

**Motion Reconstruction from a Single Video Stream
Using a Motion Library**

Min Je Park Min Gyu Choi Sung Yong Shin

CS/TR-2002-172

January 14, 2002

K A I S T
Department of Computer Science

Motion Reconstruction from a Single Video Stream Using a Motion Library

Min Je Park Min Gyu Choi Sung Yong Shin

Abstract

Videos taken from a single camera are a most common source of human motions. In this paper, we present a novel method to reconstruct the motion of a human-like figure from a single video stream. We exploit a motion library to resolve the depth ambiguity in recovering the 3D configurations from 2D features. Our reconstruction method takes three major steps: time-warping to align the reference motion with that in the video, reconstructing the joint orientations, and estimating the root trajectory. Experimental results show that our approach can reconstruct highly dynamic motions such as shooting of soccer players, which would be hard to do, otherwise.

1 Introduction

Recently, motion capture techniques have been applied successfully to animating human-like figures. With those techniques, realistic motions can be produced rapidly, which would be hard to achieve, otherwise. Live-captured motion clips are typically short and related to particular characters and environments. Thus, there have been a great deal of efforts to develop specialized tools to reuse them. With the aid of such tools, animators can adapt the captured motions to other characters and environments. However, together with special hardware devices, the motion capture process requires a puppeteer(or an actor) who performs a sequence of motions carefully under a controlled environment in accordance with a scenario. Thus, we cannot directly apply motion capture techniques to reproducing motions in actual situations such as sports events and dance performances.

Single-camera videos are a most common source of human motions. Many researchers have tried to capture a variety of human motions from videos for various purposes [1, 12, 15, 18, 19, 25, 26]. However, reconstructing motion from a single-camera video is still demanding even with the state of the art techniques.

Most of the previous techniques are inadequate to reconstruct highly dynamic motions for a variety of circumstances, and the quality of the reconstructed motion is not adequate to synthesize realistic animation.

To reconstruct human motion from a video, we should recover the 3D configuration of a figure from its 2D features such as the positions of the joints and the end-effectors. The main difficulty in motion reconstruction from a single-camera video stems from the depth ambiguity caused by the 3D projection onto 2D images. Moreover, the focal length of the camera is unavailable, and the camera may move along with the actors. Thus, the actual trajectory of an actor is hard to capture only with the information available in the video.

In this paper, we present a novel approach to reconstruct a motion of a human-like figure from a single video stream using a motion library. We assume that the scaled lengths of body segments of an actor in the video and its 2D feature positions are available. Those may be estimated automatically [18, 25] or obtained interactively. Guided by those data, we reconstruct the motion by deforming a captured motion clip in the library. Our approach requires a motion library containing a motion clip similar to the target motion in the video stream. In theory, such a motion library is not always available since human motions are too diverse to be accommodated in the library. We take a practical approach assuming that we know what to reconstruct in advance. For example, suppose that we are to reconstruct shooting motions in a soccer game recorded in a video. After building a library of appropriate live-captured motion clips for shooting, we use them to reconstruct actual players' shooting motions in the video.

2 Related Works

2.1 Motion Reconstruction

There have been research results [1, 2] to reconstruct human motions from videos automatically with image processing techniques. In those researches, statistical models are employed to estimate 3D configurations of human bodies. Azarbayejani et al. [1] proposed a method to track the motion of a human body with a monocular/stereo video stream in real time. They segmented the human body with several blobs and tracked the 3D position of each blob with a statistical dynamic model. Bregler and Malik [2] also estimated a motion from a video sequence taken from one or more cameras. Based on *twist* and *exponential maps* to represent the kinematic relationship of an articulated model, they reconstructed its 3D configuration. They tried to track the features in the video automatically.

On the other hand, some researchers have concentrated on developing efficient

ways to extract 3D information from known 2D features. Zheng and Suezaki [26] introduced a model-based approach to acquire motions of an articulated model from a single video stream. They selected several keyframes to recover the configurations of the model and interpolated them to obtain a motion. Rehg et al. [6] proposed an off-line algorithm to estimate the maximum a posteriori trajectory from the 2D measurements subject to a number of constraints such as a kinematic model and joint angle limits.

Recently, an interesting attempt has been tried to address the reconstruction problem in some different point of view. Howe et al. [11] proposed a scheme to reconstruct the 3D motion of an articulated model from a single-camera video. Relying on prior knowledge about human motion learned from training data, they resolved the inherent depth ambiguity of 2D videos. They treated 3D motion tracking as an inference problem.

2.2 Motion Reuse

Due to the success of motion capture technology, there is a vast amount of literature devoted to motion capture and reuse. Bruderlin and Williams [3] introduced the idea of displacement mapping to alter a motion while preserving its details. Witkin and Kass [23] proposed a spacetime constraint technique to produce the optimal motion which satisfies a set of user-specified features. Cohen [4] developed a spacetime control system which allows a user to interactively guide a numerical optimization to find an acceptable solution in a feasible time. Rose et al. [20] adopted this approach to generate a smooth transition between motion clips. Gleicher [9] simplified the spacetime problem by removing the physics-related aspects from the objective function and constraints for motion editing. He also applied this technique for motion retargeting [10]. For interactive performance, Lee and Shin [13] combines a hierarchical curve fitting technique with a new inverse kinematics solver for adaptively refining a motion to meet the spacetime constraints. Popovic and Witkin [16] introduced a novel algorithm for editing captured motions that takes dynamics into consideration. They simplified a complex dynamic system without losing the fundamental dynamic properties of motion. Tak et al. [22] proposed a motion balance filter that postprocesses the edited motion to keep the dynamic balance using the notion of a zero moment point (ZMP).

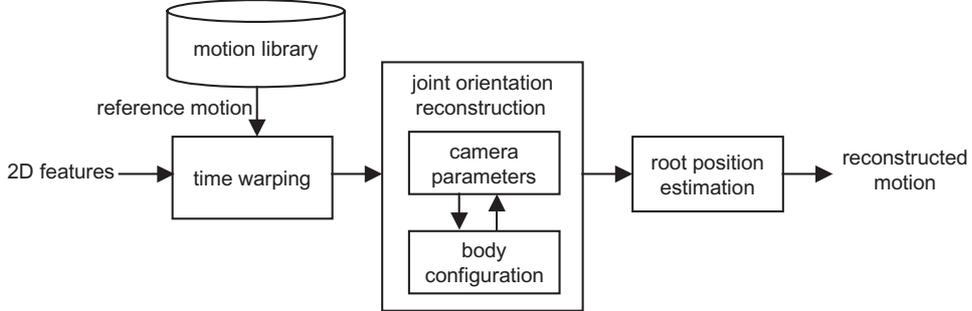


Figure 1: Block diagram of our reconstruction method

3 Overview

In this paper, we introduce a novel approach to reconstruct the motion of a human-like figure from a single video stream by exploiting a live-captured motion. We exploit a prior knowledge about the target motion to resolve the depth ambiguity in recovering the 3D configurations from 2D features. Using the motion in the library as a reference, we first warp the reference motion to establish the time correspondence with the motion in the video, and then reconstruct a motion by deforming the timewarped reference motion guided by the features in the video. Since the input video is taken from an uncalibrated camera with unknown trajectory, we can not directly acquire a root trajectory of the motion from the video. Thus, we first recover the joint orientations to achieve a sequence of postures such that their projected joint positions are coincident with the features in the video. Then, we derive a natural-looking root trajectory, exploiting the set of 2D features derived from kinematic constraints given at frames and the dynamic property of the reference motion. Figure 1 shows the block diagram that describes our reconstruction method.

To make the time correspondence between the input video and the reference motion, we start with marking their keytimes, that is, the moments of interaction between the actor and his/her surrounding environments in the video. Again, this can be done either automatically with existing image tracking methods or interactively. We timewarp the reference motion to synchronize its keytimes with those of the video. This process reparameterizes the reference motion clip. We use the timewarped motion as an initial guess for the target motion to reconstruct.

To obtain a relative posture of the 3D configuration of the articulated figure

with respect to the root segment, we establish the constraints that force their projected joint positions are coincident with the corresponding features in the video. Since there are, in general, multiple configurations that satisfy these constraints, we use an objective function to select a configuration that has minimum deviation from that of the reference motion. We propose an efficient method to acquire the articulated body configuration as well as the unknown camera parameters, simultaneously. To construct a smooth motion, we compute the posture difference between the reference motion and the reconstructed motion at each frame, and then make a posture displacement map that approximate the posture differences using the multilevel B-spline approximation technique [14]. With this map, we deform the reference motion.

Finally, we estimate a root trajectory to complete the reconstruction process. We have two different cases: In the first case, we deal with the motion that exhibits some interactions between the actor and his/her surrounding environment. In this case, we modify the root trajectory of the reference motion to acquire a feasible root trajectory of the target motion that preserves the interactions. Using the multilevel B-spline approximation technique [14], we interpolate the displacement of the root segment at each frame of the interaction while smoothly propagating approximation errors to neighboring frames. In the second case, we treat a motion that does not show such interactions. In this case, we exploit the dynamic property of the reference motion that should be preserved. In particular, assuming that the articulated figure consists of a set of rigid segments, we obtain the root trajectory of the motion from the center of gravity(COG) trajectory of the reference motion.

The remainder of this paper is organized as follows. Section 4 describes our keytime-based timewarping, and section 5 shows how we can obtain a proper camera parameters and a sequence of relative postures from the input video. In section 6, we complete the motion reconstruction process combining the relative postures and the root trajectory. Section 7 demonstrates experimental results for shooting motions of soccer players. Finally, we conclude this paper in section 8.

4 Timewarping

A motion is a time-varying function that gives the configuration of an articulated figure. For an articulated figure with n joints, we denote a motion by $\mathbf{m}(t) = (\mathbf{p}_1(t), \mathbf{q}_1(t), \dots, \mathbf{q}_n(t))^T$, where $\mathbf{p}_1(t) \in \mathbb{R}^3$ and $\mathbf{q}_1(t) \in \mathbb{S}^3$ describe the translational and rotational motions of the root segment, respectively, and $\mathbf{q}_i(t) \in \mathbb{S}^3$ give the rotational motion of the i -th joint for $2 \leq i \leq n$. We denote the features in a video stream by $\bar{\mathbf{m}}(t) = (\bar{\mathbf{p}}_1(t), \dots, \bar{\mathbf{p}}_n(t))^T$, where $\bar{\mathbf{p}}(t) \in \mathbb{R}^2$ gives the pro-

jected position of the i -th joint for $1 \leq i \leq n$. The features given by $\bar{\mathbf{m}}(t)$ represents the projection of the target motion $\mathbf{m}(t)$ onto the image at time t .

Given the reference motion $\mathbf{m}(t)$ and the 2D features $\bar{\mathbf{m}}(t)$ in the video, we establish a time correspondence between $\mathbf{m}(t)$ and $\bar{\mathbf{m}}(t)$. It is well-known that the dynamic timewarping technique gives an optimal sample correspondences between two functions [3, 5]. However, the camera parameters are not available at each frame. Thus, the dynamic timewarping technique cannot be applied directly to building a time correspondence between $\mathbf{m}(t)$ and $\bar{\mathbf{m}}(t)$, which have different dimensions.

To address this problem, we start with a set of keytimes in the video, that is, the moments of interaction between the actor and his/her surrounding environment. For example, the keytimes of human walking motion can be described as the instances of heel-strikes and toe-offs. For a kick motion in a soccer game, the most important keytime is the impact moment between a foot of a player and a ball. Such a keytime can be easily marked in the video as well as in the reference motion interactively. They may also be tracked automatically [18, 25]. Assuming that they are available, our task is to timewarp the reference motion to have the same keytimes as specified in the video.

For the i -th joint, let $K_i = \{t_{i,1}, \dots, t_{i,c}\}$ be a set of keytimes for the reference motion and $\bar{K}_i = \{\bar{t}_{i,1}, \dots, \bar{t}_{i,c}\}$ the counterpart for the video stream. To timewarp the reference motion to make K_i coincide with \bar{K}_i for all i , we reparameterize the reference motion $\mathbf{m}(t)$ by a set of linear mappings defined as follows:

$$t'(t) = \bar{t}_{i,k} + \left(\frac{\bar{t}_{i,k+1} - \bar{t}_{i,k}}{t_{i,k+1} - t_{i,k}} \right) (t - t_{i,k}) , \quad (1)$$

where $t_{i,k} \leq t \leq t_{i,k+1}$. Here, t and t' describe the original time of the reference motion and its reparameterized time, respectively. Figure 2 illustrates this procedure for a single joint. Since a human-like articulated figure has multiple joints, we apply this procedure for each joint repeatedly as shown in Figure 3.

5 Joint Orientation Reconstruction

5.1 Kinematic Constraints

At each frame, the projected joint positions of the articulated figure in the reference motion need to be coincident with their corresponding features in the video. The input video is taken from an uncalibrated camera with unknown trajectory, and the reference objects are not always available in the video. Therefore, to acquire

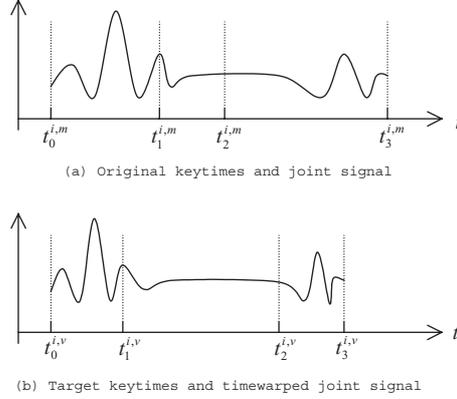


Figure 2: Timewarping of a reference motion clip using a set of keytimes in the input video sequence and the reference motion. The curve represents the change of one component of a unit quaternion with respect to time. (a) Original joint signal with a set of keytimes, $K_i = \{t_{i,0}, t_{i,1}, t_{i,2}, t_{i,3}\}$. (b) Timewarped motion by a set of keytimes, $\bar{K}_i = \{\bar{t}_{i,0}, \bar{t}_{i,1}, \bar{t}_{i,2}, \bar{t}_{i,3}\}$.

the joint positions in the global frame, we describe them relatively from the root segment. In other words, the configuration of the articulated figure is: $\mathbf{x}(t) = \mathbf{m}(t)|_{\mathbf{p}_1(t)=\mathbf{0}_3}$, that is, $\mathbf{x}(t) = (\mathbf{0}_3, \mathbf{q}(t)_1, \dots, \mathbf{q}(t)_n)^T$, where $\mathbf{0}_3$ represents the origin of the local frame of the root segment.

To describe the relationship between a 3D configuration of a figure and its projection onto an image plane, we need a camera model. In general, a camera model with full degrees of freedom is parameterized by $\mathbf{c} = (t_x, t_y, t_z, r_x, r_z, r_y, \alpha, f)$, where (t_x, t_y, t_z) and (r_x, r_y, r_z) describes the position and orientation of the camera, respectively, α describes its aspect ratio, and f is its focal length. We set the camera posture such that the projected position of the root segment is always placed at the center of the image plane. Then, we can describe the configuration of the camera by the distance to the root of the articulated figure and its orientation. Our reduced camera model is parameterized by $\mathbf{c} = (r_x, r_z, r_y, \gamma)$, where γ is the ratio of this distance to f .

With the camera parameters \mathbf{c} , the kinematic constraint on the articulated figure is defined as follows:

$$\|\bar{\mathbf{p}}_i(t) - \mathbf{P}_c \mathbf{f}_i(\mathbf{x}(t))\| = 0 \quad (2)$$

where $\mathbf{f}_i(\cdot)$ is the forward kinematic function for the i -th joint, and \mathbf{P}_c describes

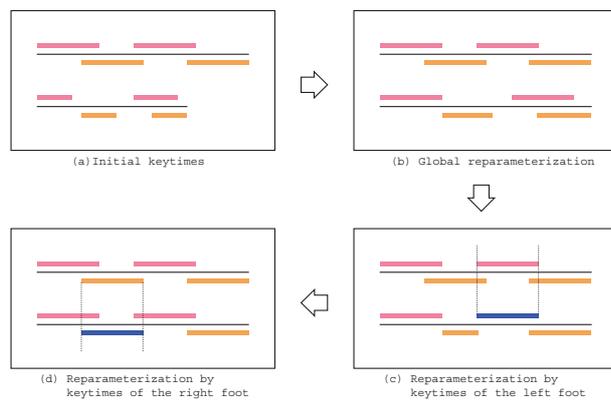


Figure 3: Timewarping of a reference motion clip using multiple sets of keytimes in the input video sequence and the reference motion. There are two sets of keytimes for left and right feet. We mark a band between two distinct keytimes for visual aids. (a) Initial sets of keytimes for the left(dark color) and the right(light color) foot in the input video(above) and the reference motion(below) with a time line. (b) Reparameterization of the reference motion to have the same duration with the input video. (c) Reparameterization of the reference motion at a set of keytimes of the left foot. (d) Reparameterization of the reference motion at a set of keytimes of the right foot.

the projection matrix.

5.2 Objective Function

Due to the excessive degrees of freedom for an articulated figure, there are typically many possible configurations that satisfy the kinematic constraints given by Equation 2. DiFranco et al [6] pointed out that this depth ambiguity can be removed partially using some additional constraints such as joint angle limits. Even with such additional constraints, the problem of motion reconstruction is still under-constrained. To achieve the best configuration, we exploit the reference motion for the articulated figure. Assuming that the reference motion is similar to the target motion in the video, the minimum change of the joint orientations from the reference motion ensures the naturalness of the reconstructed motion. Therefore, we find a configuration $\mathbf{x}(t)$ by minimizing the following objective function:

$$g(\mathbf{x}(t)) = \text{dist}(\mathbf{x}^r(t), \mathbf{x}(t)). \quad (3)$$

Here, $\mathbf{x}^r(t)$ describes the configuration of an articulated figure in the reference motion at time t , and $\text{dist}(\cdot)$ describes the distance between two orientations:

$$\text{dist}(\mathbf{x}^r(t), \mathbf{x}(t)) = \sum_{i=1}^n \|\ln((\mathbf{q}_i(t))^{-1} \mathbf{q}_i^r(t))\|^2, \quad (4)$$

where $\ln(\cdot)$ is the logarithmic map of unit quaternions [21].

5.3 Solution Method

Our joint orientation reconstruction problem is reduced to finding the configuration that minimizes the objective function $g(\cdot)$ while satisfying the constraints given by Equation 2. A typical approach to the constrained optimization is to transform the constrained problem into an unconstrained version with extra penalty functions. The objective function for the unconstrained version is

$$\hat{g}(\mathbf{x}(t)) = \sum_{i=1}^n (\|\bar{\mathbf{p}}_i(t) - \mathbf{P}_c \mathbf{f}_i(\mathbf{x}(t))\|^2) + \omega(\text{dist}(\mathbf{x}^r(t), \mathbf{x}(t))), \quad (5)$$

where ω is a weighting factor combing the two different measures. The first term of Equation 5 represents the difference between the projected joint positions of the articulated figure and the marked joint positions in the input video, and the second term measures the deviation of the posture of the figure from that of the reference

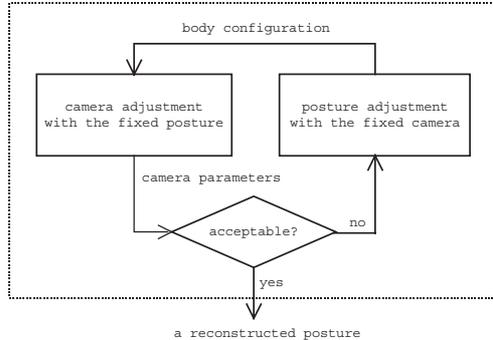


Figure 4: Iterative scheme to reconstruct the configuration of the camera and the articulated figure. This procedure starts with an initial configuration of the camera computed by 3 coplannar points specified by a user. The initial guess of the articulated figure is automatically given by the reference motion.

motion at each time instance. We adopt the conjugate gradient method to minimize this objective function [17].

The major difficulty in solving the Equation 5 stems from the excessive degrees of freedom of an articulated figure. A reasonable human model for computer animation has about 40 degrees of freedom. We specify much fewer constraints, compared to the degrees of freedom to determine. Furthermore, the unknown camera parameters complicate the problem. Typical numerical solvers such as conjugate gradient methods show the best convergence property when the objective function has a quadratic form [17]. The Equation 5 has a combination of parameters for both camera and articulated body configuration. We handle the two different sets of parameters separately during optimization to acquire better convergence. We alternately optimize the camera posture and the body configuration while fixing the other. Figure 4 shows this iterative process of acquiring both camera and body configuration, simultaneously.

It is well-known that a good initial guess for optimization problem is important to obtain a good solution with better convergence [8, 7, 17]. In our reconstruction method, we use the timewarped reference motion as the initial estimate of the target posture. Since our camera model has four degrees of freedom, we select three coplannar points lying on the center of pelvis and both hip joints of the articulated figure, respectively, to estimate the initial camera configuration with least squares approximation.

5.4 Motion Smoothing

In practice, it is very difficult to track the 2D features in a video precisely without any explicit marker on the human body. Thus, the reconstructed joint orientation contains the jerkiness caused by the noisy 2D features. To reduce such jerkiness, we combine motion displacement mapping [3, 24] with Multilevel B-spline fitting [13, 14]. A motion displacement map describes the difference between two motions. In our case, the displacement map between the configuration of the reference motion $\mathbf{x}^r(t)$ and the recovered configuration $\mathbf{x}(t)$ is defined as $\mathbf{d}(t) = \mathbf{x}(t) \ominus \mathbf{x}^r(t)$, that is,

$$\mathbf{d}(t) = \begin{bmatrix} \mathbf{0}_3 \\ \mathbf{v}_1(t) \\ \vdots \\ \mathbf{v}_n(t) \end{bmatrix} = \begin{bmatrix} \mathbf{0}_3 \\ \mathbf{q}_1(t) \\ \vdots \\ \mathbf{q}_n(t) \end{bmatrix} \ominus \begin{bmatrix} \mathbf{0}_3 \\ \mathbf{q}_1^r(t) \\ \vdots \\ \mathbf{q}_n^r(t) \end{bmatrix} = \begin{bmatrix} \mathbf{0}_3 \\ \ln((\mathbf{q}_1^r(t))^{-1} \mathbf{q}_1(t)) \\ \vdots \\ \ln((\mathbf{q}_n^r(t))^{-1} \mathbf{q}_n(t)) \end{bmatrix}, \quad (6)$$

where $\mathbf{v}_i(t) \in \mathbb{R}^3$ is the rotation vector of the i -th joint for $1 \leq i \leq n$. Thus, a new configuration can be recovered by adding the displacement map to the original motion as $\mathbf{x}(t) = \mathbf{x}^r(t) \oplus \mathbf{d}(t)$, that is,

$$\mathbf{x}(t) = \begin{bmatrix} \mathbf{0}_3 \\ \mathbf{q}_1^r(t) \\ \vdots \\ \mathbf{q}_n^r(t) \end{bmatrix} \oplus \begin{bmatrix} \mathbf{0}_3 \\ \mathbf{v}_1(t) \\ \vdots \\ \mathbf{v}_n(t) \end{bmatrix} = \begin{bmatrix} \mathbf{0}_3 \\ \mathbf{q}_1^r(t) \exp(\mathbf{v}_1(t)) \\ \vdots \\ \mathbf{q}_n^r(t) \exp(\mathbf{v}_n(t)) \end{bmatrix}, \quad (7)$$

From the joint orientation displacement $\mathbf{d}(i)$ at each frame i , we compute a smooth displacement map $\mathbf{d}(t)$ that approximates $\mathbf{d}(i)$ for all i . We employ multilevel B-spline approximation technique [14], which uses a series of B-spline functions with different knot spacings on the same interval. In contrast to the local curve fitting with B-splines, the hierarchical structure of the multilevel B-spline fitting can make a smooth shape without undulations, by globally propagating errors at coarse levels and adding details at fine levels. The function from the coarsest knot sequence provides a rough approximation, which is further refined by the functions derived from subsequent finer knot sequences. Finally, we apply the smooth displacement map $\mathbf{d}(t)$ to the reference configuration $\mathbf{x}^r(t)$ to achieve the final configuration $\mathbf{x}^c(t)$ as follows:

$$\mathbf{x}^c(t) = \mathbf{x}^r(t) \oplus \mathbf{d}(t). \quad (8)$$

This procedure is illustrated in Figure 5. Here, we need to trade off the smoothness of the recovered joint orientations against their approximation accuracy, depending

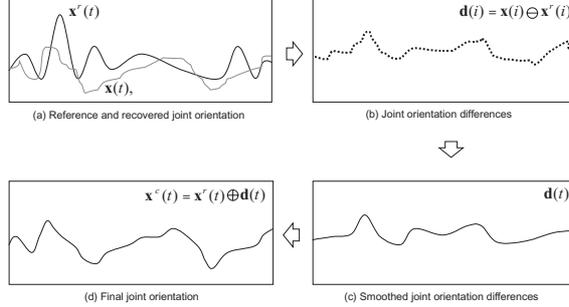


Figure 5: Smoothing the joint orientation. (a) Curves represent the one component of a unit quaternion of the reference motion and the recovered posture, respectively. (b) Joint orientation differences $\mathbf{d}(i)$ (c) Smoothed joint orientations $\mathbf{d}(t)$ that approximates the differences. (d) Sum of the reference configuration and smoothed differences.

on the quality of 2D features. By properly choosing the resolution of knots for noise filtering, we reconstruct a smooth motion even with noisy 2D features while keeping acceptable accuracy.

6 Root Position Estimation

In the previous sections, we have described a method to recover the joint orientations of the articulated figure. Now, we construct a proper trajectory of the root segment to complete the reconstruction process. The final motion $\mathbf{m}(t)$ is the direct sum of the joint orientations $\mathbf{x}(t)$ and the displacement map $\mathbf{d}(t)$ that describes only the translational movement of the root segment:

$$\mathbf{m}(t) = \mathbf{x}(t) \oplus \mathbf{d}(t) = \begin{bmatrix} \mathbf{0}_3 \\ \mathbf{q}_1(t) \\ \vdots \\ \mathbf{q}_n(t) \end{bmatrix} \oplus \begin{bmatrix} \mathbf{p}_1(t) \\ \mathbf{0}_3 \\ \vdots \\ \mathbf{0}_3 \end{bmatrix} = \begin{bmatrix} \mathbf{p}_1(t) \\ \mathbf{q}_1(t) \\ \vdots \\ \mathbf{q}_n(t) \end{bmatrix}, \quad (9)$$

where $\mathbf{p}(t)$ is the root trajectory in the global frame. Since the camera may move along with the actors, the actual trajectory of the actor is hard to capture only with the information given in the video. We try to construct a plausible root trajectory while satisfying the user-specified constraints and the dynamic property of the reference motion.

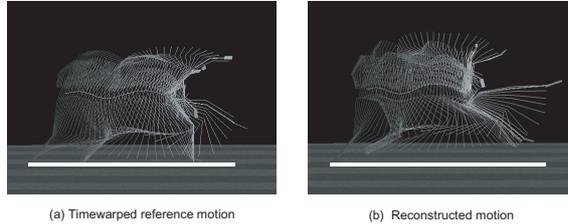


Figure 6: An artifact of direct substitution of the root trajectory. The artificial ground line is drawn for visual aids. (a) Timewarped reference motion. (b) A motion using the root trajectory of the reference motion. The motion looks like floating above the ground.

We discriminate two classes of motions according to their interaction with the environment, which is the source of constraints. In the first case, we deal with the motion that exhibits some interaction between the actor and his/her surrounding environment. Locomotion is a typical example since the feet of the actor contact with the ground. In the second case, we treat a motion that does not show such interaction, for which jumping motion is a typical example. In each case, we describe how to obtain the displacement map $\mathbf{d}(t)$, in particular, the root position $\mathbf{p}_1(t)$.

6.1 Case 1: Motion involving interaction with the environment

Consider the kick motion of a soccer player as shown in Figure 6. The reconstructed motion is so dynamic that the root trajectory is quite different in height from that of the reference motion. Therefore, we adjust the height of the root segment to make the stance foot contact with the ground at every constrained frame. We start with the motion $\hat{\mathbf{m}}(t) = \mathbf{x}^c(t) \oplus (\mathbf{p}_1^r(t), \mathbf{0}_3, \dots, \mathbf{0}_3)^T$, where $\mathbf{x}^c(t)$ and $\mathbf{p}_1^r(t)$ are the recovered joint orientations as explained in Section 5.4 and the root trajectory of the timewarped reference motion, respectively. We compute the distance $\mathbf{d}(i)$ between the stance foot in the motion $\hat{\mathbf{m}}(t)$ and the ground at every constrained frame i , and then construct a smooth displacement map $\mathbf{d}(t)$ that approximates $\mathbf{d}(i)$ using the multilevel B-spline approximation method [14]. With $\hat{\mathbf{m}}(t) \oplus (\mathbf{d}(t), \mathbf{0}_3, \dots, \mathbf{0}_3)^T$ as the initial guess, we adopt the motion retargetting technique [13] to determine the final target motion $\mathbf{m}(t)$.

6.2 Case 2: Motion without involving interaction with the environment

Unlike the previous case, a motion in this case does not involve any interaction with the environment as observed in a jump motion. We exploit a dynamics property of the reference motion to determine the root trajectory. Given an initial velocity with no external forces except for gravity, the center of gravity(COG) of an object follows a parabolic trajectory. We measure the initial velocity of the actor at the moment to start a motion. The COG trajectory $\mathbf{cog}^r(t)$ of the reference motion is defined as follows:

$$\mathbf{cog}^r(t) = \mathbf{p}^r(t) + \frac{\sum_{i=1}^n m_i \tilde{\mathbf{p}}_i^r(t)}{\sum_{i=1}^n m_i}, \quad (10)$$

where $\tilde{\mathbf{p}}_i^r(t)$ and m_i for $1 \leq i \leq n$ represent the vector from the root position $\mathbf{p}_1^r(t)$ to the COG of the i -th link and its corresponding mass, respectively. Since the reference motion is timewarped linearly as described in Section 4, we also linearly scale the COG trajectory of the reference motion to approximate the COG trajectory $\mathbf{cog}(t)$ of the motion to reconstruct as follows:

$$\mathbf{cog}(t) = s \mathbf{cog}^r(t) = \mathbf{p}(t) + \frac{\sum_{i=1}^n m_i \tilde{\mathbf{p}}_i(t)}{\sum_{i=1}^n m_i}, \quad (11)$$

where s and $\tilde{\mathbf{p}}_i(t)$ for $1 \leq i \leq n$ denote the scaling factor for timewarping at the initial frame and the vector from the root position $\mathbf{p}(t)$ to the COG of the i -th link, respectively. $\tilde{\mathbf{p}}_i(t)$ can be obtained from the recovered joint orientations $\mathbf{x}^c(t)$. Thus, the root trajectory $\mathbf{p}(t)$ is computed by combining Equations 10 and 11:

$$\mathbf{p}(t) = s \mathbf{p}^r(t) + s \left(\frac{\sum_{i=1}^n m_i (\tilde{\mathbf{p}}_i^r(t) - \tilde{\mathbf{p}}_i(t))}{\sum_{i=1}^n m_i} \right). \quad (12)$$

As shown in Figure 7, the reconstructed motion with the root trajectory of the reference motion follows an infeasible, distorted COG trajectory. However, that with the COG trajectory of the reference motion follows a smooth parabolic trajectory.

7 Experimental Results

We use a human model of 40 DOFs: 6 DOFs for the pelvis position and orientation, 3 DOFs for the chest, 3 DOFs for the neck, and 7 DOFs for each limb. The motion clips are sampled at the rate of 60 frames per second, and their keytimes are obtained by interactively specifying the moments of interaction. To demonstrate

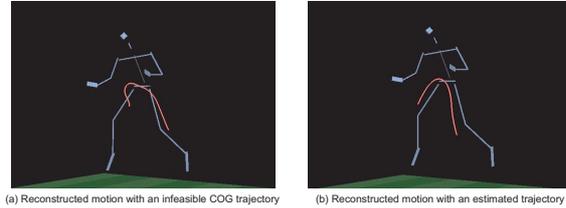


Figure 7: Estimating a trajectory of the root segment during a jump motion. The curve represents the COG trajectory of the motion. (a) The reconstructed result using the root trajectory of the reference motion. (b) The reconstructed result using the COG trajectory of the reference motion after timewarping.

the effectiveness of our approach in real situations, we recover shooting motions of soccer players from videos. Table 1 shows the repertoire of motion clips in our motion library. Each motion clip contains a highly dynamic motion of a short duration, that is, typically 2~3 seconds.

We use video clips available on public sites for soccer games. We interactively mark the 2D features in the video, each of which corresponds to a projected joint position of our human model. We also specify the keytimes interactively to build the time correspondence between the reference motion and the video stream.

Our reconstruction method is implemented in C++ on top of MS Windows 2000TM and the TGS Open InventorTM that is a convenient toolkit to support 3D graphics primitives. Experiments are performed on a Pentium PC (Intel PentiumIII 733MHz processor and 512MB memory).

7.1 Kick motion of a soccer player

Figure 8 shows the kick motion of a soccer player, which we reconstructed with a reference motion clip, “place instep kick”. The video clip used in this example contains 51 frames. The camera tracks the player from the righthand side. There are six keytimes specified in this case: four for the left foot and two for the right foot. Since the target motion takes more inclined postures than the reference motion, the root trajectory has been adjusted to keep the interaction between every stance foot and the ground. Table 2 illustrates the error $e(t)$ of our reconstruction

method defined as follows:

$$e(t) = \sum_{i=1}^n \|\bar{\mathbf{p}}_i(t) - \mathbf{P}_c \mathbf{f}_i(\mathbf{x}^c(t))\|^2, \quad (13)$$

where $\bar{\mathbf{p}}_i(t)$ and $\mathbf{x}^c(t)$ represent the 2D feature position of i -th joint in the image and the final joint configuration at time t , respectively. As shown in Table 2, the timewarping reduces the maximum error remarkably. While the unit knot spacing provides a motion that is exactly matched with the features in the video, there exists some jerkiness due to the noisy 2D features. A knot spacing of four gives a smooth motion without significant increase in the average error. It takes 19.34 seconds to reconstruct the kick motion excluding 2D feature marking. The execution time mainly depends on the number of frames in numerical optimization for joint orientation recovery.

7.2 Heading motion of a soccer player

Figure 9 shows the heading motion of a soccer player, which we reconstructed with a reference motion clip, “left jump heading with both feet”. The video clip used in this example contains 37 frames. In this example, the camera tracks the player from the rear. We specify five keytimes in this case: two for each foot, and one for the head. Since the target motion takes a free flight during the motion with the external force of gravity, the root trajectory have been adjusted to keep the dynamic balance during the jump. The error defined by Equation 13 is shown in Table 2. Excluding the time for 2D feature marking, it takes 14.86 seconds to reconstruct the heading motion.

8 Conclusion

In this paper, we have presented a novel method for reconstructing human motions from a single video sequence using a motion library. Our reconstruction method takes three major steps: timewarping to align the reference motion with that in the video, reconstructing the joint orientations, and estimating the root trajectory. Since the motion reconstruction problem is coupled with that of camera posture computation, we also describe a simple camera model to recover the camera configuration from the video sequence for our purpose. Exploiting a 3D motion in the library, we can resolve the inherent depth ambiguities in the reconstruction problem. The reconstructed motion is smooth even with noisy features. Moreover, our method can reconstruct highly dynamic motions, in which the time coherence is

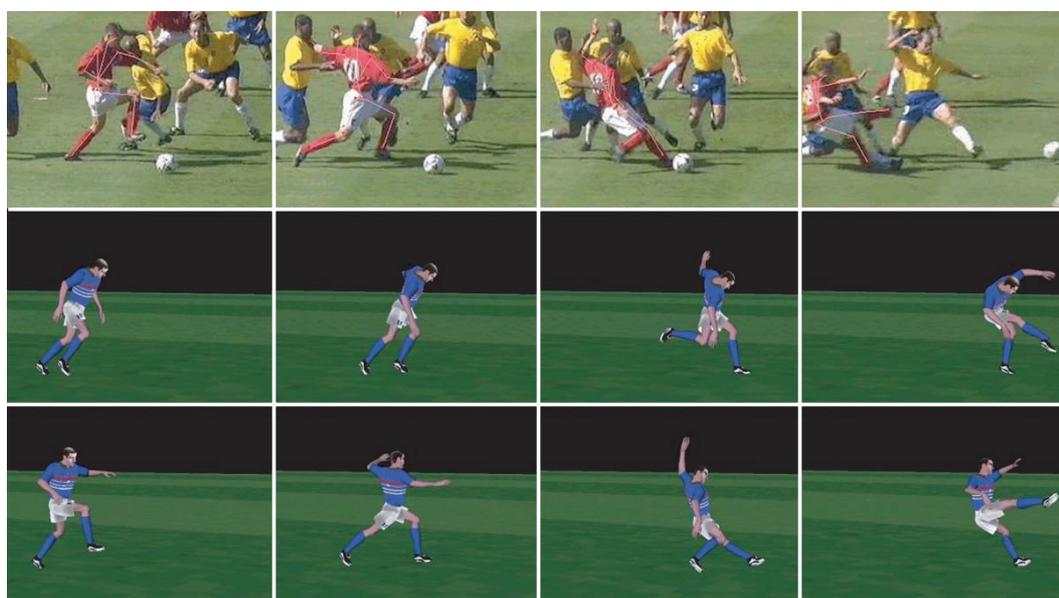


Figure 8: An example with a kick motion. (Top) Input image sequence with feature points. (Middle) Reference motion. (Bottom) Reconstructed motion.



Figure 9: An example with a heading motion. (Top) Input image sequence with feature points. (Middle) Reference motion. (Bottom) Reconstructed motion.

Table 1: Reference motions : kicking and heading motions of soccer players. († represents half-volley) This classification is based on the configuration of legs and heads with respect to ball position.

| category (kicks) | ball placement | | | | category (headings) | head direction | | |
|---------------------|----------------|------------|--------|---------|------------------------|----------------|------|-------|
| | place | volley(h)† | volley | sliding | | front | left | right |
| instep | o | o | o | - | stand | o | o | o |
| inside | o | o | - | o | jump (single foot) | o | o | o |
| outside | o | o | o | o | jump (both feet) | o | o | o |
| toe | o | - | - | - | stand back | o | o | o |
| hill | o | - | - | - | jump back | o | o | o |
| overhead | - | - | o | - | | | | |
| cut | o | - | - | - | | | | |
| turning | o | - | - | - | | | | |

Table 2: Error analysis data. We compute the minimum, maximum, and average error over all frames. The errors are measured in the normalized coordinate of the image, that is, all the coordinate values are in the interval [-1,1].

| | kick motion (51 frames) | | | heading motion (37 frames) | | |
|--------------------------------|-------------------------|--------|--------|----------------------------|--------|--------|
| | min. | max. | avg. | min. | max. | avg. |
| original | 0.0331 | 0.3049 | 0.1040 | 0.0116 | 0.1918 | 0.0846 |
| timewarped | 0.0272 | 0.1624 | 0.0915 | 0.0100 | 0.1397 | 0.0710 |
| reconstructed (knot spacing:1) | 0.0011 | 0.0100 | 0.0043 | 0.0005 | 0.0100 | 0.0051 |
| reconstructed (knot spacing:4) | 0.0013 | 0.0227 | 0.0088 | 0.0008 | 0.0159 | 0.0058 |

weak in the video. Provided with a feasible set of motions as a library, our method can be used to obtain a wide variety of motions in real situations such as sports events. We have demonstrated the effectiveness of our method by reconstructing shooting motions of soccer players from real videos.

References

- [1] A. Azarbayejani, C. Wren, and A. Pentland. Real-time 3-d tracking of the human body. *Proceedings of IMAGE'COM*, May 1996.
- [2] C. Bregler and J. Malik. Estimation and tracking kinematic chains. *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1998.
- [3] A. Bruderin and L. Williams. Motion signal processing. *Computer Graphics (Proceedings of SIGGRAPH 95)*, 29(4):97–104, 1995.
- [4] M. F. Cohen. Interactive spacetime control for animation. *Computer Graphics (Proceedings of SIGGRAPH 92)*, 26(2):293–302, July 1992.
- [5] R. Demori and D. Probst. *Handbook of Pattern Recognition and Image Processing*. Academic Press, 1 edition, 1986.
- [6] D.E. Difranco, T.J. Cham, and J.M. Rehg. Recovery of 3d articulated motion from 2d correspondences. *Cambridge Research Laboratory TR-CRL-99-7*, 1999.
- [7] R. Fletcher. *Practical Methods of Optimization*. Wiley and Sons, 1 edition, 1980.
- [8] P. E. Gill and W. Murray. *Numerical Methods for Constrained Optimization*. Academic Press, 1 edition, 1974.
- [9] M. Gleicher. Motion editing with spacetime constraints. *Proc. Symp. Interactive 3D Graphics*, pages 139–148, 1997.
- [10] M. Gleicher. Retargetting motion to new characters. *Computer Graphics (Proc. SIGGRAPH '98)*, 32:33–42, July 1998.
- [11] N.R. Howe, M.E. Leventon, and W.T. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. *Cambridge Research Laboratory TR-CRL-99-37*, 1999.

- [12] I. Kakadiaris and D. Metaxas. Model-based estimation of 3d human motion with occlusion based on active multi-viewpoint selection. *In Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 1996.
- [13] J. Lee and S. Y. Shin. A hierarchical approach to interactive motion editing for human-like figures. *Computer Graphics (Proc. SIGGRAPH '99)*, 33:395–408, August 1999.
- [14] S. Lee, G. Wolberg, and S. Y. Shin. Scattered data interpolation with multi-level b-splines. *IEEE Trans. Visualization and Computer Graphics*, 3(3):228–244, 1997.
- [15] T. Noma, K. Oishi, H. Futsuhara, H. Baba, Takeshi Ohashi, and Toshiaki Ejima. Motion generator approach to translating human motion from video to animation. *Pacific Graphics*, October 1999.
- [16] Z. Popovic and A. Witkin. Physically-based motion transformation. *Computer Graphics (Proc. SIGGRAPH '99)*, 33:11–20, 1999.
- [17] W. H. Press, Saul A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, 2 edition, 1992.
- [18] J. M. Rehg and T. Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. *European Conf. on Computer Vision*, pages 35–46, May 1994.
- [19] J.M. Rehg and T. Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. *In Proc. European Conference on Computer Vision*, 2, 1994.
- [20] C. Rose, B. Guenter, B. Bodenheimer, and M. F. Cohen. Efficient generation of motion transitions using spacetime constraints. *Computer Graphics (Proceedings of SIGGRAPH 96)*, 30:147–154, August 1996.
- [21] K. Shoemake. Animating rotation with quaternion curves. *Computer Graphics (Proceedings of SIGGRAPH 85)*, 19:245–254, 1985.
- [22] S. Tak, O. Song, and H. Ko. Motion balancing filtering. *Computer Graphics Forum*, 19(3):437–446, August 2000.
- [23] A. Witkin and M. Kass. Spacetime constraints. *Computer Graphics (Proceedings of SIGGRAPH 88)*, 22(4):159–168, August 1988.

- [24] A. Witkin and Z. Popovic. Motion warping. *Computer Graphics (Proc. SIGGRAPH '95)*, 29:105–108, August 1995.
- [25] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Trans. Pattern Analysis and Machine Intelligence*, pages 780–785, July 1991.
- [26] J. Y. Zheng and S. Suezaki. A model based approach in extracting and generating human motion. *Proceedings of Fouteenth International Conference on Pattern Recognition*, 2:1201–1205, 1998.