Relative Epipolar Motion of Tracked Features for Correspondence in Binocular Stereo

Hao Du, Danping Zou, and Yan Qiu Chen
Department of Computer Science and Engineering
Fudan University, Shanghai, China
{duhao|dpzou|chenyq}@fudan.edu.cn

Abstract

Most 3D reconstruction solutions focus on surfaces, and there has not been much research attention paid to the problem of reconstructing 3D scenes made up of large numbers of particles, while the ability to reconstruct such dynamic scenes is potentially very useful in many areas such as colony behavior research and visual modeling.

This paper proposes an approach - Relative Epipolar Motion (REM) - towards solving the correspondence problem in stereopsis by utilizing the motion clue. It matches feature trajectories instead of the features themselves as used by existing methods. The proposed method has the following new capabilities: (1) It supports reconstructing dynamic 3D scenes of large number of undistinguishable drifting particles; (2) It is applicable to correspondence establishment for dynamic surfaces made up of repetitive textures; (3) It offers an alternative way to project structured light in active mode for deforming surface reconstruction. Experiment results on both simulated and real-world scenes demonstrate its effectiveness.

1. Introduction

Our environments are full of scenes containing drifting objects, and the 3D reconstruction of such scenes are of high values. Examples include: (1) dynamic fish schools, bird flocks, bee swarms which are important subjects for behavior research [12]; (2) interactive microscopic biological particles such as enzymes, proteins, viruses, cells which are of key significance to biologists[10]; (3) falling snowflakes, exploding fireworks which are appealing scenes for visual modeling[16]. By reconstructing such dynamic particle-like scenes we mean obtaining the time-varying 3D coordinates of the particles. This poses a challenging problem different from the well-studied problem of reconstructing 3D surfaces.

Many shape from X (shading, textures, laser, structured light, motion, etc) methods in the literature that work well for surface reconstruction will encounter difficulties in particle reconstruction. Scenes containing large numbers of drifting particles provide enough salient yet nearly identical features, leading to lack of clues for correspondence establishment using existing stereo methods. Moreover, such scenes are laser, structured light and shading unsolvable. Structure from motion also fails since the scene is non-rigid. Although there exists methods for the 'trajectory-based video synchronization' problem [3, 2] that track and match such undistinguishable scene features, they are only able to tackle a few tens of feature trajectories and can not support a full 3D reconstruction task since they apply a relatively weak constraint for trajectory matching due to the synchronization and calibration information being unknown.

In this paper, we propose to utilize a motion clue, termed *relative epipolar motion* (REM), towards solving the correspondence problem in the binocular stereo setting (it can be extended to multi-view stereo). The method is based on the observation that, during a time-span in which the particles move, (1) a genuine matching of feature points of the same particle satisfies the epipolar constraint at every frame, (2) a genuine matching of feature points possesses the same motion velocity component perpendicular to the epipolar lines in the rectified common plane at every time instance. Experiments using artificial and real-world scenes containing drifting particles demonstrate the effectiveness of our method in establishing correspondence.

The proposed method can also be used for the widely studied surface reconstruction problem. In passive mode, it can establish correct correspondence of the features from repetitive natural textures. In active mode, it captures deforming textureless surfaces by projecting mono-colored particle-like active patterns, which is less affected by surface reflectance properties, and is reliable for reconstructing complex scenes with isolated parts. Experiments on surface reconstruction in passive-mode and active-mode demonstrate the applicability of the proposed REM method.

The rest of the paper is organized as follows. Section 2 lists related works on 3D reconstruction including stereo methods and monocular methods. Section 3 describes the motivation and principle. Section 4 presents the proposed method. Section 5 gives the experiment results. Section 6 discusses its characteristics. Conclusions are drawn in Section 7.

2. Related Works

Stereo 3D reconstruction methods involve feature matching and triangulation computation. The key difficulty lies in the former. Many works on passive matching (e.g. area-based correlation, dynamic programming matching[6], graph-cut matching[9], hierarchical approaches[5], trinocular and multi-view methods[6], etc) only consider single pair of images at one time-instance. These approaches make use of pixel or local texture information, as well as smooth constraints of object surfaces. False correspondences are likely to occur in dealing with complex scenes containing objects with isolated parts and repetitive textures.

Structured light methods use coded light to establish correspondence between stereo cameras or between the projector and the camera. Such methods include one-shot techniques by projecting color-coded[17], phase-coded[20] light patterns, and space-time[4, 18] methods by projecting active patterns that change over time and estimating a linear spatial-temporal window. One-shot methods tend to be affected by the reflectance properties of the target surfaces, while space-time methods require the projectors to have high refresh-rates. Another limitation with structured light is that the added lighting may be nuisance or may even be infeasible when the objects are at large distances or are photophobic.

In monocular methods, the goal is to seek simpler settings, easy calibration, etc. Shape from shading[19] and shape from defocus[11] are methods free of the stereo correspondence problem, yet, the former fails if the surfaces are not smooth, and the latter produces relatively low accuracy in depth estimation. Structure from motion[8] can recover 3D structure of rigid scenes by a moving video camera, and good dense results [1, 15] have been reported, however, these methods are limited to rigid scenes, although recent research attempts to deal with dynamic scenes by detecting and segmenting out several rigid independent moving parts[13].

There remains the challenge to reconstruct dynamic scenes made up of large numbers of drifting particles. Also difficult are dynamic scenes consisting of repetitively textured deforming surfaces and isolated parts.

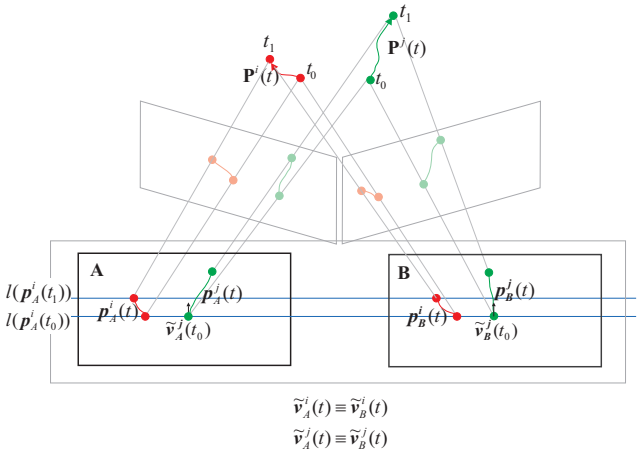


Figure 1. Relative epipolar motion. Two scene particles $\mathbf{P}^i(t)$, $\mathbf{P}^j(t)$ are moving over a time span. Let A, B be the rectified camera retinas[7]. At time t_0, the projections of $\mathbf{P}^i(t_0)$, $\mathbf{P}^j(t_0)$ on B, $\mathbf{p}^i_B(t_0)$, $\mathbf{p}^j_B(t_0)$ both lie on the associated epipolar line of $\mathbf{p}^i_A(t_0)$, denoted by $l(\mathbf{p}^i_A(t_0))$, resulting in multiple candidates for the correspondence of $\mathbf{p}^i_A$. Motion clue #1: When these scene particles move to new positions $\mathbf{P}^i(t_1)$, $\mathbf{P}^j(t_1)$ at time t_1, the correspondence for $\mathbf{p}^i_A \leftrightarrow \mathbf{p}^i_B$ is successfully established because the false candidate $\mathbf{p}^j_B$ moves out of the associated epipolar line $l(\mathbf{p}^i_A(t_1))$ while the real one still remains on the line. Motion clue #2: the velocities of feature points from a true matching $\mathbf{p}^i_A \leftrightarrow \mathbf{p}^i_B$ or $\mathbf{p}^j_A \leftrightarrow \mathbf{p}^j_B$ maintain the property that their vertical components perpendicular to the epipolar lines remain the same at every time instance, denoted by $\tilde{v}^i_A(t) \equiv \tilde{v}^i_B(t)$, $\tilde{v}^j_A(t) \equiv \tilde{v}^j_B(t)$.

3. Motivation and Principle

The proposed method makes use of two relative epipolar motion clues.

Motion clue #1: The basic principle towards utilizing the *relative epipolar motion* is based on the observation that, given a moving scene point, its two projections on the binocular retinal planes always stay on the same epipolar plane containing the scene point no matter how it moves, i.e. in the rectified common plane [7], the two projections always lie on the same epipolar line, as illustrated in Figure 1. Existing matching methods using single image pair come across ambiguity if more than one moving scene particle once coincide on the same epipolar line, while these points may move to new positions where their epipolar lines separate, and the matching ambiguity vanishes. Figure 1 illustrates this situation.

Motion clue #2: For a pair of feature points of a true matching, the vertical velocity components perpendicular to the epipolar lines on the rectified common plane remain the same at every time instance. Although this is deducible from motion clue #1, it has independent importance in real tasks. Due to distortion from lens and sensor planes, and the inaccurate feature detection, a width to tolerate the resultant epipolar line inaccuracy is needed, thus for matching ambi-

guities that cannot be overcome by the width, motion clue #2 is helpful.

The above observation motivates us to track the trajectory of each feature point and then to establish correct correspondence of two trajectories, each on one of the two retinal planes, by examining (1) whether the two moving feature points making up the trajectories are on the associated epipolar lines at every frame, and (2) whether the velocity component perpendicular to the epipolar lines are similar enough. Once the correspondence for trajectories is established, feature points that make up the trajectories are matched.

4. The Method

In this section, we formulate the 3D particle reconstruction problem and present our method, termed *relative epipolar motion*(REM).

4.1. The Problem of 3D Particle Reconstruction

Suppose $\mathcal{P} = \{\mathbf{P}^k(t)\}$, $\mathbf{P}^k(t) \in \Re^3$ is the set of particles drifting in the world space. They are captured at discreet time instances by a pair of video cameras with intrinsic and extrinsic parameters calibrated. Let A, B be the rectified image planes of the two cameras, and the particles project onto A, B as feature points denoted by $\mathcal{P}_A = \{\mathbf{p}_A^k(t)\}$, $\mathcal{P}_B = \{\mathbf{p}_B^k(t)\}$, where each $\mathbf{p}_\Pi^k(t), \Pi = A|B$, forms a trajectory over the time sequence. The task is to reconstruct the original $\mathcal{P}$ given the observations and the calibration parameters.

Several factors make the extracted trajectories in $\mathcal{P}_\Pi$ less than perfect. Firstly, some observed trajectories have to be divided into segments to eliminate tracking ambiguity when collisions of scene particles or occlusions of projected feature points occur. Secondly, coordinate errors exist because the detected centroid of a feature point in 2D-projection image may not coincide with the actual centroid of the particle, especially when the scene particle is non-regular in shape and its projection takes a region covering several pixels in the image. Other aspects such as the particles' moving in-and-out of the camera image plane, and uncalibratable geometrical distortions caused by the lens also reduce the observation accuracy.

Suppose there are a number of n_Π extracted trajectories on the rectified image plane Π. Each trajectory $\mathbf{T}_\Pi^i$ has its start time δ_Π^i, end time ζ_Π^i and image coordinate $\mathbf{q}_\Pi^i(t) = (x_\Pi^i(t), y_\Pi^i(t))^T$ on Π during the time span:

$$\mathbf{T}_\Pi^i \quad = \quad \big(\mathbf{q}_\Pi^i(t), \delta_\Pi^i, \zeta_\Pi^i\big), \tag{1}$$

where,

$$t \in [\delta_\Pi^i, \zeta_\Pi^i], \qquad i = 1, 2, ..., n_\Pi, \qquad \Pi = A|B. \tag{2}$$

The common time span $\eta(i, j)$ of trajectories $\mathbf{T}_A^i$ in A and $\mathbf{T}_B^j$ in B is:

$$\eta(i, j) = \{t | \max(\delta_A^i, \delta_B^j) \le t \le \min(\zeta_A^i, \zeta_B^j)\}. \tag{3}$$

In addition, local descriptions of a feature point such as color, intensity or texture might be available, thus, let the window $w_\Pi(\mathbf{q}_\Pi^i(t))$ be a local image patch on Π at time t at the coordinate $\mathbf{q}_\Pi^i(t)$.

Notice that, in the above statement, the epipolar constraint tells that if feature points $\mathbf{q}_A^i(t)$ and $\mathbf{q}_B^j(t)$ are a genuine match at time t, we have

$$|y_A^i(t) - y_B^j(t)| < \varepsilon, \tag{4}$$

where y_Π^i is the vertical coordinate of $\mathbf{q}_\Pi^i$, and ε is a width to tolerate the inaccuracy.

4.2. Matching Score between Trajectories

The proposed method matches feature trajectories extracted from the image sequence instead of the feature points between one pair of images at one time as adopted by many existing stereo methods.

A matching score between a pair of trajectories is defined by combining the motion clues and local texture description,

$$s = \alpha s_1 + \beta s_2 + \gamma s_3, \tag{5}$$

where s_1, s_2 are the scores relating to two motion clues #1,#2, and s_3 is the score relating to texture description, α, β, γ are the weights to balance the contributions of each term. The matching score describes the similarity between two trajectories with smaller score indicating higher similarity.

The matching score between two trajectories that share no common time span is set to infinity,

$$s(\mathbf{T}_A^i, \mathbf{T}_B^j) = \infty, \qquad \text{if } \eta(i, j) = \emptyset, \tag{6}$$

We then only consider trajectories that share common time span.

The first score s_1 is based on motion clue #1 that, during a time span when particles move, a true matching of their projections (features) satisfies the epipolar constraint at every time instance. If at any time, the features of candidate trajectory pair $\mathbf{T}_A^i$ and $\mathbf{T}_B^j$ go to epipolar planes with large difference, $s_1(\mathbf{T}_A^i, \mathbf{T}_B^j)$ would be large, indicating that the two trajectories are not likely to be genuine matching. Let

$$e(\mathbf{T}_A^i, \mathbf{T}_B^j) = \max_{t \in \eta(i,j)} \{|y_A^i(t) - y_B^j(t)|\}. \tag{7}$$

Score s_1 is defined as,

$$s_1(\mathbf{T}_A^i, \mathbf{T}_B^j; \varepsilon) = \begin{cases} e(\mathbf{T}_A^i, \mathbf{T}_B^j) & e(\mathbf{T}_A^i, \mathbf{T}_B^j) < \varepsilon \\ \infty & otherwise \end{cases}, \tag{8}$$

where ε is the threshold to tolerate the inaccuracy.

The second score s_2 derives from the velocity. On the rectified projection planes A and B, the vertical velocity components perpendicular to the epipolar lines of two genuinely matched features remain similar, thus,

$$s_2(\mathbf{T}_A^i, \mathbf{T}_B^j) = \frac{1}{|\eta(i,j)|} \sum_{t \in \eta(i,j)} |\widetilde{v}_A^i(t) - \widetilde{v}_B^j(t)|^2, \quad (9)$$

where,

$$\begin{cases} \widetilde{v}_A^i(t) = y_A^i(t) - y_A^i(t - \triangle t) \\ \widetilde{v}_B^j(t) = y_B^j(t) - y_B^j(t - \triangle t) \end{cases}, \quad (10)$$

and $|\eta(i,j)|$ is the length of the common time-span for normalization.

The third score s_3 describes the similarity of the local image areas of the features if local texture is available,

$$s_3(\mathbf{T}_A^i, \mathbf{T}_B^j) = \frac{1}{|\eta(i,j)|} \sum_{t \in \eta(i,j)} C\big(w_A(q_A^i(t)), w_B(q_B^j(t))\big), \quad (11)$$

where C is some correlation or similarity measure of texture patches, with small value indicating high similarity.

From Equations (5,6,8,9,11), the final integrated matching score s between two trajectories is,

$$s(\mathbf{T}_A^i, \mathbf{T}_B^j) = \begin{cases} \alpha s_1(.;\varepsilon) + \beta s_2(.) + \gamma s_2(.) & \eta(i,j) \neq \emptyset \\ \infty & otherwise \end{cases}. \quad (12)$$

4.3. Correspondence Establishment

After obtaining the matching scores between trajectories, the scene can be reconstructed frame-by-frame. The key is to establish the correspondences between trajectory pairs at each frame, then feature points making up these trajectories at that frame are matched and the 3D scene can be reconstructed by stereo triangulation.

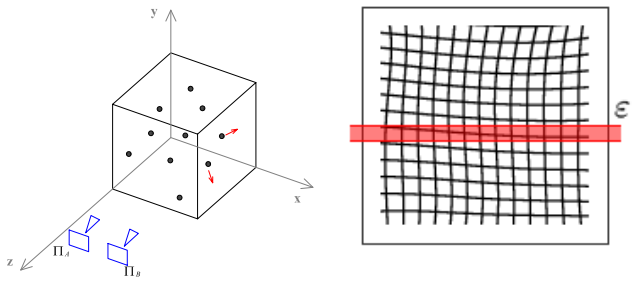
The correspondence matching can be formulated as a *maximum weighted bipartite matching* problem[14] with trajectories $\mathbf{T}_\Pi^i$ be the nodes and matching weights ρ be the edges, which is defined for each frame t,

$$\rho(\mathbf{T}_A^i, \mathbf{T}_B^j; t) = \begin{cases} e^{-\lambda s(\mathbf{T}_A^i, \mathbf{T}_B^j)} & t \in \eta(i,j) \\ 0 & otherwise \end{cases}, \quad (13)$$

where $\lambda > 0$ is a parameter controlling the slope of the exponential.

The objective is to maximize the total weight,

$$h(t) = \sum_i \rho(\mathbf{T}_A^i, g(\mathbf{T}_A^i); t), \quad (14)$$



(a) (b)

Figure 2. The simulation settings. (a) N particles drifting in a bounding unit cube with a limited velocity and acceleration, are captured by a pair of virtual video cameras. (b) A tolerance width ε is set to accept the distortion caused by lens and feature detection errors.

subject to the constraint that the matching is one-to-one, where $g(T_A^i)$ denote the resultant match of T_A^i.

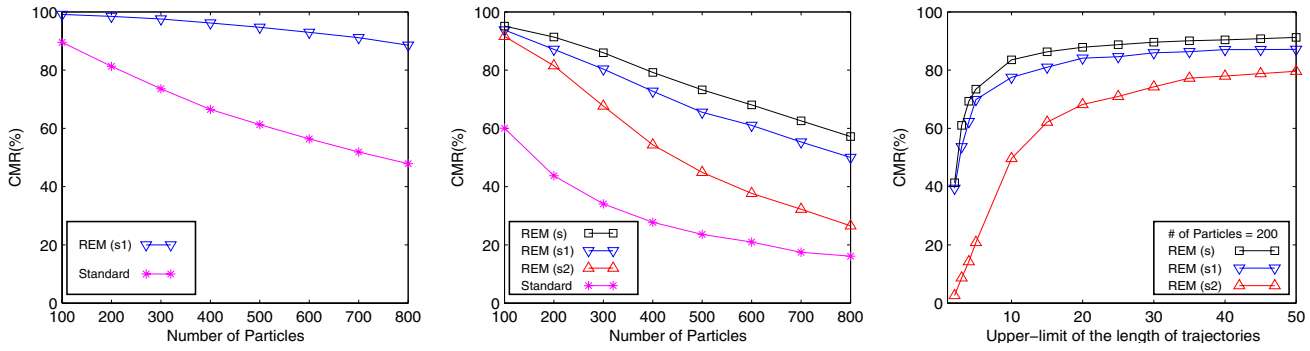
We add "dummy" nodes to each trajectory set with a constant weight ϵ_d. In this case, a trajectory will be matched to a "dummy" whenever there is no true matching available at higher weight than ϵ_d, which happens when the true matching is out of the image plane of the other camera, or it is an outlier. Similarly, since the input to the algorithm should be a square weight matrix, "dummy" nodes are added to the smaller trajectory set.

Notice that, in Equation (13), the matching weight relating to a trajectory with time-span out of the currently considered frame t is set to zero, i.e., they are ignored and can be extracted from the node set before running the algorithm. In addition, the number of edges E can be reduced by holding a small number of potential matching candidates with high weights, which may sacrifice the matching accuracy a little.

This matching process can be solved in $O(V^2E)$ time using modified Bellman-Ford algorithm, or in $O(V^2log(V) + VE)$ time with the Dijkstra algorithm.

5. Experiments

The effectiveness and performance of the proposed *Relative Epipolar Motion* (REM) method is to be tested by a simulation and several real-world 3D reconstruction tasks. The imaging system includes two synchronized digital video cameras operating at the 800×600 resolution and 25Hz frame-rate, and an LCD projector working at the 1024×768 resolution. For camera geometric calibration, Zhang's method [21] is used with a printed chessboard pattern $37.5cm \times 37.5cm$ pasted on a flat acrylic sheet.



(a) (b) (c)

Figure 3. The result of the simulation. (a) In the ideal case with zero distortion, given different number of particles, the proposed REM method shows good performance while the CMR (correct-matching-rate) using standard matching method decreases rapidly. (b) Under a 5% projection distortion, both the two motion clues REM(s_1), REM(s_2) perform better than the standard method, and their combination REM(s) performs even better. (c) Under a 5% projection distortion, for 200 particles, the longer the length of tracked trajectories, the better the performance of REM method.

5.1. Simulations

This simulation is to evaluate the correspondence establishment ability of the proposed method. The classical matching method based on single image pairs using the epipolar constraint is implemented for comparison.

As shown in Figure 2(a), the scene contains N particles drifting in a bounding unit cube. Each particle $\mathbf{p}^i$ is initiated at a random location in the cube, with a random starting velocity $\mathbf{v}_0^i$ and a small random acceleration $\mathbf{a}^i(t)$ at each time step t,

$$\mathbf{v}_0^i \sim \mathcal{N}(\mathbf{0}, 0.05I) \qquad \mathbf{a}^i(t) \sim \mathcal{N}(\mathbf{0}, 0.001I), \tag{15}$$

where $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes that $\mathbf{x}$ obeys a normal random distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. The particle rebounds when hitting the walls.

The scene is captured by two virtual video cameras at resolution of 800×600, i.e., the particles are perspectively projected and quantized onto the image planes of the virtual cameras. A random distortion after the projection is applied to simulate the inaccuracy caused by uncalibratable lens, retina distortions and feature detection errors, as shown by Figure 2(b), where large distortions require a large tolerance width ε in Equation (8,12).

In the above setting, the projected feature points are moving at a speed no more than 4 pixels per frame. During tracking of the feature trajectories, if more than two feature points move too close (less than 8 pixels), the trajectory is divided to eliminate tracking ambiguity.

We record a video sequence of 200 frames and the various number of particles N ranges from 100 to 800.

The first set of simulations is in the ideal case with no projection distortion. Table 1 shows the number of tracked trajectory pieces on the two rectified image planes with their average length (only trajectories with $length \geq 5 frames$ are considered). Figure 3(a) is the result represented in correct-matching-rate (CMR) of the proposed REM method using motion clue $\#1$ (score s_1) as compared to standard matching method using only epipolar constraint considering single image pairs.

Table 1. The number of tracked trajectories and their average length given different number N of points. Only trajectories with $length \geq 5 frames$ are recorded.

N	100	200	300	400	500	600	700	800
n_A	254	698	1391	2213	3297	4314	5489	6594
n_B	241	725	1384	2218	3288	4331	5535	6555
$\bar{l}$	79.2	54.6	41.1	33.3	27.4	24.4	21.9	20.1

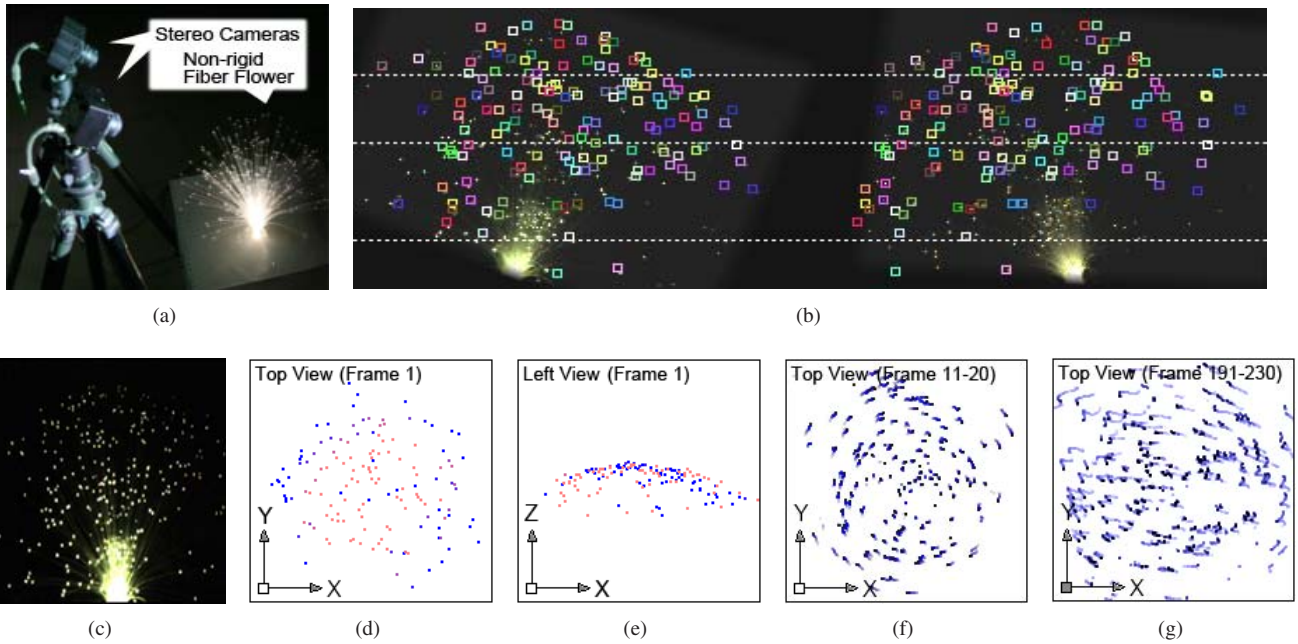
N : Number of particles in the scene.
n_Π : Number of tracked trajectories on the projection plane Π. ($\Pi = A|B$)
$\bar{l}$: The average length of trajectories.

Then, we test the performance of the methods given a considerable distortion of 5%, which means a tolerance width of $\varepsilon = 30\ pixels$ is required. As shown in Figure 3(b), the proposed REM method using both the motion clue $\#1$ and $\#2$ performs better than the standard matching method, and the REM method using integrated score (s) performs the best.

Third, the effect of the trajectory length is tested. Given 5% distortion, we fix the number of particles $N = 200$, and set the upper limit of the length of tracked trajectory pieces from 1 to 50. As shown in Figure 3(c), the CMRs using REM method with motion clues $\#1$, $\#2$, and the both increase with the length of trajectories. At the extreme case when all length of trajectories are limited to 1, the proposed REM method reduces to standard matching method.

5.2. A Non-Rigid Waving Optical Fiber Flower

This experiment evaluates the ability of the proposed method in establishing correspondence for the reconstruc-



(a) (b) (c) (d) (e) (f) (g)

Figure 4. The reconstruction of an non-rigid optical fiber flower. (a) The experiment setting. (c) A sample captured image where more than 200 bright-spots are detectable. (b) The matching results of the 1^{st} frame using the proposed REM method shown in rectified image pair, where a pair of rectangles in the same color represents a resultant correspondence. (d) Reconstructed depth result from top-view (red to blue $\leftrightarrow$ near to far). (e) Reconstructed depth result from left-view. (f) Reconstruction result of Frame 1-10 from top-view, where the fiber flower is rotating. (g) Reconstruction result of Frame 191-230 from top-view, where the fiber flower is waving wildly.

tion of real-world drifting particles. Figure 4(a) shows the setting. A flexible waving optical fiber flower illuminated from the bottom is captured by a pair of calibrated cameras. Figure 4(c) is a sample captured image where more than 200 bright-spots at the end of optical fibers can be detected as salient features. The task is to locate the 3D positions of these moving spots over time.

The scene is captured for 256 frames with the flower waving. Each bright spot takes up an average region of 4 pixels in radius on the captured image. The tolerance width of the epipolar line is set to 5 pixels.

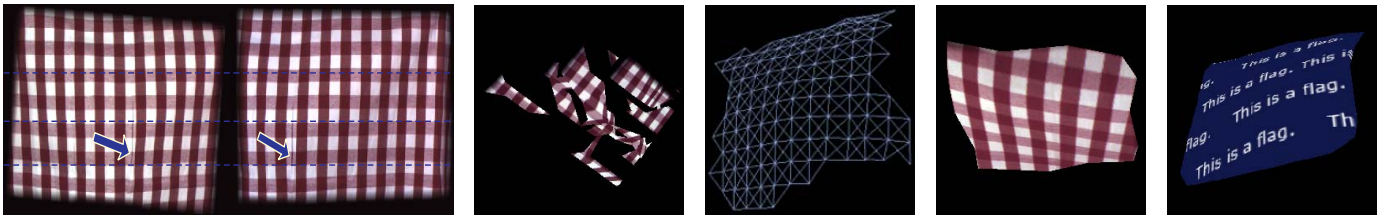
Classical matching methods considering only image pair at one time instance fail to establish the correspondence since more than one bright-spots appear on almost every epipolar line width. Smooth assumptions and area-correlation are also non-applicable in this situation. The proposed REM method offers good performance. Figure 4(b) shows the result of the 1^{st} frame pair using REM method, where the image pair is rectified and each pair of rectangles in the same color representing a resultant correspondence. Figure 4(d) 4(e) are the reconstructed depth map looking from top and left. Figure 4(f) is the reconstructed results of Frame 11-20 looking from top during which the optical fiber flower is rotating. Figure 4(g) is the reconstructed results of Frame 191-230 looking from top, where the flower is waving wildly.

5.3. A Waving Flag

This experiment is to reconstruct a waving flag making use of the trackable features from repetitive natural texture. The proposed method works in the passive mode.

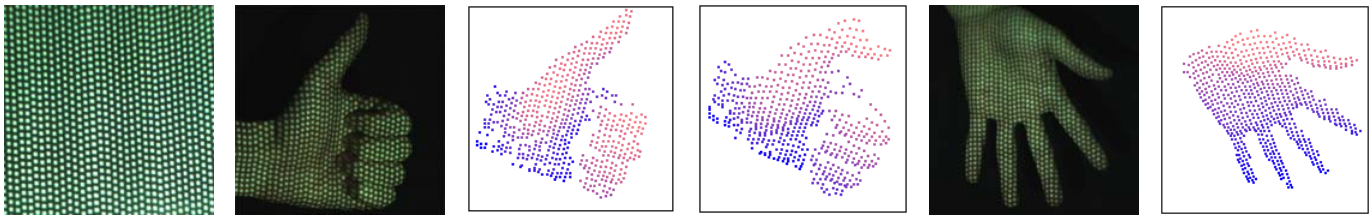
The flag contains regular patterns with trackable feature points (corners) on the surface, waving in wind. A video sequence of 256 frames is captured. Figure 5(a) shows the rectified [7] 1^{st} frame pair. It is difficult for existing algorithms, even for a human, to identify the correspondence based only on this one-shot frame pair, where a small tuck on the surface pointed by arrows indicates the underlying true correspondence. The matching and reconstruction result by dynamic programming based fusion technique[6] is shown in Figure 5(b). It can be seen that it did not work for this situation.

For the proposed REM method, we set the toleration width ε to 5 pixels. Although many feature points along an epipolar line are not distinguishable for matching in the 1^{st} frame, these points will go to separate epipolar line over time while the surface deforms, and the proposed method uses this information to establish correct correspondence. Figure 5(c) is the successfully reconstructed mesh for the 7^{th} frame using the proposed method. Figure 5(d) and 5(e) are another two frames (19^{th} and 37^{th}) reconstructed by the proposed method rendered with original and new texture.



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

Figure 5. The reconstruction of a piece of deforming flag waving in wind. (a) Rectified 1^{st} frame pair. A small tuck pair pointed by the arrows indicates the underlying true matching. (b) The reconstruction result by dynamic programming based fusion. (c) The mesh of the 7^{th} frame reconstructed using the proposed method. (d)(e) Another two frames (19^{th} and 37^{th}) reconstructed by the proposed method rendered with original and new textures.



(a) (b) (c) (d) (e) (f)

Figure 6. The reconstruction of a gesturing hand by projecting active particle-like pattern. (a) The active particle pattern in which each column moves vertically and independently. (b) The source image captured by the left camera at 1^{st} frame. (c)(d) The reconstructed point cloud of 1^{st} and 10^{th} frame shown in depth map. (e) Another source image at frame 370^{th} with fingers opened. (f) The reconstructed point cloud of 370^{th} frame.

5.4. A Gesturing Human Hand

This experiment shows an alternative way to reconstruct textureless surfaces by projecting active particle-like patterns. The scene surface can either be static or dynamic.

The active particle pattern is designed to possess two properties: (1) It creates independent relative epipolar motion for each particle; (2) The particles do not overlap during the motion so that they can be easily tracked. Figure 6(a) is a sample pattern projected on a flat sheet, where each column of particles independently moves a vertical step randomly chosen from $\{-1, 0, 1\}$ pixel at each time. Thus, after tracking n frames, it rarely happens that two particles reamin on the same epipolar lines over the whole time span.

The advantages of the active particle-like pattern as compared to traditional structured light of color strip patterns are: (1) Compared to color-coded patterns, it is less affected by the surface reflection properties since no recognition of the color is required; (2) Compared to non-coded patterns, it is more reliable when the scene is complex with occlusion and isolated surfaces, since it does not use smooth and order consistency assumptions in matching.

The target human hand deforms over a time span of 512 frames. Figure 6(b) is one captured image of the 1^{st} frame. Figure 6(c)6(d) are the reconstructed point cloud of 1^{st} and 10^{th} frame shown in depth map. Figure 6(e) is another captured image with five fingers opened. Figure 6(f) is the point cloud looking from a different viewpoint.

6. Discussions

We now discuss several characteristics of the proposed REM method.

• Motion clues provide essential information for correspondence establishment. Two scene particles may happen to coincide on the same epipolar line, leading to correspondence ambiguity. However, the probability of two features that remain unmatchable after tracking n frames tends to be low since it rarely happens that two moving particles remain on the same epipolar line during the whole time span if they are not strictly dependent.

• Tracking failures do not harm too much to the REM method. The proposed REM method matches trajectories based on motion clues as well as local texture features (if available). If tracking of trajectories fails, i.e. trajectories are of length equal to one frame, either because of the scene moving too fast or occlusion of features being severe, the REM method would reduce to traditional feature-to-feature matching.

• Static scenes can be reconstructed with a moving stereo-rig. For the reconstruction of completely static scenes, or when the motion is too small to provide enough relative epipolar motion information, an alternative implementation is to fix the calibrated camera pair onto a stick.

The relative epipolar motion of features can then be created by moving (rotating) the camera stick, which equals to the scene moving while fixing the cameras. A little swing of the stick creates significant relative epipolar motion information.

• Colored particle-like active pattern is an option. Experiment 5.4 shows an example to use only the relative epipolar motion clue by projecting mono-colored active particle-like pattern, however, if the surface reflectance properties is unanimous, Bayer-like pattern with more than one color could be applied, which both helps for trajectory tracking and matching.

• Trinocular and multi-view are supported. The proposed REM method can be extended to stereo settings using multiple calibrated cameras. The motions of features relative to epipolar lines of each pair of cameras collectively help to determine the correct correspondence.

7. Conclusion

We have developed an approach and an experimental system targeting the correspondence problem in binocular stereo setting for 3D dynamic scene reconstruction. The proposed method utilizes the relative epipolar motion of tracked features, offering a remarkable correspondence matching performance in reconstructing dynamic scenes containing trackable but undistinguishable features such as scenes containing large amount of drifting particles, which cannot be solved by using structured light and local texture information. The proposed method is applicable also to surface reconstruction in passive mode relying on repetitive natural textures with trackable but undistinguishable features, and in active mode by projecting particle like active patterns.

Acknowledgements

The research work presented in this paper is supported by National Natural Science Foundation of China, Grant No. 60575022; and Specialized Research Fund for the Doctoral Program of Higher Education, Grant No. 20050246063. The authors would like to thank Changyin Zhou, Jiajun Zhu and Stephen Lin for valuable discussions on this work.

References

[1] T. Brodsky, C. Fermuller, and Y. Aloimonos. Shape from video. In *Proc. CVPR*, 1999. 2

[2] R. Carceroni, F. Padua, G. Santos, and K. Kutulakos. Linear sequence-to-sequence alignment. In *Proc. CVPR*, volume 1, pages 746–753, 2004. 1

[3] Y. Caspi, D. Simakov, and M. Irani. Feature-based sequence-to-sequence matching. *International Journal of Computer Vision*, 68(1):53–64, 2006. 1

[4] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz. Spacetime stereo: A unifying framework for depth from triangulation. *IEEE Trans. Pattern Anal. Machine Intell.*, 27(2):296–302, 2005. 2

[5] U. Dhond and J. Aggarwal. Structure from stereo - a review. *IEEE Trans. Syst., Man, Cybern.*, 19(6):1489–1510, 1989. 2

[6] D. Forsyth and J. Ponce. *Computer vision : a modern approach*. Prentice Hall series in artificial intelligence. Prentice Hall, Upper Saddle River, N.J. ; London, 2003. 2, 6

[7] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, 2000. 2, 6

[8] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, Cambridge, UK ; New York, 2nd edition, 2003. 2

[9] V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *Proc. ECCV*, 2002. 2

[10] D. McDonald, M. A. Vodicka, G. Lucero, T. M. Svitkina, G. G. Borisy, M. Emerman, and T. J. Hope. Visualization of the intracellular behavior of hiv in living cells. *Journal of Cell Biology*, 159(3):441–452, 2002. 1

[11] S. K. Nayar, M. Watanabe, and M. Noguchi. Real-time focus range sensor. *IEEE Trans. Pattern Anal. Machine Intell.*, 18(12):1186–1198, 1996. 2

[12] K. S. Norris and C. R. Schilt. Cooperative societies in three-dimensional space: On the origins of aggregations, flocks, and schools, with special reference to dolphins and fish. *Ethology and Sociobiology*, 9(2-4):149–179, 1988. 1

[13] K. E. Ozden, K. Cornelis, L. Van Eycken, and L. Van Gool. Reconstructing 3d trajectories of independently moving objects using generic constraints. *Computer Vision and Image Understanding*, 96(3):453–471, 2004. 2

[14] C. H. Papadimitriou and K. Steiglitz. *Combinatorial optimization : algorithms and complexity*. Dover Publications, Mineola, N.Y., 1998. 4

[15] M. Pollefeys, L. V. Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232, 2004. 2

[16] C. B. Wang, Z. Y. Wang, T. Xia, and Q. S. Peng. Real-time snowing simulation. *Visual Computer*, 22:315–323, 2006. 1

[17] L. Zhang, B. Curless, and S. M. Seitz. Rapid shape acquisition using color structured light and multi-pass dynamic programming. In *Proc. Int. Symposium on 3D Data Processing Visualization and Transmission (3DPVT)*, pages 24–36, 2002. 2

[18] L. Zhang, B. Curless, and S. M. Seitz. Spacetime stereo: shape recovery for dynamic scenes. In *Proc. CVPR*, volume 2, pages 367–374, 2003. 2

[19] R. Zhang, P. S. Tsai, J. E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE Trans. Pattern Anal. Machine Intell.*, 21(8):690–706, 1999. 2

[20] S. Zhang and P. Huang. High-resolution, real-time 3d shape acquisition. In *Proc. IEEE Computer Vision and Pattern Recognition Workshops*, 2004. 2

[21] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *Proc. ICCV*, volume 1, pages 666–673, 1999. 4