Non-rigid structure from motion from 2-D images using Markov chain Monte Carlo

Huiyu Zhou, Xuelong Li, Senior Member, IEEE and Abdul Sadka, Senior Member, IEEE

Abstract—In this paper we present a new method for simultaneously determining three dimensional (3-D) shape and motion of a non-rigid object from uncalibrated two dimensional (2-D) images without assuming the distribution characteristics. A non-rigid motion can be treated as a combination of a rigid rotation and a non-rigid deformation. To seek accurate recovery of deformable structures, we estimate the probability distribution function of the corresponding features through random sampling, incorporating an established probabilistic model. The fitting between the observation and the projection of the estimated 3-D structure will be evaluated using a Markov chain Monte Carlo based expectation maximisation algorithm. Applications of the proposed method to both synthetic and real image sequences are demonstrated with promising results.

Index Terms—Non-rigid, structure from motion, Markov chain Monte Carlo, uncalibrated.

I. INTRODUCTION

Digital video cameras are widely used, and the quantity of digital videos has dramatically increased recently. For reuse and storage purpose, consumers have to retrieve a video from a large number of multimedia resources. To recognise specific events from a definite database, information retrieval systems have been established with promising performance in searching accuracy and efficiency, e.g., [23]. Many of these established systems attempt to search for events that have been annotated with metadata a priori (e.g. [11]). Nevertheless, there are still a significant number of footages that have been shot but not ever used [1]. These footages normally have not been properly annotated, and hence the retrieval can only be carried out according to the video contents rather than the annotated information.

Of the generic video contents, the need for the ability to retrieve 3-D models from databases or the Internet has gained prominence. Content-based 3-D model retrieval currently remains an active research area, and has found its tremendous applications in information retrieval, medical imaging, and security. To extract a 3-D object, shape-based 3-D modelling (e.g. [29]) and similarity or dissimilarity (distance) computation (e.g. [22]) are two of the main research schemes. Recovery of 3-D structure (or shape) and motion from 2-D images is a critical problem in computer vision [3]. As a solution, studies on camera calibration have been launched in order to relate image coordinates to 3-D coordinates (e.g., [23]). This strategy with promising accuracy and consistency, however, is time consuming while requiring necessary camera calibration.

To release the demand of camera calibration, recent research work has been focused on non-metric reconstruction from uncalibrated image data [34]. Technically, different camera models, e.g. orthographic and perspective projection, have been investigated so that structure and motion recovery can be effectively achieved [18]. In spite of promising performance, many problems still remain open. One of the unsolved problems is that most of these algorithms cannot be directly applied to non-rigid deforming objects (or scenes), as they presumably rely on the assumption of rigid motion [31]. As a result, appropriate approaches to solve the problem of non-rigid structure and motion need to be explored [34], [18].

A number of solutions to this particular problem have been reported in the literature. For example, Szeliski and Kang [28] proposed a batch analysis of image streams in the sense of nonlinear least squares in order to simultaneously recover 3-D structure and motion. Carceroni and Kutulakos [7] stated that structure and motion estimation was equivalent to the recovery of differential properties of the spatial-temporal curve manifold that described the curve’s trace in space-time. To quantify a shape of rank 1 or 2, Xiao and Kanade [37] presented a linear approach that imposed the positive semi-definite constraint to determine the desired solution for this approximation. Similarly, Torresani et al. modeled shape motion as a rigid rotation plus a non-rigid deformation [32], [33]. To handle the arbitrary deformation problem, the object shape at each time instant was extracted from a Gaussian distribution.

Prior knowledge of shape and motion can be used to augment the accuracy of shape and motion recovery. For example, Shaji and Chandran [26] introduced a Riemannian Newton algorithm while using a variant of the non-rigid factorisation schemes for non-rigid structure from motion. Seo and Hong [25] applied the expectation maximisation algorithm based on an extended Kalman smoother to impose the time-continuity of the motion parameters. In [13], a smoothing penalty has been added to the camera trajectory in order to apply prior knowledge of camera pose to the structure recovery. Yan and Pollefeys [38] proposed an approach based on the modeling of the articulated nonrigid motion as a set of intersecting motion
II introduces a generalised 3-D structure recovery method. This is followed by a description about the minimization of the fitting residuals in Section III. In Section IV, we will introduce individual mathematical models in the cases of Gaussian and non-Gaussian distributions. Experimental work shown in Section V demonstrates the performance of the proposed algorithm against a classical approach in synthetic and real data. An application of the proposed algorithm in event recognition is also reported. Finally, conclusions and future work are given in Section VI.

II. GENERALISED 3-D STRUCTURE RECOVERY

It is assumed that a number of image feature points have been tracked throughout a 2-D image sequence. These tracked features are obviously visible while may be of outliers due to the noise and multiple candidates in the same image. Let \( P \) be the observed projection, \( S \) the time-varying shape and \( S = \{S_0, ..., S_t\} \) where \( S_t \) is a shape at time \( t \), \( R \) a \( 2 \times 3 \) matrix combining rotation with the orthographic projection \( [31] \) and \( T \) a \( 2 \times 1 \) translation vector. These parameters, in fact, should have come with time stamps; however, for convenience we omit the time stamps and hereafter. Then, we have the projection after rotating and translating the shape \( S \) with additive noise:

\[
P = R(S + T) + N,
\]

where \( N \) is noise. Taking similar assumptions made in [24],[32], we here consider that a non-rigid shape \( S \) can be represented as a shape average \( \bar{S} \) plus a deformation which has the following form:

\[
S = \bar{S} + \sum_{i=1}^{k} w_i s_i,
\]

where \( i \) is the index of an image point. \( w_i \) are scalar (or weights) that determine the contributions of individual deformation components to each shape. \( \bar{S} \) and \( s \) are regarded as the shape basis and \( k \) is an order. It is recognised that the space of shapes is the linear combinations of basis shapes [6].

Suppose that a shape \( S \) is drawn from a probability distribution \( p(S|\alpha) \), given its parameter \( \alpha \). Knowing the projection \( P \), our intention is to estimate the motion parameters \( R \) and \( T, \alpha \) and the parameter of a probability distribution, \( \sigma^2 \). This can be achieved by maximizing the following conditional probability according to Bayes’ rule:

\[
p(R, T, \alpha, \sigma^2 | P) \propto \int p(P|R, T, S, \sigma^2)p(S|\alpha)dS.
\]

Eq. (3) reveals that the estimation of motion and shape can be performed deducing the parameters of \( p(S|\alpha) \) over the unknown shapes \( S \). An optimal solution can be found when the probability distribution of the observations fits that of the mean shape \( \bar{S} \) over a historic period with a global minimal variance between two probability estimations.

Alternatively, an optimal solution to the 3-D reconstruction of the non-rigid shape can be obtained if the following weighted sum of squares is satisfied:

\[
\min_{R,T,S} \sum_{j=1}^{m} (P - R(S + T))^T \Phi^{-1}(P - R(S + T)),
\]
where \( m \) is the number of the image points, and

\[
\Phi = \text{cov}(P, P) + R \text{cov}(S + \mathbf{T}, S + T)R^T - R \text{cov}(S + T, P) - \text{cov}(S + T, P)^T R^T, \tag{5}
\]

where \( \text{cov} \) is the variance-covariance matrix. To simplify the computation, \( \Phi \) can be approximated as \( \Phi \approx E[(P - R(S + T))(P - R(S + T))^T] \). \( j \) indicates the \( j \)-th image feature point. Eqs. (4)-(5) describe an optimisation procedure, which is used to search for an appropriate solution to the shape and motion recovery problem. Unfortunately, this process cannot guarantee a globally optimal solution. In fact, this optimisation can be biased due to the local minimum in the residuals. To prevent this situation, we take a closer look at Eq. (3).

III. MINIMIZATION OF RESIDUALS

If the residual of the shape estimation is non-Gaussian, then the posterior probability \( p(S|\alpha) \) is also non-Gaussian [19], and vice versa. A variance (also a dominant element of a covariance matrix) is normally used to evaluate the discrepancy between the distribution of the estimated and the true shapes. The residual \( (P - R(S + T)) \) with its variance will be further investigated in this section, where a smaller variance indicates a better fit between the estimated and the true shape. We may never be able to recover the true shape in a real circumstance. To minimise the shape fitting errors in the optimisation, we consider the mean of the estimated shapes in a fixed period as the approximation of the real shape. In the following sub-sections, the composites of the right hand side of Eq. (3) will be analysed individually.

A. Derivation of \( p(P|R, T, S, \sigma^2) \)

The conditional probability \( p(P|R, T, S, \sigma^2) \) is associated with the Euclidean distance between the projections using the estimated \( R, T \) and the observations. This probability employs a form of calculating the Euclidean distance \( \overline{D} \), which is different from the spherical Gaussian form [32]:

\[
p(P|R, T, S, \sigma^2) \sim \overline{D}^{-1}_{|P - R(S + T)|} \tag{6}
\]

Eq. (4) shows that the variance-covariance matrix is fully independent of the residual characteristics (i.e. Gaussian or non-Gaussian). In such a generalisation case, an unbiased and consistent estimation of the variance \( \sigma^2 \) is obtained using the pooled variance estimator [36], which is expressed as

\[
\sigma^2 \propto \frac{\sum_k ((n_k - n_0)(P - R(S + T))^T(P - R(S + T)))}{\sum_k (n_k - n_0)}, \tag{7}
\]

where \( n_k \) is the number of the samples collected from the image space to simulate the shape’s PDF, and \( n_0 \) is the number of the unknown components of the motion and shape estimation. Eq. (7) denotes that a larger \( (n_k - n_0) \) leads to a smaller variance, if the numerator of Eq. (7) is unvaried.

Our aim is to search for the smallest variance during the optimization, which corresponds to the maximisation of \( p(R, T, \alpha, \sigma^2|P) \). To prevent a large \( (n_k - n_0) \) affecting the estimation of the variance, a relatively small number of image points \( n_i \) needs to be chosen for computing a variance. Nevertheless, the estimation based on such a group of points is unstable due to noise or corresponding errors. For example, a group of image points may lead to a very small residual, while others can cause very large residuals. To avoid this situation, a robust estimation of the variance will be sought using a random sampling scheme, e.g., [14]. In other words, an ideal variance will be the one chosen from a group of estimates using Eq. (7), whereas a number of image points are randomly sampled.

The reader can refer to [39] for intermediate procedures of calculating the variance \( \sigma^2 \). The final variance to be determined will be the mean of individual estimates over the overall corresponding features: \( \sigma^2 = \frac{1}{l} \sum_{i=1}^{l} \sigma^2_i \), where \( l \) is the number of the estimated variances (usually \( l \geq 20 \)). This averaging operation is necessary. This process actually boosts the optimisation practice when it falls in a “trap” due to biases or local minimisation. Also, the mean value of the variances within a neighborhood can be used as a robust indicator without making any assumption to the characteristics of the residuals.

B. Derivation of \( p(S|\alpha) \)

Now, we analyse the conditional density function \( p(S|\alpha) \). Again, we start the derivation by decomposing the joint probability into the multiplication of a prior and a posterior:

\[
p(S, \alpha) = p(S)p(\alpha|S), \tag{8}
\]

where the prior can be determined by investigating the correlation of the same shape in two neighbouring image frames:

\[
p(S) = \prod_{t=1}^{t_p} p(S_t|S_{t-1}), \tag{9}
\]

where \( t_p \) is the sample size. Eq. (9) reveals that the probability of a shape depends on the historic estimates of the shape. The shape can move or deform in several possible directions over time. Therefore, assuming that the shapes with their parameters at different time instances are independent of each
other, we shall have: \( p(\alpha | S) = \prod_{i=1}^{\Omega} p(\alpha_i | S_i) \). According to the Bayes’ theorem, we have the posterior density be reformed as follows:
\[
p(S|\alpha) = \frac{p(S, \alpha)}{\int p(S, \alpha) dS} = \frac{p(S)p(\alpha | S)}{\int p(S)p(\alpha | S) dS}.
\] (10)

Without loss of generality, we use a functional \( q(S_i) \) to replace the shape \( S \). This replacement is very important because it is essential to obtain a description of the shape (e.g. shape \( S \)). We here utilise the Metropolis-Hastings algorithm (M-H) in the most right hand side of Eq. (12). For mathematical tractability, the random sampling process. In the M-H algorithm, let \( p(\alpha | S) \) be the proposal density (or called the sampling density), its density function becomes proportional to the numerator of the expression shown in Eq. (11). Unfortunately, it is very difficult to directly generate random draws of \( \alpha \) from \( p(\alpha | S) \). This is due to the fact that \( p(\alpha | S) \) itself is an estimate (see the previous sub-section for details). Assume that \( \beta \) is a sample from \( p(\alpha | S) \), whose probability needs to be determined in the estimation. Therefore, to apply the Gibbs sampler, we derive the conditional density function of \( \beta \) (given \( S \) and \( \alpha \)) as follows:
\[
p(\beta | S, \alpha) = \frac{p(\beta)p(S | \beta)p(\alpha | \beta, S)}{\int p(\alpha | S) d\beta} \propto \frac{p(S)p(\alpha | S)}{\int p(\alpha | S) d\beta}.
\] (12)

When the optimisation iterates for a number of times, the density function becomes proportional to the numerator of the most right hand side of Eq. (12). For mathematical tractability, we here utilise the Metropolis-Hastings algorithm (M-H) in the random sampling process. In the M-H algorithm, let \( q(z|x) \) be the proposal density (or called the sampling density), where \( x \in \mathbb{R}^d \) (\( d \) is dimension size) and \( z \) is a sample. For this proposal density \( q(z|x) \), we then define the acceptance probability \( Q(x, z) \) as follows:
\[
Q(x, z) = \begin{cases} 
\min \left( \frac{p(x | S, \alpha)q(z|x)}{p(x, z | S, \alpha, \phi)}, 1 \right), & \text{if } p(x | S, \alpha)q(z|x) > 0 \\
1, & \text{otherwise}.
\end{cases}
\]

That is to say, the random samples whose probability is larger than \( Q(x, z) \) will have the same effect.

Using a Kalman filter (e.g. [2]), initial shape and motion parameters can be obtained. This provides an approximation of the final solution towards the problem of optimal structure and motion recovery. The M-H algorithm is then used to generate a random draw \( \beta \) from the distribution \( p(S | \alpha) \). This process is summarised in Algorithm 1 [8].

**Algorithm 1** A summary of the M-H algorithm.

1. Repeat for \( i = 1, 2, ..., N \).
2. Generate \( v \) from \( q(z|x^i) \) and \( u \) from \( U(0,1) \) (Gaussian distribution).
3. If \( u \leq Q(x^i, z) \)
   4. - set \( x^{i+1} = v \).
   5. Else
5. - set \( x^{i+1} = x^i \).
6. Return the values \( \{x^1, x^2, ..., x^N\} \).

**Fig. 2.** Illustration of applying the method of [32] to the “Shark” sequence with additive Poisson random noise. Mean = 0 and variance is (a) 1, (b) 5, and (c) 10.

IV. SOLUTIONS OF SHAPE AND MOTION RECOVERY

A. Generalised estimation

In the previous sections, the estimation of non-rigid shape and motion has been introduced without assuming the properties of the PDF. In the presence of clutters or occlusions, we may observe multiple peaks in the estimation of residuals or tilted normal distributions. For example, Fig. 2 illustrates the residual distributions of 50 features randomly selected, where the technique of [32] was used to recover the shape and motion from the “Shark” sequence with different additive Poisson random noise.

This non-Gaussian problem cannot be effectively handled using the classical Gaussian based algorithms (e.g. [32]), as the saddle points in the optimisation hold local minimum/maximum energy that may drive the search towards erroneous solutions. To estimate the motion and deformation model simultaneously, a generalised expectation maximization (EM) algorithm reported in [16], [32] is taken into account in this section. This EM algorithm enables us to reconstruct a 3-D structure from missing data, e.g., occlusions and lighting change in our applications.

An EM algorithm consists of two processing steps, namely E- and M-stages. The generalised EM approach starts from the E-stage. In this stage, the probability of each component that contributes to the shape estimation is formulated. Unfortunately, it is extremely difficult to calculate the posterior in a closed form. A simplified posterior can be formed as follows:
\[
p_i(w | F, \phi) \propto \frac{p_i(F | w, \phi)p_i(w | \phi)}{p_i(F | \phi)} \exp \left( -\frac{\mathcal{F}^T \Omega^{-1} \mathcal{F}}{\sum_{i=1}^{N} \mathcal{F}^T \Omega^{-1} \mathcal{F}} \right).
\] (13)

where \( \mathcal{F} = F - RS - D \), \( \zeta \) is a scalar, \( T \) is transpose, \( F \) is the vectorised form of \( P \), \( w \) represents the weight, the diagonals of the variance-covariance matrix \( \Omega \) are \( \sigma \), \( N \) is the number of
the neighborhood of the feature point, and $D_i$ is the estimated projection.

In the M-stage, we design an objective function that will be optimised in order to find the minimum variance between the target and the estimate. Using a pseudo-linear form, the objective function is described as follows:

$$Q(P, \phi) = \sum E_{p_i(W|F, \phi)} \lambda_1 + \lambda_2,$$

(14)

where $\phi$ is the estimation variance, $E_{p_i(W|F, \phi)}$ is the expectation of $p_i(W|F, \phi)$, and $\lambda_1$ and $\lambda_2$ are two coefficients. Of these parameters, $\lambda_1$ can be determined using the residual between the expectation and the projection with the updated motion parameters:

$$\lambda_1 = \epsilon_i^T \Omega^{-1} \epsilon_i,$$

(15)

where the error residual $\epsilon_i = F_i - \tilde{F}_i - D_i$. $\tilde{F}$ is the vectorised form of $RS$. Also, $\lambda_2$ refers to the difference of the estimates at two consecutive times:

$$\lambda_2 = Q(P, \phi)|_\gamma - Q(P, \phi)|_{\gamma - 1},$$

(16)

where $\gamma$ is the iteration number. Eq. (16) clearly shows the Markovian property in the residual minimisation.

**B. Gaussian distributions**

This is a special case of the generalised form introduced in the previous sub-section. In this special case, Eq. (13) can be further simplified considering the following objective function:

$$\min(Q(P, \phi)) = \min \left( \sum W_i ||F - vec(RS) - D||^2 \right),$$

(17)

where $W_i$ are weight coefficients, which can be determined using the Euclidean distance between the estimated and the averaged projections.

Assuming $N$ is a Gaussian distribution, then we have the following form [32]:

$$p(F|\phi) = N(F|D + \tilde{F}; MM^T + \sigma^2 I),$$

(18)

where $\phi$ encapsulates $R, T, \alpha$ and $\sigma^2$. The displacement $D = RS$, $M$ is the vector of the multiplication of $R$ and $s$, and $I$ is the identity matrix.

The EM algorithm is applied herein as some correspondences in the matching stage may be missing due to occlusions and illumination change. To ensure systematic efficiency, this EM algorithm can be further simplified: In the E-stage, the parameters to be used in the M-stage will be roughly optimised, given initial motion and shape estimations. Especially, we pay more attention to the probability of the weights $w([w_1, ..., w_i])$ over the available $F$ and $\phi$:

$$p(w|F, \phi) = N([w, 1]^T - F^{-1} \theta^T; \sigma^2 I),$$

(19)

and $\theta = [F_{i+1} - F_{i-1} - D_{i+1-1}, 1]$.

In the M-stage, we have a motion estimation form similar to Eq. (14) but the corresponding parameters are estimated using the following equations: $\lambda_1 = \frac{||F_{i+1} - vec(RS_{i+1}) - D_{i+1}||^2}{2\sigma^2}$, and $\lambda_2 = 2ab \log \sqrt{2}\pi\sigma^2$, where $a$ and $b$ are two intermediate variables [16]. To update the estimated parameters, one can compute the partial derivative of the log likelihood with respect to each parameter, and then set it to zero before solving it. Fig. 3 demonstrates such a process that starts from an approximate estimate of the structure and then iterates in the proposed MCMC scheme to reach the final settlement.

**C. Algorithmic summary**

The proposed algorithm starts from corner feature detection in the first frame of an image sequence, followed by feature tracking using a standard tracker reported in [30]. Here, 150 corner features are extracted from the first image in order to guarantee that there are sufficient features to be tracked throughout the entire sequence. The features are gradually lost in the tracking procedure due to the absence of some features from the region of view. The proposed algorithm applied to two neighboring images is presented in Algorithm 2.

It is worth pointing out that EM is an iterative procedure that is very sensitive to initial conditions. Evidence shows that a poor starting of the structure/motion parameters will end up with significantly large errors when the solution is settled. Therefore, we need a good and efficient initialisation procedure. In the literature, a number of approaches have been proposed to successfully deal with this issue. For example, Coleman and Woodruff [9] introduced a classification technique that started from a random partition of a subsample and finished with a clustering of the subsamples, which was the starting point for EM. Similarly, Biernacki et al. [5] reported a scheme that used the parameters from a sequence of a classification EM algorithm, a stochastic EM algorithm and previous iterations of EM itself. As stated in [21], a good initial point for EM targets at the compromise of accuracy and computational costs. In our application, after the Markov random sampling process with randomly selected features, we will have well initialised structure/motion parameters before applying the generalised EM scheme to the overall corresponding features.

**D. Computational complexity analysis**

In this subsection, we use the following conventions for the analysis: $n_{im}$ is the number of image frames within a sequence, $n_F$ is the number of features used for corresponding two neighboring images, $n_g$ is the number of feature groups used within MCMC and $n_f$ is the number of features used within MCMC.
Algorithm 2 Outline of the proposed non-rigid shape and motion recovery algorithm.

1. Feature correspondence across two neighboring images.
2. Initialisation of structure and motion parameters.
3. Randomly choose 100 groups of feature points (each group has 50 features) from the overall correspondences.

repeat
   3.1. Employ the Metropolis-Hastings algorithm with the Gibbs sampling.
   3.2. Maximisation of \( p(R, T, \alpha, \sigma^2|P) \) using Eq. (3) and the associated equations.
until
4. Choose the feature group with its recovered structure and motion parameters, which lead to the minimum mean error.
5. Apply the generalised EM procedure to the overall feature correspondences using the outcomes of Step 4.
repeat
   5.1. In the E-step, compute the expectation \( E(w) \) of the residuals using Eq. (13).
   5.2. In the M-step, Eq. (14) is deployed with the updated shape basis, error variance, translation and rotation.
until
6. Stop the iteration if the average variance is less than a pre-defined threshold. Otherwise, go back to Step 3.

Fig. 4. Four frames of a synthetic sequence (“face sequence”). Image courtesy of Heriot-Watt University, UK.

Fig. 5. Some clips of a chess-board sequence, which are superimposed by the detected/tracked feature points (red squares).

Organised as follows: (1) Motion estimation and 3-D shape recovery are achieved using a synthetic sequence of a deforming shark provided in [32] and a synthetic face sequence created by ourselves. (2) Using publicly accessible image sequences, we evaluate the performance of the proposed algorithm, Zhou’s and Torresani’s approaches.

A. Synthetic images

The motion of the shark is a rigid rotation in combination with deformations generated by \( K = 2 \) basis shapes [32]. We apply average errors that are defined as the mean distance of the reconstructed point to the expected point divided by the size of the shape. To evaluate the proposed method in a noisy situation, we add two noise sources to the original data. One of them is (mean, variance) Gaussian type, and the other is (mean, variance) uniformly distributed noise. In the first case, the variance of noise levels ranges from 0 to 16 in steps of 2 whilst all means are zero (units: intensity levels). In the second case, to describe the degrees of added noise, we adopt signal-to-noise ratios. Tables I and II illustrate that with smaller noise levels these three algorithms possess comparable performance. However, as the noise levels increase (occlusions also occur), the proposed 3-D recovery algorithm has much smaller errors than Torresani’s and Zhou’s approaches. For example, the proposed method is of 3.48 against 5.41 of Torresani’s model and 3.95 of Zhou’s model in the case of uniform noise. Therefore, the proposed method is more robust than the others in different noise circumstances.

The second synthetic sequence, namely face sequence, was created via a ray-tracing technique. Some clips are shown in Fig. 4. The motion of the face has been known a priori. In order to evaluate the proposed shape recovery algorithm, we add Gaussian and uniformly distributed noise to the facial images individually. The noise levels are the same as illustrated in Tables I and II. From the experiments, the results can be summarised as follows: (1) In the case of Gaussian noise, the proposed method has the error range of (2.01, 15.77)\%, while Zhou’s method has the error range of (2.12, 16.40)\%, and Torresani’s method (2.09, 16.82)\%, and (2) in the case of uniformly distributed noise, the proposed technique results in the error range of (2.02,9.04)\%. Comparably, Zhou’s technique leads to an error range of (2.12,9.56)\%, while Torresani’s method has a larger error range (2.09, 13.48)\% compared to the other two methods. In summary, the proposed algorithm has better accuracy than these two classical techniques in this experiment.

Referred to Algorithm 2, the computational costs of the proposed MCMC scheme for an entire video sequence are as follows: (1) The MCMC step takes \( O((n_{im} - 1)n_gn_f^2) \), and (2) the original EM stage takes \( O((n_{im} - 1)n_f^2) \). Therefore, the entire proposed algorithm adds extra computational efforts for MCMC onto the top of the original EM and hence reaches a computational level of \( O((n_{im} - 1)n_gn_f^2 + (n_{im} - 1)n_f^2) \). Choosing smaller \( n_g \) and \( n_f \) will lead the computational costs to being dramatically reduced. However, this feature reduction may bring a negative impact on the optimality of the structure and motion parameters. In general, the more features and image frames are involved, the higher computational costs are required.

V. EXPERIMENTAL WORK

We compare the proposed algorithm against the state of the art techniques developed by Torresani et al. [32] and Zhou et al. [39],[40], respectively. The former has achieved successful performance in Gaussian estimates, while the latter pursued a solution towards the case of non-Gaussian distributions but did not apply any MCMC procedure. To seek a consistent estimation, the experimental work described in this section is

(a) (b) (c) (d)

(a) Frame 1 (b) Frame 2 (c) Frame 3
TABLE I
ERROR STATISTICS FOR DIFFERENT GAUSSIAN NOISE LEVELS. UNIT OF VARIANCE: INTENSITY – LEVEL².

<table>
<thead>
<tr>
<th>Variance</th>
<th>[32] (%)</th>
<th>[39] (%)</th>
<th>Proposed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no additive noise</td>
<td>1.46</td>
<td>1.47</td>
<td>1.41</td>
</tr>
<tr>
<td>2</td>
<td>2.98</td>
<td>3.17</td>
<td>2.9</td>
</tr>
<tr>
<td>4</td>
<td>5.68</td>
<td>5.17</td>
<td>4.94</td>
</tr>
<tr>
<td>6</td>
<td>7.72</td>
<td>7.05</td>
<td>6.53</td>
</tr>
<tr>
<td>8</td>
<td>8.84</td>
<td>8.11</td>
<td>7.46</td>
</tr>
<tr>
<td>10</td>
<td>10.05</td>
<td>9.11</td>
<td>8.66</td>
</tr>
<tr>
<td>12</td>
<td>11.24</td>
<td>9.98</td>
<td>9.23</td>
</tr>
<tr>
<td>14</td>
<td>12.01</td>
<td>10.73</td>
<td>9.5</td>
</tr>
<tr>
<td>16</td>
<td>12.99</td>
<td>11.48</td>
<td>10.22</td>
</tr>
</tbody>
</table>

TABLE II
ERROR STATISTICS FOR DIFFERENT UNIFORM NOISE LEVELS. UNIT OF SNR: dB.

<table>
<thead>
<tr>
<th>SNR</th>
<th>[32] (%)</th>
<th>[39] (%)</th>
<th>Proposed (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>no additive noise</td>
<td>1.46</td>
<td>1.47</td>
<td>1.43</td>
</tr>
<tr>
<td>35.35</td>
<td>1.86</td>
<td>1.84</td>
<td>1.68</td>
</tr>
<tr>
<td>29.33</td>
<td>2.50</td>
<td>2.41</td>
<td>2.25</td>
</tr>
<tr>
<td>25.84</td>
<td>3.13</td>
<td>2.92</td>
<td>2.36</td>
</tr>
<tr>
<td>23.31</td>
<td>3.65</td>
<td>3.24</td>
<td>2.79</td>
</tr>
<tr>
<td>21.12</td>
<td>4.27</td>
<td>3.37</td>
<td>2.94</td>
</tr>
<tr>
<td>20.03</td>
<td>4.52</td>
<td>3.45</td>
<td>3.17</td>
</tr>
<tr>
<td>19.17</td>
<td>4.87</td>
<td>3.69</td>
<td>3.24</td>
</tr>
<tr>
<td>18.26</td>
<td>5.41</td>
<td>3.95</td>
<td>3.48</td>
</tr>
</tbody>
</table>

### B. Real images

The ground truth of the real image sequences is not available and therefore the evaluation can only be based on the subjective judgments. In this paper, to provide a comprehensive evaluation of the involved algorithms, we categorise the testing results into two classes: The first one corresponds to the 2-D projections of the recovered structure, and then second one is related to the depth estimation. Three image sequences are here tested using the three approaches respectively but we will use image pairs to demonstrate the performance of the shape recovery. Note that missing data has been recognised in two of the image sequences due to the exit of the feature points from the camera view.

The first sequence is about a chess-board symbol printed on an subject’s back (Fig. 5) while he repeatedly rotates the trunk.

![Fig. 6. Performance comparison of the two methods in the chess-board sequence: (a)-(d) 3-D recovery by [32] using the correspondences between frames 1(or 2) and 2(or 3); (b)-(e) 3-D recovery by the proposed method using the correspondences between frames 1(or 2) and 2(or 3); (c)-(f) 3-D recovery by [39] using the correspondences between frames 1(or 2) and 2(or 3).](image)

This sequence was kindly provided by the authors of [17]. The chess-board actually is of non-rigid motion, as the worn shirt is soft and flexible. The experimental results, depicted in Fig. 6, show that the proposed algorithm has generated 3-D points of larger amplitudes along the z-axis (vertical to the image plane), compared to Torresani’s and Zhou’s methods. The feature points collected from the central areas of frames 1 and 2 actually hold different depths along the view direction of the camera. However, the estimation of Torresani’s and Zhou’s approaches does not reflect these depth variations, and hence ends up with relatively large errors.

Fig. 7 demonstrates the performance comparison of depth estimation from images 2 and 3 using different algorithms in this chess-board image sequence. The chess-board is nearly a plane, and therefore the recovered structure must be planar. In Fig. 7, the outcome of the proposed MCMC-EM algorithm demonstrates linear changes from blue to red colours, resulting in small errors in structure from motion, while the others sustain significant peaks in the recovered plane.

The second example is a hand sequence that is available from an unknown provider. The hand is tightly held and then released. Fig. 8 shows three example frames of the image sequence and the corresponding 3-D recovery is illustrated in Fig. 9. Similar to the previous example, the proposed method results in correct feature locations. Interestingly, the recovered 3-D points shown in Figs. 9(a)-(d), (b)-(e) and (c)-(f) are significantly different from each other in terms of their positions. In fact, Figs. 9(b)-(e) appear to follow the correct movements of the hand but the other sub-figures do not. Therefore, the proposed method has better accuracy in shape recovery than Torresani’s and Zhou’s methods.

![Fig. 7. Performance comparison of depth estimation by different algorithms in the chess-board image sequence.](image)

Fig. 7 illustrates the performance comparison of depth estimation from images 2 and 3 using different algorithms in this hand sequence. In some particular areas, the hand appears to be concave. Consequently, the recovered structure must correspond to this shape. In Fig. 10, the results of the proposed MCMC-EM algorithm simulates this hand movement, where the light blue colour indicates the concave of the hand. Unfortunately, the other two algorithms have less accuracy in the depth estimation, although Zhou’s method is better than Torresani’s algorithm in the sense that a concave is observed from the outcome of Zhou’s method.

The third example is a plastic sheet sequence provided by Salzmann et al., EPFL, Lausanne, Switzerland (see Fig. 11). The sheet is stretched and tilted to form different shapes at different angles. Fig. 12 shows the reconstruction comparison...
between our new method and Torresani’s model. Comparing Figs. 12(a) and (b), one can observe that the recovered surface shown in Fig. 12(b) is more flat and wider than that shown in Fig. 12(a). In terms of Figs. 12(d) and (e), the former is tilted at a totally different angle from the latter which matches the true rotation. This justifies that the proposed method correctly reflects the angle variation of the object but Torresani’s and Zhou’s approaches have worse performance in this example.

Fig. 13 shows the performance comparison of depth estimation from images 2 and 3 using the three algorithms in this plastic sheet sequence. In most circumstances, the plastic sheet appears to be planar. Therefore, the recovered structure must be planar. In Fig. 13, the results of the proposed MCMC-EM and Zhou’s algorithms share a very similar structure, while Torresan’s algorithm possesses larger variations on the recovered surface. Looking closely at these resulting images, it is observed that Zhou’s method has a larger error range than the proposed MCMC-EM algorithm, indicating that the proposed algorithm has the best capability among the three algorithms in this example.

C. Applications in event recognition

In this subsection, we address the potential applications of the proposed structure and motion recovery scheme in event recognition. Event recognition is to recognize which events are occurring in a video stream. As an example, we here focus on the hand sequence, where the state changes during the period of hand holding and releasing need to be detected. In a 2-D based classification system, one has to design a classification system to measure the changes of the marked hand area. However, it is not easy to identify whether or not these changes are due to hand folding/releasing, mutual occlusion or pose change. In a 3-D based classification system, this uncertainty can be dealt with if one follows the variations of the angle \( \psi \) formed by three points, which are extracted from the marked area. The 3-D solution does not require any prior knowledge of the hand motion (e.g. planar or non-planar). Instead, it only needs to monitor the state change of the angle \( \psi \) during the entire session (i.e. from holding to releasing or vice versa).

Fig. 14 illustrates an example for event detection (e.g. hand holding/releasing). After a hand has been detected using the well-known OpenCV package, the proposed structure from motion algorithm is used to provide the angle \( \psi \) measurements using the area centre and other two surrounding feature points. It is observed that the detected events well match the true situations, except the case where the structure/motion scheme becomes less efficient to catch up the transition from holding to releasing. This is because it is extremely difficult to identify which frame out of images 16-18 is the turning point due to the unchanged angle.

D. Further discussion

One of the main contributions of the proposed MCMC-EM algorithm is the capability of effectively handling the 2-D observations without assuming the distribution characteristics. In the synthetic cases, we have run the proposed algorithm over the image data with additive Gaussian or non-Gaussian noise. In the real datasets, missing data and non-Gaussian distributions continuously challenged the proposed algorithm. For example, in the hand sequence, the marks on the hand gradually disappeared and re-appeared as the hand folded and then released. On the other hand, in the chess-board
sequence, since all the regions of the chess-board are of similar texture, the used tracking system was reluctant to find correct correspondences from multiple candidates in the matching stage. Regardless of these difficult circumstances, however, the proposed algorithm has achieved better performance than the other two state of the art approaches. This attributes to the fact that the proposed approach used an MCMC procedure plus EM refinements, which had been applied to general distributions and did not require the log-concavity condition for the final solution.

VI. CONCLUSIONS AND FUTURE WORK

We have presented a solution to the non-rigid shape and motion recovery problem. This new algorithm has allowed us to perform more accurate 3-D structure recovery than two classical approaches in different distributions. A novel MCMC-EM algorithm was proposed without assuming the statistical properties. By addressing the orthogonal reconstruction of the structure in a constrained minimization function, we can directly render the 3-D shape, motion parameters and variance values, based on the rigid motion using Tomasi-Kanade’s rigid structure-from-motion factorization. Numerical experiments have demonstrated that the proposed method produces better results from clean or noisy synthetic data, and real image sequences. An application of the proposed algorithm in event recognition was also introduced.

In spite of its success, this approach has an obvious limitation: Gibbs sampling in the computation takes inevitable computational efforts. To address this weakness in the future work, we are working towards an optimisation technique to reduce the number of random sampling, which may compromise the convergence accuracy and efficiency in due course.

ACKNOWLEDGMENTS

This work was in part supported by the National Basic Research Program of China (973 Program) (Grant No. 2012CB316404), the National Natural Science Foundation of China (Grant No. 61072093) and the Open Project of State Key Laboratory of Industrial Control Technology (Zhejiang University) (Grant No. ICT1105). The authors would like to thank the handling editor and the anonymous reviewers for their valuable and constructive comments. The unknown provider of the hand sequence and other image creators in the experimental work are also appreciated.

REFERENCES


Huiyu Zhou received his BEng degree in radio technology from Huzhong University of Science and Technology of China in 1990. He was awarded an Master of Science degree in biomedical engineering from the University of Dundee, and an PhD degree in computer vision from the Heriot-Watt University, Edinburgh, United Kingdom. Currently, he holds a permanent post in ECT, Queen’s University Belfast, United Kingdom. His research interests include computer vision, human motion analysis, intelligent systems and human computer interface. He has published widely on these topics.

Xuelong Li (M’02-SM’07) is a professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi’an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi’an 710119, Shaanxi, P. R. China.

Abdul H. Sadka (SM’07) is chair, head of Electronic and Computer Engineering and the director of the Centre for Media Communications Research of Brunel University West London with 15-year experience in academic leadership and excellence. He received his PhD in Electrical and Electronic Engineering from Surrey University UK in 1997. He is an internationally renowned expert in visual media processing and communications with an extensive track record of scientific achievements and peer recognised research excellence. He has managed so far to attract nearly 2.5M worth of research grants and contracts in his capacity as principal investigator. He has published widely in international journals and conferences and is the author of a highly regarded book on “Compressed Video Communications” published by Wiley in 2002. He holds three patents in the video transport and compression area. He acts as scientific advisor and consultant to several key companies in the international Telecommunications sector. He is a Fellow of IET and senior member of IEEE.