic motion. These works gain some remarkable results without error accumulation.

Method based on registration among reconstructed 3D shapes is also proposed to estimate 3D motions. 3D active appearance models (AAMs) are often used for facial motion estimation [4]. Parametric models which are used to encode facial shape and appearance are fitted to several images. Most recently, Bradley $et\ al.$ [14] deploy a camera array and multi-view reconstruction to capture the panoramic geometry of garments during human motion. Work by Furukawa $et\ al.$ [15] uses a polyhedral mesh with fixed topology to represent the geometry of the scene. The shape deformation is captured by tracking its vertices over time with a rigid motion model and a regularized nonrigid deformation model for the whole mesh. An efficient framework for 3D deformable surface tracking is proposed in [5], this work reformulate the SOCP feasibility problem into an unconstrained quadratic optimization problem which could be solved efficiently by resolving a set of sparse linear equations.

These registration methods achieve impressive effort especially in garment motion capture. However, the reconstruction and registration process are performed separately. In this case, solving 3D shape reconstruction and motion tracking alone, ignoring their interrelationships, is rather challenging, which leads imprecise motion estimation. One way to improve these methods is to draw dependency between the reconstruction and tracking by integrate them within a unified model.

1.2 Proposed Approach and Contribution

This paper addresses elastic motion estimation from a synchronized and calibrated stereo sequences, which is treated as a joint tracking and reconstruction problem. A novel generative model - Symmetric Hidden Markov Models(SHMMs) is proposed to model the spatio-temporal information with stereo constraint of whole sequences. The main contribution of this paper is: the proposed globally optimal approach can handle fast, complex, and highly nonrigid motions without error accumulation over a large number of frames. This involves several key ingredients: (a) a generative model, which fully exploits the spatio-temporal information to simultaneously recover the 3D surface and obtain its motion over time; (b) nonrigid transformation functions, which effectively describe the elastic deformation; (c) a common spatial structure - a mean shape, which is automatically learned and utilized to establish the correspondence among the shapes in each image(details in Sec. 2).

2 Symmetric Hidden Markov Models and Problem Formulation

2.1 Basic Idea

The problem of 3D elastic motion estimation consists of two subproblems: shape reconstruction and motion tracking. On one hand, the method for 3D shape reconstruction has been fully developed (based on stereo vision), and an intuitive idea is to solve the problem of 3D motion estimation in two steps: first to reconstruct surface frame by frame and then to track points on the surface over time. On the other hand, the optical flow based methods for 2D motion tracking have matured as well. Straightforwardly, scene flow approaches coupled optical flow with disparity for 3D motion estimation between adjacent frames. Both the approaches obtained dense 3D motion, however, suffered from the accumulation of both reconstruction and tracking errors.

Inspiring by the work on Di $et\ al.$ [2,17], we assimilate its idea of groupwise shape registration which avoids error propagation. A straightforward approach is: first independently apply [2] to each of the stereo sequences which obtains the 2D elastic motion in both views, and then reconstruct the 3D motion via stereo matching. In this way, reconstruction and registration are still performed separately which lead to imprecise result(discussed in Sec. 4.2). Our idea is to couple the two image sequences together with stereo constraint by integrating a stereo term into an unified tracking model–Symmetric Hidden Markov Models(SHMMs). This stereo term bridges the left and right sequences, through which spatio-temporal information from both sequences could pass to each other(see fig.1). In other words, our globally optimal approach fully exploits the spatio-temporal information in both views via stereo constraint, therefore draws dependency between the registration and reconstruction.

We define a 2D shape in each image which is represented as a set of sparse points, say L points. In each image, these L points are clustered from the edges by assuming they are the L centers of a Gaussian Mixture Models(GMMs) that generates the edges. Inspired by [19], a 2D mean shape is introduced into SHMMs as a common spatial structure, which is also represented by L points, and the spatial constraints associated to SHMMs are imposed by registering the mean shape to 2D shape in each image

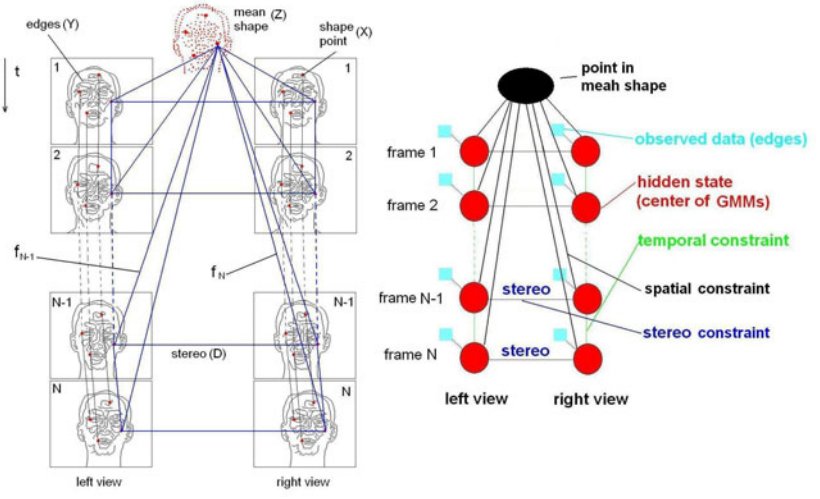

Fig. 1. Our basic idea: Symmetric Hidden Markov Model with stereo, spatial and temporal constraints

through a smooth nonrigid transformation f (see fig.1). Our tracking is performed by the registering of 2D shapes to our mean shape over time in both views. Meanwhile, the reconstruction is accomplished through the matching of the 2D shapes in stereo image pairs. In short, our task is simultaneously registering all the 2D shapes to the mean shape in both image sequences through smooth nonrigid transformations.

2.2 Symmetric Hidden Markov Models

Since we are working on the whole stereo image sequences, we first rectify the stereo images and estimate the disparity field using a stereo algorithm [18] for each stereo image pairs as the stereo constraint in SHMMs. As mentioned in sec. 2.1, the shape is represented as L points. Then these L points come into being L trajectories or Hidden Markov Models(HMMs) along the time domain in each image sequence, and therefore form L Symmetric Hidden Markov Models(SHMMs) in the stereo sequences. In our SHMMs, L points representing the shape are L hidden states; the edges extracted in each frame [16] are our observations.

Our SHMM is displayed in right fig.1, a hidden state stands for a center of GMM while the edges are treated as its observation. Each hidden state belongs to a HMM while the rest of the hidden states in the same HMM delegate the corresponding states in the temporal domain. A couple of corresponding HMMs in the stereo sequences enforced by the stereo constraints form a SHMM. The correspondence among all hidden states in one SHMM is founded by registering them to a identical point in the mean shape via transformations. Note that each SHMM essentially represents a trajectory of a 3D point. 3D shape deformation can be achieved when all L SHMMs are inferred, and the dense 3D motion is obtained through the TPS(see Sec. 2.3.2).

2.3 Problem Formulation

Before giving the formulation of our SHMMs, let us introduce the following notation for better understanding. Assume that there are N frames in each image sequence, t denotes the index of frames, $t \in \{1, 2, \ldots, N\}$, k denotes the viewpoint, $k \in \{l, r\}$; Let $I_{k,t}$ be the t^{th} image in the left($k = l$) or right($k = r$) sequence; $D = \{D_t\}$ be the disparity maps drew from the stereo image pairs; $Y_{k,t} = \{Y^i_{k,t} | i = 1, 2, \ldots, N_{k,t}\}$ be the edge point set of image $I_{k,t}$, where $N_{k,t}$ is the number of edge points.

Now we want to obtain the 3D shape motion. From the stereo vision we know that a 3D point could be reconstructed by two corresponding points in stereo images; The two corresponding points also could be seen as the projection of a 3D point onto two images. Since we have rectified the images sequences, the corresponding points in two views have the same y. So we define our 3D points(shape) as $X_t = \{[x^j_{l,t}, x^j_{r,t}, y^j_t] | j = 1, 2, \ldots, L\}$ which is clustered from edge set Y_t, where L is the number of the points in the shape. Actually $X^j_{l,t} = [x^j_{l,t}, y^j_t]$ is the 2D projection in the left image and $X^j_{r,t} = [x^j_{r,t}, y^j_t]$ is its corresponding point in the right image. Let V^j_t be the velocity of X^j_t; $Z = \{Z^j | j = 1, 2, \ldots, L\}$ be the common spatial structure - mean shape. The variable $X^j_{k,t}$ stands for the position of a hidden state in a SHMM. Each shape $X_{k,t}$ in our SHMMs matches to the mean shape Z via a smooth nonrigid transformation, $f = \{f_{k,t}\}$(see fig.1). We can simply write the equation:

$$X_{k,t} = f_{k,t}(Z) \tag{1}$$

note that X^j_t and V^j_t are 3D vector denoting the position and velocity of the points; $X^j_{k,t}$ is the 2D projection of X^j_t; Y^i_t and Z^j are 2D vector denoting the position of the edge points and mean shape; $f_{k,t}$ is $\mathbb{R}^2 \rightarrow \mathbb{R}^2$ transformation.

Our problem is: Given disparity map sets D and edge point sets Y, we want to obtain an optimal solution of mean shape Z, nonrigid transformation f and the velocity V. Then each shape X can be directly obtained by eqn. 1.

The global energy based on our SHMMs includes four terms: a data term $E_d(Z, f; Y)$, which models the inherent relations between observation Y and hidden states X under GMM; a spatial term $E_{sp}(f)$, which enforces the smoothness of the transformations; a temporal term $E_t(Z, f, V)$, which encodes the temporal continuity by modeling the kinematics motion of points; and a stereo term $E_{st}(Z, f; D)$, which embeds the stereo constraint of the hidden states in stereo images. Note that the position of hidden states X can be represented as $f(Z)$ in terms of eqn. 1. Now we write the global energy as:

$$\begin{aligned} E(Z, f, V; D, Y) = E_d(Z, f; Y) + \alpha E_{sp}(f) \\ + \beta E_t(Z, f, V) + \gamma E_{st}(Z, f; D). \end{aligned} \tag{2}$$

α, β and γ are weight parameters that control the proportion of each term in the global energy.

Date Term. The data term encodes the relation between observation and hidden state node in all frames and it is defined as:

$$E_d(Z, f; Y) = \sum_{k,t} \sum_{i=1}^{N_{k,t}} \sum_{j=1}^{L} \left(Q(m_{k,t}^i = j) \| Y_{k,t}^i - f_{k,t}(Z^j) \|^2 / {\sigma_{k,t}^j}^2 \right) \tag{3}$$

where $\sigma_{k,t}^j$ is the variance of Gaussian distribution, $k \in \{l, r\}$, $t \in \{1, 2, \ldots, N\}$; $m_{k,t}^i$ is a discrete variable introduced to denote an index of the Gaussian mixture, which generates the i^{th} edge point $Y_{k,t}^i$ and $m_{k,t}^i \in \{1, 2, \ldots, L\}$; $Q(m_{k,t}^i = j)$ is the probability of $m_{k,t}^i = j$, where $\sum_{j=1}^{L} Q(m_{k,t}^i = j) = 1$.

The reasoning of the data term is depicted below: In a single image(say $I_{k,t}$), as mentioned in section 2.1, the observation is considered as a GMM with L centers. Then the complete density function of $Y_{k,t}^i$ is then given by:

$$P(Y_{k,t}^i | m_{k,t}^i = j, f_{k,t}(Z^j)) = \sum_{j=1}^{L} Q(m_{k,t}^i = j) N(Y_{k,t}^i, f_{k,t}(Z^j), {\sigma_{k,t}^j}^2) \tag{4}$$

where $N(X, \mu, \sigma)$ denote the Gaussian distribution on X with mean μ and variance σ. Note that $f_{k,t}(Z^j)$ essentially represent the j^{th} center $X_{k,t}^j$ of the GMMs in terms of eqn. 1. Assuming that all the edge points are independent and identically distributed (i.i.d.) with distribution eqn.4. Therefore, our data energy term eqn. 3 comes from the negative log likelihood of the joint $posterior\ \ probability$ of all the edge points. This term indicates how well the observation and the hidden states fit the GMMs.

Spatial Term. Spatial smoothness is considered in a way that each shape should be smoothly transformed from the mean shape through a smooth nonrigid transformation $f : \mathbb{R}^2 \to \mathbb{R}^2$, the smoothness of the transformation f can be measured by:

$$\|\triangle f\|^2 = \int \int \left[(\frac{\partial^2 f}{\partial x^2})^2 + 2(\frac{\partial^2 f}{\partial x \partial y})^2 + (\frac{\partial^2 f}{\partial y^2})^2 \right] \tag{5}$$

which is one of popular choices and invariant under rotations and translations, where $\triangle$ is the $Laplace\ \ operator$. When minimizing the functional related to eqn. 5, the optimal f will be the well known thin-plate spline (TPS) [20] and the measure in eqn. 5 will be the bending energy of a thin plate of infinite extent. And the $E_{sp}(f)$ term can be written as:

$$E_{sp}(f) = \sum_{k,t} \|\triangle f_{k,t}\|^2 \tag{6}$$

when considering all the transformation function f, where $k \in \{l, r\}$, $t = 1, 2, \ldots, N$. This term enforces the spatial smoothness of the transformations, which transfer the mean shape to each shape. Dense points matching between the stereo images and among the temporal images can be achieved by defining a dense mesh in the mean shape and warping it to all the images via these transformations f.

Temporal Term. In order to enforce the temporal constraint, we define the temporal term $E_t(Z, f, V)$ as:

$$E_t(Z, f, V) = \sum_{t=1}^{N} \sum_{j=1}^{L} \left(tr(S_t^j - S_{t-1}^j A) \Psi^{-1} (S_t^j - S_{t-1}^j A)^T / {\tau_t}^2 \right) \tag{7}$$

where tr stands for the trace operation of a matrix. The symbol S_t^j is a vector which combines the position of a hidden point with its velocity.

$$S_t^j \equiv [X_t^j \;\; V_t^j] \tag{8}$$

A and Ψ are both matrixes, their form will be given in the reasoning of eqn. 7. τ_t is standard deviation of a normal distribution.

Inspired by the idea of Kalman filter [21], the temporal continuity is enforced by modeling the state transition in the tracking of each HMM. The overall motion may not be easily described, but if we focus on one particular particle on the object, its motion however can be defined under kinematics. Without loss of generalization, we assume that between the $(t-1)^{th}$ and t^{th} frame the j^{th} HMM undergoes a constant acceleration that is normally distributed, with zero mean and standard deviation τ_t. From kinematics we conclude that

$$[X_t^j \;\; V_t^j] = [X_{t-1}^j \;\; V_{t-1}^j] A + [a_p \;\; a_v] G \tag{9}$$

where X_t^j is the position of the 3D point at frame t, V_t^j is its velocity. a_p and a_v are the point's accelerations of displacement and velocity, respectively. Under an ideal case of exact constant acceleration, the random acceleration variable a_p and a_v are perfectly correlated, i.e. their correlation equals to 1(a_p=a_v). But in practice, they may not be perfectly correlated, and this is why two separated accelerations a_p and a_v rather that one are used in eqn. 9. The matrixes A and G are defined as

$$A = \begin{bmatrix} 1 & 0 \\ \triangle t & 1 \end{bmatrix} \qquad G = \begin{bmatrix} \frac{\triangle t^2}{2} & 0 \\ 0 & \triangle t \end{bmatrix} \tag{10}$$

The $\triangle t$ is taken as 1 here as frame index is used as time, both A and G are 2×2 matrixes. Therefore the distribution of $[a_p \;\; a_v]G$ is a Gaussian with zero mean and covariance ${\tau_t}^2 \Psi$. Then eqn .9 can be written as

$$\begin{aligned} &P(X_t^j, V_t^j, X_{t-1}^j, V_{t-1}^j) \\ &= N(S_t^j - S_{t-1}^j A, 0, {\tau_t}^2 \Psi) \end{aligned} \tag{11}$$

where Ψ is calculated from G

The temporal energy term eqn. 7 directly comes from the negative log likelihood of $posteriori$ (eqn. 11) by considering all the HMMs together. This term enforces the temporal continuity via modeling the kinematics motion of points.

Stereo Term. So far we haven't model the relationship between the stereo sequences. Since the two 2D shapes $X_{l,t}, X_{r,t}$ at any time t are the two projections of the same 3D shape X_t, their deformations are inherently related. In order to model this inherent spatio-temporal relationship, we encode stereo constraint as the stereo term $E_{st}(Z, f; D)$ to force a restriction of the motion between the corresponding stereo points.

$$E_{st}(Z, f; D) = \sum_{t=1}^{N} \sum_{j=1}^{L} \rho_d(X_{l,t}^j, X_{r,t}^j, D_t^j) \tag{12}$$

where function ρ_d punish the inconsistent matching of the hidden state between the stereo images under the disparity map, D_t^j gained from D_t is the disparity between $X_{l,t}^j$ and $X_{r,t}^j$. Here we choose a broadly used quadratic cost function, similar to the one used in [24]:

$$\rho_d(P_l, P_r, d) = ||P_l - P_r| - d|^2 \tag{13}$$

This term plays a crucial role in the model that it bridges the information of the two sequences so that shape matching between two views and registration along the time domain could be simultaneously achieved.

The combination of eqn. 3, 6, 7 and 12 is our symmetric model. We now describe an iterative optimization algorithm to minimize the global energy eqn. 2.

3 Inference and Optimization

3.1 Inference Under EM Algorithm

Directly minimizing the global energy in eqn.2 is intractable, as many terms are coupled together. Using the same divide-and-conquer fashion in [22], the optimization problems can be split into two slightly simpler sub-problems. The idea is that we first treat X and V as a whole S(defined in eqn.8) and minimize $E(Z, f, V; D, Y)$ (eqn.2) w.r.t. S, then find f and Z which achieve the optimal X(eqn.1) by solving a fitting problem. In this regard we have $X = SB$, where $B = [1\ \ 0]^T$. Then the two sub-problems are given as

$$\begin{aligned} SP1 : \min_{S} & \sum_{k,t} \sum_{j=1}^{L} \sum_{i=1}^{N_{k,t}} (Q(m_{k,t}^i = j) \|Y_{k,t}^i - S_{k,t}^j B\|^2 / {\sigma_{k,t}^j}^2) \\ & + \sum_{t=1}^{N} \sum_{j=1}^{L} tr(S_t^j - S_{t-1}^j A) \Psi^{-1} (S_t^j - S_{t-1}^j A)^T / {\tau_t}^2 \\ & + \gamma \sum_{t=1}^{N} \sum_{j=1}^{L} \| \|S_{l,t}^j B - S_{r,t}^j B\| - D_t^j \|^2 \end{aligned} \tag{14}$$

$$SP2 : \min_{f,Z} \sum_{k,t} \left(\alpha \|\triangle f_{k,t}\|^2 + \sum_{j=1}^{L} \|f_{k,t}(Z^j) - S_{k,t}^j B\|^2 \right) \tag{15}$$

SP1 is a symmetric trajectory tracking problem, and SP2 is a groupwise shape registration problem. The solution of these two sub-problems can be obtained by an iterative deterministic annealing algorithm under EM framework present in [17]. We refer to sec. 4.2 and 4.3 of [17] for full details on how to achieve the optimal solution of these two energies. Note that S in $SP2$ is obtained from $SP1$ in each iteration and β in eqn.2 is merged into τ_t.

3.2 Outlier and Missing Data Handling

In our SHMMs, if one hidden state $X^j_{k,t}$ is inferred inaccurately due to the outliers (noises and occlusions) or data missing, it will be temporally inconsistent with the others in the same HMM, spatially inconsistent with Z (mapping from Z to X will be non-smooth) and symmetrically inconsistent with the corresponding state in the SHMM. Although constraints of stereo, temporal continuity and spatial smoothness will help in pulling $X^j_{k,t}$ away from the outliers and missing data, too strict constraints may introduce a bias in the estimation of $X^j_{k,t}$. During the EM iteration, the temporal and stereo parameter $1/\tau_{k,t}$ and γ in eqn.14 together with spatial parameter α in eqn.15 will be decreased from an initial value to a small value. Thus towards the end, the spatiotemporal and stereo constraints will have a very tiny bias effect.

In order to account for outliers and missing data, occlusions O is first computed from disparity D using the similar principle in sec. 3.3 [7]. A hidden binary variable $w^i_{k,t}$ is further introduced so that $w^i_{k,t} = 0$ if the edge point $Y^i_{k,t}$ is an outlier, and $w^i_{k,t} = 1$ if $Y^i_{k,t}$ is generated by the GMM. As mentioned in section 2.3.1, $Q(m^i_{k,t} = j)$ is the posterior probability that $Y^i_{k,t}$ is generated by $X^j_{k,t}$, then we have $\sum_{j=1}^{L} Q(m^i_{k,t} = j) = P(w^i_{k,t} = 1|Y^i_{k,t})$, for occlusion, $P(w = 0|Y \in O) = 1$. By introducing $w^i_{k,t}$, outliers can be suppressed and missing data can be handled. The complete EM algorithm for the SHMMs is shown in Table 1. ρ is the correlation between a_p and a_v.

Table 1. Our full iterative algorithm

Initialize f,Z,Y,D
Initialize parameters α,λ,ρ
Begin: Deterministic Annealing
Calculate new $Q(m)$,σ and τ
Symmetric trajectory tracking problem: Solve $SP1$
Groupwise Shape Registration: Solve $SP2$
Update Z and f
Decrease annealing parameters
Repeat until converge
End

4 Experimental Results and Discussion

4.1 Implementation and Datasets

Given the stereo video sequences mentioned above, the edge features and disparity maps, which serve as the inputs, are obtained by using edge detection algorithm [16] for all

Table 2. Characteristics of the six datasets: N, L, E and M are the numbers of frames, shape points, average edge points and vertices on the mesh we use to reconstruct the 3D surface; w and h are the width and the height of input images in pixels; The initial parameters: ρ is fixed to 0.1; α, λ and γ are 5, 0.2 and 10, they decrease during the iterations

Data	$face1$	$face2$	$cloth$	$flag$	$sphere$	$balloon$
N	60	40	50	30	120	20
L	500	500	500	700	300	250
E	4140	4000	5210	6540	3100	2500
M	5200	5200	6600	7010	6200	3200
w	640	640	1024	1024	640	320
h	480	480	768	768	480	240

images and by applying stereo algorithm [18] to all stereo image pairs respectively. Although [18] is not the best stereo algorithms according to the Middlebury benchmarks [23], it achieves fast speed in implementation. The output of the proposed approach is a set of 4D vectors (m_x, m_y, m_z, t), denoting the 3D motion field over time. Our algorithm is implemented in MATLAB and run with almost same speed on a 2.8GHZ CPU. The speed of the algorithm depends on the number of the points L representing the shape, number of frames N and the number of input edge points. For instance, it needs about one and a half hour on a stereo sequence where L equals 500, N equals 40 and there are about 4000 edge points in each image, average 68 seconds per frame.

Six real datasets are used for the experiments: waving $flag$, flapping $cloth$, inflating $balloon$, deforming $sphere$(courtesy of J. Wang, K. Xu [25]), talking $face1$ and $face2$ with different expression. The characteristics of these datasets and the parameter values used in our experiments are given in Table. 2. The motions in $face1$ are slow, but the mouth and head motions in $face1$ are challenging. Motions are fast in $flag$ and $balloon$, but relatively simple. The $sphere$ deforms quickly and dramatically, which makes it hard to track the points on its highly deforming surface. $Cloth$ and $face2$ are quite challenging datasets involving complex motions during their deformation. In $cloth$, the textures are weak compared to the others and the motions are very fast. Moreover, there are occlusions in some part of the video due to some folds. Motions in $face2$ are relatively slow than in $cloth$, but with different facial expressions, the mouth, eye and eyebrow shapes change distinctly and irregularly which makes motion estimation difficult especially in these regions.

4.2 Results and Evaluation

Left most of Fig.2 gives some result on $face1$ from some frames in the left stereo sequences along with the estimated deformation field. The center and right most of Fig.2 demonstrates a inflating $balloon$ and a deforming $sphere$. The red points are some of the tracked particles during the inflating motion and the quick deformation. Scene flow methods which use the intensity as their observation and assume the local consistency of the motion failed on the large scale motion especially when the size or shape of the tracking object varies greatly. However, our method makes use of the

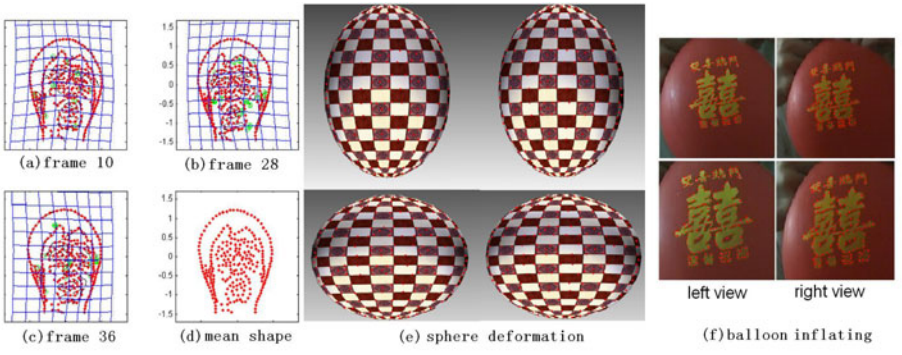

Fig. 2. Several achieved results (a)-(c): edge maps of a face sequence along with estimated deformation field; (d): learned mean shape; (e)-(f): sphere deformation and balloon inflation, the time interval between the top and the bottom stereo image pairs is 0.3s

explicit geometric information among the edges and allows us dealing with large scale motion(see Fig.4 for detail).

Left of Fig.3 shows our result on dataset $flag$, $cloth$ and $face2$, including a sample image from left sequence, the corresponding 3D surface with and without texture mapping and the estimated motion field which is rendered by line segments connecting the positions of sample points in the previous frame (red) to the current ones (green). Textures are mapped onto the reconstructed surface by averaging the back-projected textures from the corresponding images, which is a good way to visually assess the quality of the results, since textures will only look sharp and clear when the estimated shape and motion are accurate throughout the sequence. As shown by the figure, the reconstructed surfaces with sharp images looking close to the originals. Of course, there are some unshapely regions. For instance, the mouth of $face2$ and some part of the fold structure of the $cloth$. Overall however, our algorithm has been able to accurately capture the cloth's and face's complicated shape and motion. Right of Fig.3 gave two texture-mapped mean shapes computed by our method and [2] respectively. In the selected region (red rectangle) the mean shape calculated by the method [2] is much more blurred than the one calculated by our approach, which indicates the registration error of [2] is much bigger than that of our method.

Whereas there are numerous datasets with ground truth for various algorithms in computer vision, the 3D motion estimation problem is probably not mature enough to deserve a proper evaluation benchmark. In order to evaluate our method and get a comparison with other algorithms, we test our algorithm on two typical datasets with ground truth: we manually align 57 points across the stereo sequence of $face2$(use 40 frames)(landmark: 1-8: right eye, 9-16: left eye, 17-29: nose, 30-41: mouth, 42-49: right eyebrow, 50-57: left eyebrow) since face deformation is highly irregular which makes motion estimation difficult and we also use the synthetic, textured $sphere$ in [25] as our test data since its quick and dramatic deformation. We compared three methods on each dataset: (a) our globally optimal method based on SHMM; (b) tracking both sequences using our method in [2], then using stereo information to match shape; (c) pixel-wise

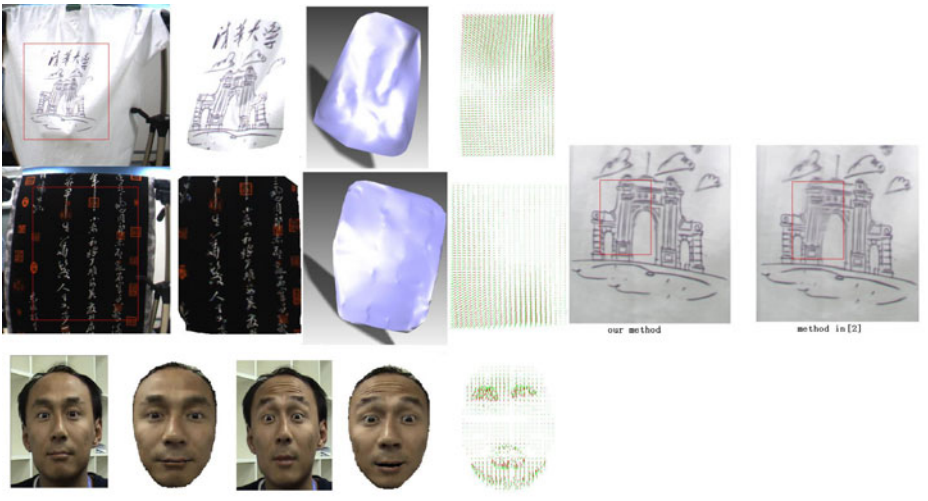

Fig. 3. From left to right in the left top and left mid($cloth$ and $flag$): an input image, a reconstructed surface with and without texture-mapping, and the corresponding motion field; $face2$ is shown at the left bottom of the figure. Our texture-mapped model is indeed very close to the corresponding input image, but there are moderate flaws in some places, in particular in the mouth region of $face2$ dueing to the complex expression and in some folded area of $cloth$. A comparison of the texture-mapped mean shape computed by our method and [2].

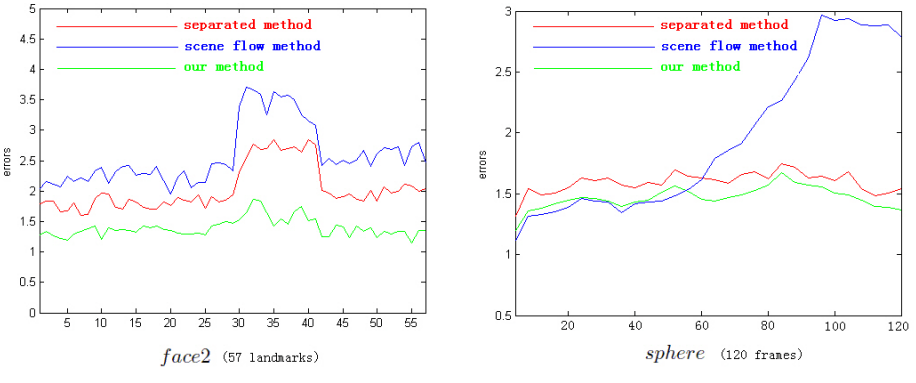

Fig. 4. Comparison of three methods on ground truth data in terms of RMS errors in pixels. Left: landmark-based RMS errors in $face2$. Right: time-based RMS errors in $sphere$.

scene flow algorithm(we simply re-implemented the method in [7]), which couples the optical flow estimation in both image sequence with dense stereo matching between the images.

Fig.4 gives the comparison of the three methods by computing the RMS errors in terms of pixels based on the ground truth data. The results showed that our approach achieved more accurate motion than the other two methods without error accumulation.

Our approach outperforms (b) because method (b) estimated the 3D motion in a separated way. In other words, (b) didn't combine the stereo and the motion together, the reconstructing and tracking were achieved separately. Although scene flow methods gained almost the same accuracy as ours in the first few frames in the $sphere$ data(see Fig.4 right), with the passage of time, the accumulated errors becomes notable since this method only computed the adjacent flow and then concatenated them into long trajectories. Suffering from error accumulation, scene flow methods could not achieve the same accuracy as our approach.

5 Conclusions

In this paper, we proposed a globally optimal approach for 3D nonrigid motion estimation from a stereo image sequences. We embed spatio-temporal information with stereo constraints of whole sequences into a novel generative model - Symmetric Hidden Markov Models. A global energy of the model is defined on the constraints of stereo, spatial smoothness and temporal continuity, which is optimized via an iterative algorithm to approximate the minimum.

Our approach is inspired by Di $et\ al.$ [2] on 2D groupwise shape registration. Experiments on real video sequences showed that our approach is able to handle fast, complex, and highly nonrigid motions without error accumulation. However, our method has two main limitations. First, our method cannot handle the topological interchange of shapes, for instance the object surface comes in contact with itself, e.g. a sleeve touches the torso in a cloth flapping. The extracted edges in these situations can tangle up or be covered with each other, which is not handled in the clustering process. Second, large occlusion leads the disparity range being comparable to the size of the objects in the image, which will not only cause many multi-resolution stereo algorithms hard to obtain accurate disparity but also make our approach difficult to matching shapes between the stereo images. A feasible scheme for dealing with this limitation is to extend our work to multi-view based motion capture [14,15].

Acknowledgments

This research was supported in part by the National Natural Science Foundation of China under Grant Nos. 60873266 and 90820304. The authors would like to thank Kun Xu and Liang Li for their valuable suggestions and helps towards this research.

References

1. White, R., Crane, K., Forsyth, D.: Capturing and animating occluded cloth. ACM Transactions on Graphics (2007)
2. Di, H., Tao, L., Xu, G.: A Mixture of Transformed Hidden Markov Models for Elastic Motion Estimation. IEEE Trans. PAMI 31(10), 1817–1830 (2009)
3. Park, S.I., Hodgins, J.K.: Capturing and animating skin deformation in human motion. ACM ToG 25(3) (2006)

4. Koterba, S.C., Baker, S., Matthews, I., Hu, C., Xiao, J., Cohn, J., Kanade, T.: Multi-view AAM fitting and camera calibration. In: Proc. ICCV (2005)
5. Jianke, Z., Steven, C.H., Zenglin, X., Lyu, M.R.: An Effective Approach to 3D Deformable Surface Tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 766–779. Springer, Heidelberg (2008)
6. Carceroni, R.L., Kutulakos, K.N.: Multi-view scene capture by surfel sampling: From video streams to non-rigid 3D motion, shape and reflectance. IJCV (2002)
7. Huguet, F., Devernay, F.: A variational method for scene flow estimation from stereo sequences. In: Proc. ICCV (2007)
8. Wedel, A., Rabe, C., Vaudrey, T., Brox, T., Franke, U., Cremers, D.: Efficient Dense Scene Flow from Sparse or Dense Stereo Data. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 739–751. Springer, Heidelberg (2008)
9. Vedula, S., Baker, S., Kanade, T.: Image-based spatiotemporal modeling and view interpolation of dynamic events. ACM ToG (2005)
10. Vedula, S., Baker, S.: Three-dimensional scene flow. IEEE Trans. PAMI (2005)
11. Li, R., Sclaroff, S.: Multi-scale 3D scene flow from binocular stereo sequences. In: WACV/MOTION (2005)
12. Pons, J.-P., Keriven, R., Faugeras, O.: Modelling dynamic scenes by registering multi-view image sequences. In: Proc. CVPR (2005)
13. Neumann, J., Aloimonos, Y.: Spatio-temporal stereo using multi-resolution subdivision surfaces. In: IJCV (2002)
14. Bradley, D., Popa, T., Sheffer, A., Heidrich, W., Boubekeur, T.: Markerless Garment Capture. ACM Trans. on SIGGRAPH (2008)
15. Furukawa, Y., Ponce, J.: Dense 3D Motion Capture from Synchronized Video Streams. In: Proc. CVPR (2008)
16. Canny, J.: A computational approach to edge detection. PAMI (1986)
17. Di, H., Iqbal, R.N., Xu, G., Tao, L.: Groupwise shape registration on raw edge sequence via a spatio-temporal generative model. In: CVPR (2007)
18. Felzenszwalb, P.F., Huttenlocher, D.P.: Efficient belief propagation for early vision. IJCV (2006)
19. Chui, H., Rangarajan, A., Zhang, J., Leonard, C.: Unsupervised learning of an atlas from unlabeled point-sets. IEEE Trans. PAMI (2004)
20. Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Trans. PAMI (1989)
21. Forsyth, D., Ponce, J.: Chapter tracking with linear dynamic models, computer vision: A modern approach. Prentice Hall, Inc., Englewood Cliffs (2003)
22. Chui, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration. CVIU (2003)
23. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV (2002)
24. Jian, S., Heung-Yeung, S., Nanning, Z.: Stereo Matching Using Belief Propagation. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2351, pp. 510–524. Springer, Heidelberg (2002)
25. Wang, J., Xu, K., Zhou, K., Lin, S., Hu, S., Guo, B.: Spherical Harmonics Scaling. In: Pacific Conference on Computer Graphics and Applications (2006)