ACCV2002: The 5th Asian Conference on Computer Vision, 23--25 January 2002, Melbourne, Australia.

Recovering Human Motions by Mobile Cameras and Factorization

Tomonori TABUSA*, Joo Kooi TAN**, Seiji ISHIKAWA**

**Department of Information Science*
Yuge National College of Maritime Technology
Shimoyuge 1000, Yuge, Ehime 794-2593, Japan
tabusa@info.yuge.ac.jp

***Department of Mechanical and Control Engineering*
Kyushu Institute of Technology
Tobata, Kitakyushu 804-8550, Japan

Abstract

This paper describes a technique for recovering human motion employing mobile video cameras. Human motion recovery is a demanding technique in many fields nowadays. One of the established techniques for the recovery is stereo vision. It cannot however recover wide range motions such as field athletics, since it employs fixed video cameras. In the proposed technique, multiple uncalibrated video cameras mounted on a mobile frame track a subject and capture his/her video images from different directions. These video images are employed for recovering motion of the subject by factorization. The motion recovery is realized by calculating at every sample time 3-D locations of the feature points specified on the subject. The recovered motion is what one can observe on the moving frame. In order to show performance of the technique, motion recovery is performed of a person playing basketball. Experimental results are shown and discussion is given.

1. Introduction

Three-dimensional human motion recovery has ever increasing needs in various fields. It is employed in human 3-D models creation in video games, virtual reality spaces, man-machine communication systems, or even in traditional skills or folk dance preserving in an electronic museum. Human motion analysis in a 3-D way is as well an expected technique for forms or movement analysis of various human activities including sports, working, rehabilitation, etc.

A motion capture system with magnetic sensors realizes real time motion recovery. It restricts motion of the subject, however, since he/she should remain in a limited magnetic field. The technique is therefore not applicable to recovery of a wide range motion such as field athletics. A motion that needs movement and/or panning of a video camera for its tracking is called a wide range motion in the present paper. Stereo vision [1] is another established technique for performing such motion recovery. It always necessitates calibration of the employed fixed cameras before use. This makes the technique difficult in recovering wide range motions, since there is no way of calibrating arbitrarily moving stereo cameras. Tan & Ishikawa [2,3] propose an optical 3-D motion recovery technique based on more than three uncalibrated cameras. This technique is employable in any place, only if one can take at least three image streams of the subject from different views. Having escaped from camera calibration, it still suffers from fixed cameras when one wants to apply it to wide range motions.

In the present paper, a technique is proposed for achieving 3D recovery of wide range human motions employing a mobile set of cameras. The technique proposed in the literature [2] is extended by a novel idea in defining a measurement matrix from obtained video images. In order to show performance of the technique, 3-D motion recovery is done of a person playing basketball in front of an image capturing frame which tracks the person.

The recovery technique based on uncalibrated video cameras [2,3] is introduced in 2, followed by the description of the proposed technique in 3. Experimental results are shown in 4. The presented technique is discussed in 5 and finally the paper is concluded.

2. Shape Recovery of Non-Rigid Objects by Uncalibrated Cameras

The technique proposed in [2,3] is overviewed which recovers 3-D shape of non-rigid objects by the

employment of uncalibrated multiple cameras at fixed locations and by factorization [4].

An object O in a 3-D space is taken images from F orientations by the same number of fixed cameras. Feature points $P_p(t)$ (p=1,2,…,P_t) are defined on the object O at sample time t (t =1,2,…,T). They must be commonly visible by the F cameras at time t and are projected onto the points ($x_{fp}(t)$, $y_{fp}(t)$) (p=1,2,…, P_t; t=1,2,…,T) on the image plane of the fth (f=1,2, …,F) camera. Correspondence of those projected feature points are examined among the F images obtained from the cameras at each sample time t and it is written into a measurement matrix $W(t)$ of the following form;

$$W(t)=\begin{pmatrix} x_{11}(t) & x_{12}(t) & \cdots & x_{1Pt}(t) \\ x_{21}(t) & x_{22}(t) & \cdots & x_{2Pt}(t) \\ \vdots & \vdots & \vdots & \vdots \\ x_{F1}(t) & x_{F2}(t) & \cdots & x_{FPt}(t) \\ y_{11}(t) & y_{12}(t) & \cdots & y_{1Pt}(t) \\ y_{21}(t) & y_{22}(t) & \cdots & y_{2Pt}(t) \\ \vdots & \vdots & \vdots & \vdots \\ y_{F1}(t) & y_{F2}(t) & \cdots & y_{FPt}(t) \end{pmatrix}, \quad (1)$$

where the rows represent camera f(f=1,2,…,F), and the columns correspond to the feature points $P_p(t)$ (p=1,2,…, P_t).

The matrices $W(t)$ (t=1,2,…,T) are merged into a single matrix of the form

$$W=\left(W(1)\,|\,W(2)|\ldots|\,W(T)\right), \quad (2)$$

which is called an extended measurement matrix.

The entire number of feature points denoted by Q is given by

$$Q=\sum_{t=1}^{T}P_t\,. \quad (3)$$

The world origin is transformed to the centroid of the Q feature points. The 3-D position of the feature point $P_p(t)$ is then denoted by vector $\boldsymbol{s}_p(t)$, and its projected position on the image plane is represented by $(\tilde{x}_{fp},\tilde{y}_{fp})$. This is displacement of the projected feature point from the projected centroid (x_f, y_f), *i.e.*,

$$\tilde{x}_{fp}(t)=x_{fp}(t)-x_f,\ \tilde{y}_{fp}(t)=y_{fp}(t)-y_f\ , \quad (4)$$

where

$$x_f=\frac{1}{Q}\sum_{t=1}^{T}\sum_{p=1}^{Pt}x_{fp}(t),\quad y_f=\frac{1}{Q}\sum_{t=1}^{T}\sum_{p=1}^{Pt}y_{fp}(t)\,. \quad (5)$$

Using Eq.(4), Eq.(1) is rewritten in the form

$$\tilde{W}(t)=W(t)-W(t)\cdot E_{Pt}/Q \quad (6)$$

or

$$\tilde{W}(t)=\begin{pmatrix} \tilde{x}_{11}(t) & \tilde{x}_{12}(t) & \cdots & \tilde{x}_{1Pt}(t) \\ \tilde{x}_{21}(t) & \tilde{x}_{22}(t) & \cdots & \tilde{x}_{2Pt}(t) \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{x}_{F1}(t) & \tilde{x}_{F2}(t) & \cdots & \tilde{x}_{FPt}(t) \\ \tilde{y}_{11}(t) & \tilde{y}_{12}(t) & \cdots & \tilde{y}_{1Pt}(t) \\ \tilde{y}_{21}(t) & \tilde{y}_{22}(t) & \cdots & \tilde{y}_{2Pt}(t) \\ \vdots & \vdots & \vdots & \vdots \\ \tilde{y}_{F1}(t) & \tilde{y}_{F2}(t) & \cdots & \tilde{y}_{FPt}(t) \end{pmatrix}, \quad (7)$$

Here E_{pt} is a P_t*P_t square matrix whose entries are all unity. Then we have

$$\tilde{W}=\left(\tilde{W}(1)\,|\,\tilde{W}(2)|\ldots|\,\tilde{W}(T)\right) \quad (8)$$

instead of Eq.(2).

Let the lens coordinate system of the fth video camera be denoted by three mutually orthogonal unit column vectors $\boldsymbol{i}_f$, $\boldsymbol{j}_f$, and $\boldsymbol{k}_f$ ($\boldsymbol{k}_f \equiv \boldsymbol{i}_f * \boldsymbol{j}_f$ being coincident with the light axis of the camera). Then, if one assumes that the imaging of objects through a camera lens is done orthographically, the followings hold;

$$\tilde{x}_{fp}(t)=(\boldsymbol{i}_f,\boldsymbol{s}_p(t)),\quad \tilde{y}_{fp}(t)=(\boldsymbol{j}_f,\boldsymbol{s}_p(t))\ . \quad (9)$$

Substitution of Eq.(9) into Eq.(8) yields

$$\tilde{W}=M\cdot S\,, \quad (10)$$

where

$$M=\left(\boldsymbol{i}_1,\boldsymbol{i}_2,...,\boldsymbol{i}_F,\boldsymbol{j}_1,\boldsymbol{j}_2,...,\boldsymbol{j}_F\right)^T, \quad (11)$$

and

$$S=\left(S(1),\,S(2),...,\,S(T)\right) \quad (12a)$$

$$S(t)=\left(\boldsymbol{s}_1(t),\,\boldsymbol{s}_2(t),\ldots,\,\boldsymbol{s}_{Pt}(t)\right). \quad (12b)$$

In this way, matrix $\tilde{W}$ given by Eq.(8) is decomposed into two matrices M and S. The former is called a camera orientation matrix giving the orientations of F cameras, whereas the latter is called a shape matrix which provides the 3-D coordinates of the Q feature points. Matrix $\tilde{W}$ is decomposed in practice by singular value decomposition and other linear computation algorithms [4].

It should be noted that, for the decomposition of Eq.(10), matrix $\tilde{W}$ of Eq.(8), therefore matrix W of Eq.(2), needs to be a full matrix. If there are any vacant entries in the matrix, the feature points corresponding to the columns containing the vacant entries cannot be employed for the motion recovery.

3. Employing Mobile Cameras

In the proposed technique, $F(\geq 3)$ cameras are used for image capture. They are fixed on a mobile frame and the frame traces the subject interested. **Figure 1** illustrates the

idea of mobile frame with fixed cameras. Since the cameras are fixed on the frame, the angles their light axes make take the values specified in advance. The frame itself can move along with the subject. The motion of the frame may contain rotation around a vertical line as well as parallel translation.

Let us specify some feature points on the subject concerned. Then tracking of their projected positions on F images of the cameras at time t yields the following matrix $W(t)$

$$W(t) \equiv \begin{pmatrix} W_x(t) \\ W_y(t) \end{pmatrix}. \qquad (13)$$

Here $W_x(t)$ is upper F rows and $W_y(t)$ is lower F rows of Eq.(1). Then the entire matrix W^{in} called an initial measurement matrix is defined as

$$W^{in} = \begin{pmatrix} W_x(1) & & & \\ & W_x(2) & & \\ & & \ddots & \\ & & & W_x(T) \\ \hline W_y(1) & & & \\ & W_y(2) & & \\ & & \ddots & \\ & & & W_y(T) \end{pmatrix}. \qquad (14)$$

Here the f 'th row of sub-matrix $W_x(t)$ and $W_y(t)$ correspond to the f 'th camera 's observation on the mobile frame at time t. It is denoted by $C_f(t)$ (f=1, 2,…,F; t=1,2,…,T). Gray rectangles contain acquired coordinates of the projected feature points, whereas other entries of the matrix remain vacant. Static feature points, if any, are regarded here for simplicity as different feature points at different sample times.

Equation (14) tells that the camera-mounted frame changes its location at every sample time. This formulation is done only for simplicity. The frame of course doesn't have to be moving all the time. It can naturally repeat stay and move. (Or even it can stay still at a certain location during observation, which is the case explained in Section 2.)

Equation (14) is obviously not a full matrix and therefore the shape recovery technique presented in Section 2 cannot be immediately applied to matrix W^{in} The idea of the proposed technique is to shift sub-matrices $W_x(t)$ and $W_y(t)$ (t =1,2,…,T, $t \neq s$ to an arbitrary camera locations $C_f(s)$ (f=1,2,…,F) at time s in matrix W^{in}. This yields a matrix W of the form

$$W = \begin{pmatrix} W_x(1) & W_x(2) & \cdots & W_x(T) \\ W_y(1) & W_y(2) & \cdots & W_y(T) \end{pmatrix}. \qquad (15)$$

Obviously this matrix is a full matrix and, after having been transformed into matrix $\tilde{W}$ by the procedure stated in the former section, it can be decomposed as

$$\tilde{W} = M \cdot R, \qquad (16)$$

where

$$M = (\boldsymbol{i}_1, \boldsymbol{i}_2, \boldsymbol{i}_3, \boldsymbol{j}_1, \boldsymbol{j}_2, \boldsymbol{j}_3)^T, \qquad (17)$$

and

$$R = (R(1), R(2), ..., R(T)), \qquad (18)$$

$$R(t) = (\boldsymbol{r}_1(t), \boldsymbol{r}_2(t), ..., \boldsymbol{r}_{Pt}(t)). \qquad (19)$$

Here t=1,2,…,T and $\boldsymbol{r}_p(t)$ (p=1,2,…,P_t) are the feature points on the subject observed at sample time t.
Equation (18) therefore gives recovered 3-D motion. This is a relative motion observed by the F cameras at the s'th location of the mobile frame.

4. Experimental Results

In order to show performance of the proposed technique, an experiment is performed in which a person playing with a basketball is captured images by a mobile frame with three mounted video cameras. The frame has the form of a triangular prism and the video cameras are fixed at each corner of the upper triangle (See **Figure 1**). The central angle of the triangle is 120 degrees and the angle the light axes of the adjacent video cameras make is approximately 40 degrees. The adjacent video cameras are distant from each other by 2 meters.

Several markers are placed on the subject as feature points. They are employed for the 3-D recovery along with the feature points specified at the basketball. The mobile frame is conveyed manually following and capturing the subject. The feature points observable from all the three video cameras are chosen at each sample time in order to make the initial measurement matrix given by Eq.(14).

The above manual operation to the frame is often not very smooth and it may cause some noise in the recovered motion. Although a certain kind of post-processing is indispensable for removing the nose, the recovery result without the noise reduction is reported in this particular paper, since the main objective of the paper is to give a basic idea of the new type of motion recovery.

Employed video images at some sample times are shown in **Figure 2**. In all, 170 shots were sampled on the video images with the interval of 1/30 second and 18 feature points were chosen and tracked half-automatically by calculating the correlation function on the images. The recovered result is depicted in **Figure 3**, where time proceeds as indicated by arrows.

5. Discussion

The proposed technique contains a novel idea of taking images of a moving object by following it with a mobile camera frame and recovering its 3-D shape. This recovery is realized by producing a full matrix from an initial measurement matrix W^{in} whose components are sparse. The idea of shifting the sub-matrices in matrix W^{in} of Eq.(14) means that the mobile camera frame is regarded as settled fixed at a certain location and the frontal scene relatively moves against the frame. This enables the present technique to observe a moving object by following it and recovers its 3-D motion as if the object makes a motion all the time in front of virtually fixed cameras. The effect can be seen in **Figure 3**. This provides us with recovery results different from those obtained from fixed video cameras mentioned in Section 2, in front of which an object may pass away.

The performed experiment shows acceptable results, but some issues remain to be solved. Since the cameras mounted on the mobile frame are fixed their orientations as well as positions, the distance between the frame and the object should be kept constant so that the cameras always observe the object in the center of their visual planes. Manual movement of the frame is not very easy to achieve this. Movement of the frame is also likely to be disturbed by a rough floor. This may result in unstable location of the recovered object. Mechanical realization of the image taking needs to be considered for the experiment.

The recovery error mainly comes from the assumption of orthographic projection of the employed factorization technique and tracking errors of the projected feature points on the obtained video images. The former may be improved by introducing an approximating model of perspective projection [5], for example, in lens imaging system, whereas the latter needs further investigation. Average recovery error of approximately 4% is empirically obtained in case the frame does not move.

Because of the mobility of the proposed system, the present technique may have applications in various ways. Wide range motion recovery may be one of the main areas of application, such as field as well as track athletics.

6. Conclusions

A technique was proposed for recovering 3-D human motions employing mobile cameras. Advantages of the present technique over others are that it is applicable to wide range motions and camera calibration is not necessary. Instead, a certain amount of recovery errors cannot be avoided because of the approximation of the imaging by orthographic projection and the disturbance on the captured images caused by movement of the image-taking frame. Evaluation of recovery errors is now under study.

7. Acknowledgements

The authors are grateful to Mr. T. Yamashita, Technical Staff of Yuge National College of Maritime Technology, for his help in producing the mobile frame. They are also thankful to the students who joined in the experiment.

8. References

[1] Jain, A. K.: *Fundamentals of Image Processing*.

[2] Tan, J. K., Kawabata, S., Ishikawa, S.: "An efficient technique for motion recovery based on multiple views", *Proc. IAPR Workshop on Machine Vision Applications*, 270-273(Nov.,1998).

[3] Tan, J. K., Ishikawa, S.: "Human motion recovery by the factorization based on a spatio- temporal measurement matrix", *Computer Vision and Image Understanding*, **82**, 2, 101-109 (May, 2001).

[4] Tomasi, C., Kanade, T.: "Shape and motion fom image streams under orthography: A factorization method", *Int. J. Comput. Vision*, **9**, 2, 137/154 (1992).

[5] Poelman, C. J., Kanade, T.: "A paraperspective factorization method for shape and motion recovery", *IEEE Trans. Patttern Anal. Mach. Intell*., **19**, 3, 206- 218(March, 1997).

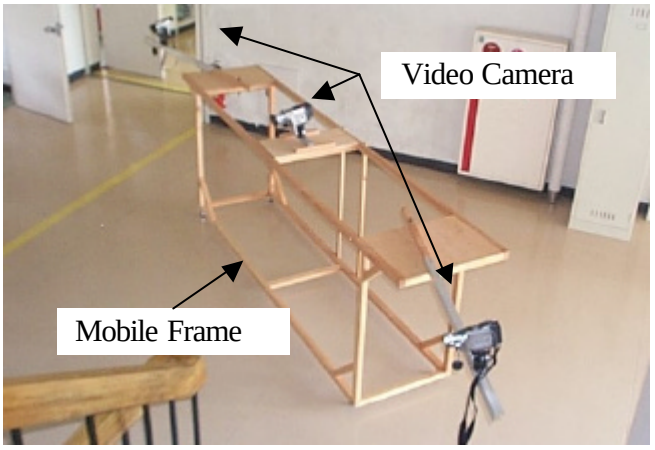

Figure 1. Mobile frame with fixed cameras.

Figure 2. Original images at some sample times.

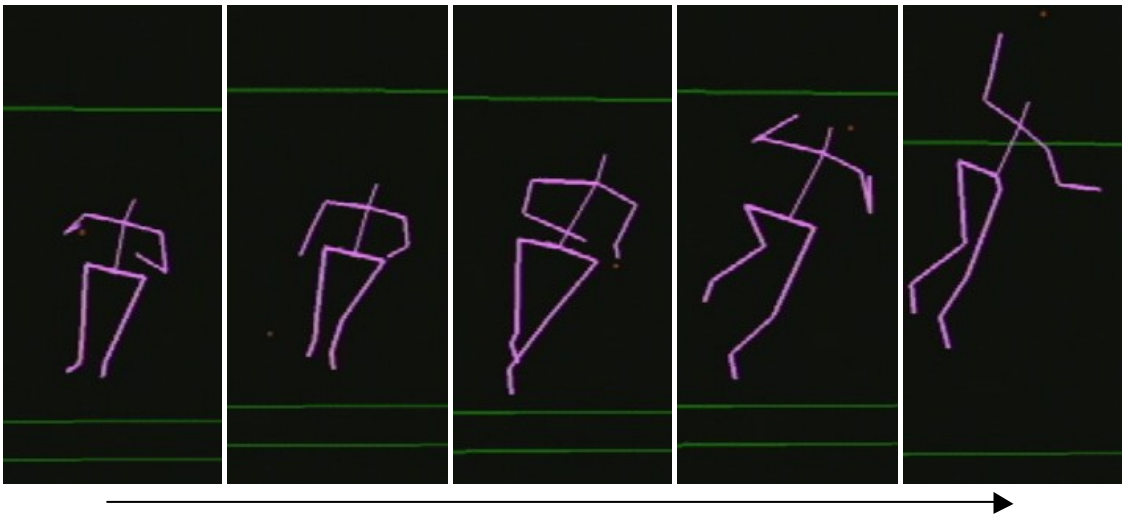

Figure 3. Some shots of the recovered motion.