The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

Three-dimensional Human Motion Modeling by Back Projection Based on Image-based Camera Calibration

S. Masaoka, J. K. Tan, H. Kim and S. Ishikawa

Kyushu Institute of Technology,
1-1 Sensui-cho, Tobata-ku, Kitakyushu-shi, Fukuoka, 804-8550, Japans
(Tel : 81-93-884-3191; Fax : 81-93-884-3191)
(masaoka@ss10.cntl.kyutech.ac.jp)

Abstract: This paper proposes a back projection technique for 3-D human motion modeling that performs camera calibration using obtained multiple video sequences. This technique calculates an affine camera matrix from the factorization method and performs back projection under the affine camera model. The proposed technique needs neither 3-D camera calibration tool nor markers for shape recovery, and can recover a human motion from silhouette images. In this paper, we also propose a shadow detector and eliminator using color information and normalized cross correlation for robust extraction and elimination of shadows. Experimental results show effectiveness of the proposed technique.

Keywords: motion capture, back projection, factorization, affine camera model

I. INTRODUCTION

Human motion recovery from video sequences is an important as well as interesting subject of study in computer vision with various applications in the fields of video media and sports. One of the advantages of the optical technique employing cameras should be its limitless nature with respect to human motion range and its environment. However many of conventional techniques need camera calibration using a 3-D tool in order to acquire three-dimensional coordinates in a real space. Thus these techniques are not suitable for human motion modeling in outdoor scenes. A technique of 3-D modeling without camera calibration [1] and a mobile stereo technique employing image-based self-calibration [2] have already been proposed as calibrationless techniques before image capture. They are suitable in various outdoor scenes. However, its post-processing is complicated because of the modeling based on markers. As a markerless technique, the back projection technique is proposed [3]. The post-processing of this technique is simpler than other techniques because it performs 3-D modeling from only a silhouette image. But, this technique need to perform camera calibration employing a 3-D tool before video capture.

This paper proposes a back projection technique that performs camera calibration using obtained video sequences without a 3-D calibration tool and markers. It employs the factorization technique [4] for deriving camera orientations and performs back projection of the multiple silhouette images of the object interested by making use of the camera orientations.

We use background subtraction to acquire silhouette images. However, this technique cannot be simply employed in outdoor scenes because of shadows. We then propose a shadow detector and eliminator using color information and normalized cross correlation for robust extraction and elimination of shadows. The proposed technique is improved based on a normalized RGB color space compared to the existing normalized cross correlation technique [5]. By employing the color information, the shadow detection gains much robustness. Furthermore, the threshold is set adaptive with respect to various scenes in each frame by approximating the normalized cross correlation by a Gaussian distribution.

II. PROPOSED TECHNIQUE

1. Affine camera model and the factorization

As a projection model, we consider an affine camera model such as week-perspective projection. When a three-dimensional point $\boldsymbol{X}=(X,Y,Z)$ in the world coordinate system is projected on an image point $\boldsymbol{x}=(x,y)$ in an image coordinate system, it is expressed in a homogeneous coordinate system as follows;

$$\tilde{\mathbf{x}} = P_a \tilde{\mathbf{X}} \tag{1a}$$

$$P_a = \begin{pmatrix} \hat{P}_a & \mathbf{t} \\ \mathbf{0} & 1 \end{pmatrix} \tag{1b}$$

where P_a is a 3×4 affine camera matrix. It consists of a 2×3 matrix $\hat{P}_a$ and a 2-component column vector $\boldsymbol{t}$.

Equation (1a) is expressed in the Euclid coordinate system as follows;

$$\mathbf{x} = \hat{P}_a \mathbf{X} + \mathbf{t} \tag{2}$$

In Eq. (2), when $\boldsymbol{X}$=0, $\boldsymbol{t}$ is a projected point of the origin of the world coordinate system to the camera image, because $\boldsymbol{x}$=$\boldsymbol{t}$. We can therefore assume that the origin of the camera image is $\boldsymbol{t}$. Let $\boldsymbol{x}$ be deviation from $\boldsymbol{t}$, and the following equation holds.

$$\mathbf{x} = \hat{P}_a \mathbf{X} \tag{3}$$

Suppose that the points $\boldsymbol{X}_p$ (p=1,2,…,P) in the space is projected to the points $\boldsymbol{x}_p$ on the image plane of a fixed camera f (f=1,2,…,F). Then, by Eq. (3),

$$\mathbf{x}_{fp} = \hat{P}_{a,f} \mathbf{X}_p \tag{4}$$

In the case of F cameras, we have

$$\begin{pmatrix} \mathbf{x}_{11} & \mathbf{x}_{12} & \cdots & \mathbf{x}_{1P} \\ \mathbf{x}_{21} & \mathbf{x}_{22} & \cdots & \mathbf{x}_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_{F1} & \mathbf{x}_{F2} & \cdots & \mathbf{x}_{FP} \end{pmatrix} = \begin{pmatrix} \hat{P}_{a,1} \\ \hat{P}_{a,2} \\ \vdots \\ \hat{P}_{a,F} \end{pmatrix} \left(\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_P \right) \tag{5}$$

Let the left-hand side matrix be denoted by W. W is the $2F \times P$ matrix. By performing the factorization to W, the following decomposition is obtained;

$$W = MS \tag{6}$$

where M is a camera orientation matrix, and S is a 3-D shape matrix of points $\boldsymbol{X}_i$. The matrix M is given by

$$M \equiv \begin{pmatrix} \hat{P}_{a,1} \\ \hat{P}_{a,2} \\ \vdots \\ \hat{P}_{a,F} \end{pmatrix} \tag{7}$$

On the plane which is perpendicular to an optical axis via a lens center of camera f, let $\boldsymbol{i}_f$ be a horizontal unit vector to the lens center, and $\boldsymbol{j}_f$ be a vertical unit vector. Then $\hat{P}_{a,f}$ is given by;

$$\hat{P}_{a,f} = \begin{pmatrix} \mathbf{i}_f^{\mathrm{T}} \\ \mathbf{j}_f^{\mathrm{T}} \end{pmatrix} \tag{8}$$

So, if the points $\boldsymbol{x}_{fp}$ $(f = 1,2,...,F;\ p = 1,2,...,P)$ on the image are obtained, $\hat{P}_{a,f}$ is acquired by Eqs. (6) and (8).

2. The back projection technique based on the affine camera model

The back projection technique is a modeling technique that projects silhouette images on camera image planes back to the 3-D space and obtain a 3-D object model by taking the common region of the back projected images. The conventional back projection technique needs to perform camera calibration using a 3-D calibration tool because of a camera projection model. Here we propose a back projection technique based on the affine camera model. It performs the factorization and the back projection employing the calculated affine matrices. By this approach, the proposed technique needs neither 3-D camera calibration tool nor markers for shape recovery, and can recover a human motion from silhouette images. This technique is described in the following.

A pixel $\boldsymbol{x}_f \in S_f$ in the silhouette S_f on the fixed camera image f is projected back to the 3-D space by Eq. (1a). Then the set of the points is given as follows;

$$BP(\mathbf{x}_f) = \left\{ \mathbf{X} \mid \lambda \tilde{\mathbf{x}}_f = P_{a,f} \tilde{\mathbf{X}} \right\} \tag{9}$$

where P_f is the affine camera matrix of camera f. The set of the points $BP(S_f)$ that contains all the projected pixels in the S_f is defined by

$$BP(S_f) = \bigcup_{\mathbf{x}_f \in S_f} BP(\mathbf{x}_f) \tag{10}$$

A 3-D model V employing the back projection is finally obtained by

$$V = \bigcap_{f=1,2,...,F} BP(S_f) \tag{11}$$

3. Shadow detection and elimination

In this paper, we perform a silhouette extraction by using a background subtraction technique. However, it is often difficult to extract a silhouette image because of the shadow of the object. Therefore, we propose a shadow detection and elimination technique using color information and normalized cross correlation. This is more robust than employing existing techniques.

In the first place, the proposed technique extracts a foreground area including the shadow as follows;

$$\begin{cases} 1 & if\ L_{th} < |T - B| < H_{th} \\ 0 & otherwise \end{cases} \tag{12}$$

where T and B are the sum of each channel of RGB in an input image and in a background image, respectively. L_{th} and H_{th} are the thresholds for extracting the area including the shadow. The shadow detection is applied to all the pixels that satisfy Eq. (12). In the second, we calculate a feature quantity that mixes the color

information with the texture information. In the color information, we use the normalized RGB color space. The conversion from RGB color space to normalized RGB color space is given by

$$\begin{cases} r = \dfrac{R}{R+G+B} \\ g = \dfrac{G}{R+G+B} \\ b = \dfrac{B}{R+G+B} \left(= 1 - r - g\right) \end{cases} \tag{13}$$

In the texture information, we use the normalized cross correlation which is defined in the normalized RGB color space as follows;

$$NCC(i,j) = \frac{\sum_{k=r,g}\sum_{n=-N}^{N}\sum_{m=-N}^{N} B_k(i+n,j+m)T_k(i+n,j+m)}{\sqrt{\sum_{k=r,g}\sum_{n=-N}^{N}\sum_{m=-N}^{N} B_k(i+n,j+m)}\sqrt{\sum_{k=r,g}\sum_{n=-N}^{N}\sum_{m=-N}^{N} T_k(i+n,j+m)}} \tag{14}$$

where $T_k(i,j)$ is the value of r and g in a pixel $T(i,j)$ of the input image, and $B_k(i,j)$ is that in a pixel $B(i,j)$ of the background image. N is the size of a window. Furthermore, the threshold is set adaptive with respect to various scenes in each frame by approximating it by a Gaussian distribution as follows;

$$f(x,\sigma^2/n) = \frac{1}{\sqrt{2\pi(\sigma^2/n)}} \exp\left(-\frac{(x-1)^2}{2(\sigma^2/n)}\right) \tag{15}$$

where σ^2 is a variance of the correlation value, and n is the number of samples. Finally the shadow area is defined as the pixel that satisfies the following relation;

$$NCC(i,j) > 1.0 - c\frac{\sigma}{\sqrt{n}} \tag{16}$$

where c is a parameter. An example of the proposed shadow detection and elimination is shown in Fig.1.

III. EXPERIMENTAL RESULTS

We performed an experiment of a human motion recovery by employing 4 cameras in an outdoor environment. We fixed the distance from the cameras to a human about 5*m*, and the angle between the set cameras about 45 degrees. We extracted 39 points manually for the factorization method. A sample of the video sequences is shown in Fig.2. The result of the shadow detection and elimination is shown in Fig.3. Examples of silhouette images are shown in Fig.4.

The result of the 3-D recovery is depicted in Fig.5. In order to evaluate the proposed technique, we examined the number of true positive voxels contained in the recovery results employing ground truth data. To compare the proposed technique with a conventional technique, we calibrated the video taking system with the DLT technique using a 3-D calibration tool. The precision of the 3-D recovery is shown in Table.1.

IV. DISCUSSION

The precision of the motion recovery was 83.0% in the proposed technique, whereas it was 86.4% in the back projection technique employing the DLT method. The result is not much difference in spite of the approximation by an affine camera model. It should be noted that the proposed technique is more adaptable to various usages including outdoor motion capture than the existing back projection technique. It is indeed convenient for the modeling of transient, or unrepeated motions, since one only has to take their images on the spot and can calibrate the cameras afterword in the lab using the obtained videos. Thus the proposed technique may contribute to expanding motion capture technology to various new fields.

An increasing error in Camera2 has been attributed to the effect of a moving distance of a human in the direction of the optical axis. Because the human in Camera2 is moving largely in the direction of the depth compared with other cameras, it may result in larger computational errors than other cameras. On the other hand, in Camera4, the error is the smallest, since there was the least moving distance by the human.

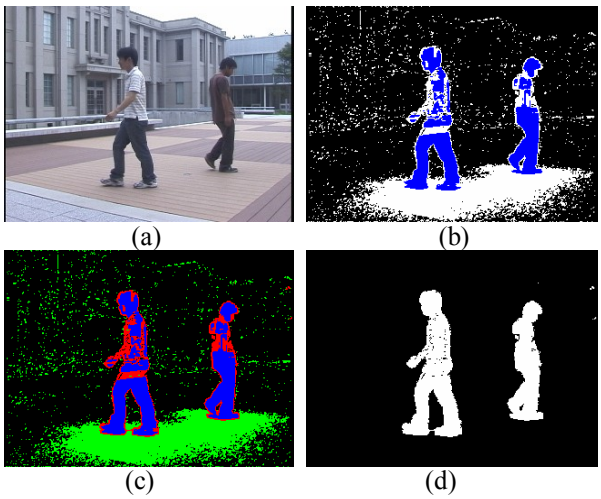


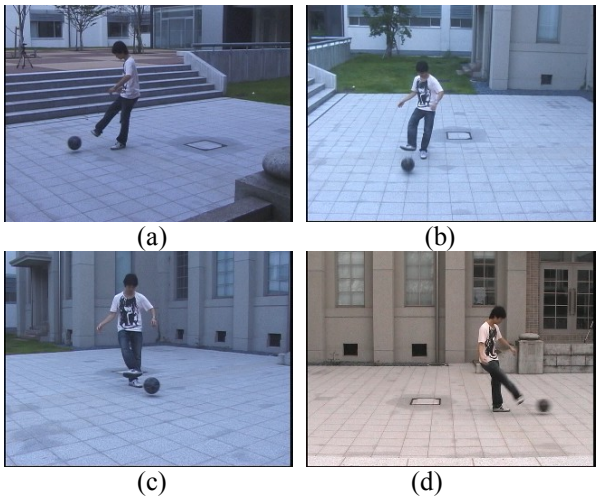
Fig.1 Experimental result of shadow detection and removal. (a) Input image. (b) Foreground image. (c) Result of the shadow detection. (d) Result of the shadow removal.

V. CONCLUSIONS

In this paper, we proposed a back projection technique with simpler camera calibration. The proposed technique does not need a 3-D calibration tool. Instead it performs camera calibration by employing video sequences. The technique calculates an affine camera matrix from the factorization method, and performs back projection under the affine camera model. We also proposed a shadow detection and elimination technique using color information of the scene. This is more effective than employing light intensity to a textured scene. Future works include the recovery of a human motion in various outdoor scenes.

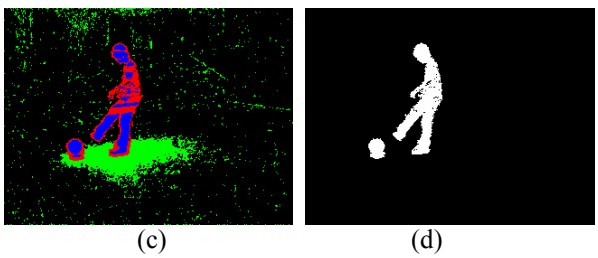
Table.1 True Positive Rate

	The DLT technique	The Proposed technique
Camera1	86.3 %	84.7 %
Camera2	87.8 %	80.1 %
Camara3	86.4 %	82.1 %
Camera4	85.0 %	85.0 %
Average	86.4 %	83.0 %



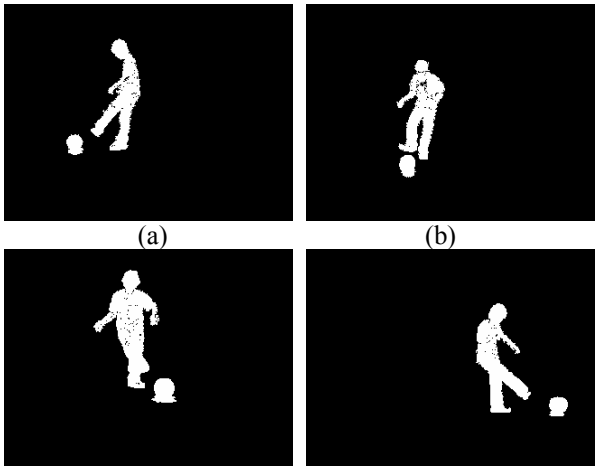
(a) (b) (c) (d)

Fig.2 Video sequence. (a)Camera1, (b)Camera2, (c)Cam era3, (d)Cmaera4.



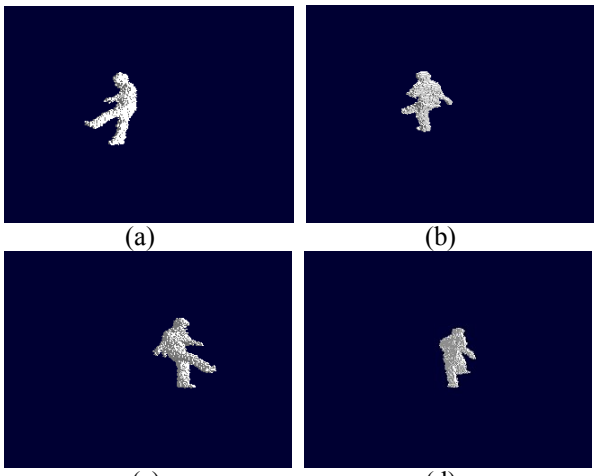
(c) (d)

Fig.3 Experimental result of shadow detection and elimination. (a)Shadow detection,(b)shadow elimination.



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

Fig.4 Silhouette image. (a) Camera1, (b) Camera2, (c) Camera3, (d) Cmaera4.



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

Fig.5 Experimental result of human motion recovery. (a) View1, (b) View2, (c)View3, (d)View4.

REFERENCES

[1] Tan J. K., Ishikawa S.: "Deformable shape recovery by factorization based on a spatiotemporal measurement matrix", *Computer Vision and Image Understanding*, Vol. 82, No. 2, pp.101-109, 2001.

[2] Yamaguchi I., Tan J. K., Ishikawa S.: "A mobile motion capture technique excelling in 3-D modeling of temporary events", *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, pp.1613-1617, 2006.

[3] Uchinoumi M., Tan J. K., Ishikawa S.: "A simple structured real-time motion capture system employing silhouette images", *Proc. IEEE Int. Conf. on Systems, Man and Cybernetics*, pp. 3094-3098 (Oct., 2004).

[4] Tomasi C., Kanade T.: "Shape and motion from image streams: a factorization method", Technical Report CMU-CS-91-172, CMU, 1991.

[5] Jacques Jr., J. C. S., Jung C. R., Musse S. R.: "Background subtraction and shadow detection in grayscale video sequences", *Proceedings of SIGGRAPH*, pp.189-196, 2005.