3D-2D Spatiotemporal Registration for Sports Motion Analysis

Ruixuan Wang, Wee Kheng Leow, Hon Wai Leong
Dept. of Computer Science, National University of Singapore, Computing 1, Singapore 117590
{wangruix, leowwk, leonghw}@comp.nus.edu.sg

Abstract

Computer systems are increasingly being used for sports training. Existing sports training systems either require expensive 3D motion capture systems or do not provide intelligent analysis of user's sports motion. This paper presents a framework for affordable and intelligent sports training systems for general users that require only single camera to record the user's motion. Sports motion analysis is formulated as a 3D-2D spatiotemporal motion registration problem. A novel algorithm is developed to perform spatiotemporal registration of the expert's 3D reference motion and a performer's 2D input video, thereby computing the deviation of the performer's motion from the expert's motion. The algorithm can effectively handle ambiguous situations in a single video such as depth ambiguity of body parts and partial occlusion. Test results show that, despite using only single video, the algorithm can compute 3D posture errors that reflect the performer's actual motion error.

1. Introduction

Computer systems are increasingly being used for sports training. Two kinds of computer-aided sports training systems are commercially available: 3D motion-based systems and 2D video-based systems. A 3D motion-based system [20, 29] uses multiple cameras to track the motion of reflective markers attached to the performer's body. The markers' 3D positions are recovered and used to compute the performer's 3D motion, which can be analyzed by the coach or compared with a 3D reference motion of an expert. Such a system can provide accurate motion analysis. However, it is very expensive and difficult to use for the general users.

A 2D video-based system [14, 21, 24, 28] captures the performer's motion using an off-the-shelf video camera and loads the video into a computer system. The system displays the performer's video and a pre-recorded expert's video side by side, and provides tools for the user to manually compare the performer's motion with the expert's motion. The system is affordable to general users. However, it cannot perform detailed motion analysis automatically.

To overcome the shortcomings of existing systems, this paper proposes a framework for affordable and intelligent sports training systems for general users that require only single stationary camera to record the user's motion. Sports motion analysis is formulated as a 3D-2D spatiotemporal motion registration problem (Section 3). A novel algorithm is developed to perform spatiotemporal matching of the 3D reference motion of an expert and the 2D input video of a performer, thereby computing the deviation of the performer's motion from the expert's motion (Sections 4–7). The algorithm can effectively handle ambiguous situations in single video such as depth ambiguity of body parts and partial occlusion. It can be applied to analyze different types of sports motion. Extensive test results show that the algorithm can compute 3D posture errors that reflect the performer's actual motion error using only single video.

In principle, videos of human motion can be recorded by multiple cameras, which may remove depth ambiguity. Nevertheless, we propose to work on the case of single-video motion analysis, which is technically more challenging. Once the algorithm is developed, extending it to multiple-video analysis would be a relatively simple task. To our best knowledge, this is the first attempt at automatic computer analysis of 3D sports motion in a single video.

2. Related Work

Our 3D-2D spatiotemporal registration problem for sports motion analysis is closely related to several known research topics, namely human body tracking, human posture estimation, and video sequence alignment. However, there are fundamental differences between them. Human body tracking [3, 19, 23], in general, performs spatial matching between consecutive images in the input sequence without using 3D reference motion. Human posture estimation infers the 2D or 3D body posture from single or multiple images without solving temporal correspondence. Human body tracking methods often apply human posture estimation techniques [13, 22, 27]. Video sequence alignment [4, 16] solves for the temporal correspondence between two sequences without posture matching and 3D motion information. Our proposed problem involves both temporal cor-

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

respondence and posture matching, which is much more complex than the related problems. In the following, existing work in the most related area, human posture estimation, is discussed in more detail.

Two general approaches exist for solving human posture estimation problem: model-free and model-based. Model-free approach does not use explicit human body model. One method is to train a nonlinear mapping function to map from image features to body postures [1, 6, 13, 27]. The other method is to store a set of exemplar images with known 3D postures, and estimate the posture in the input image by searching for the exemplar that is most similar to the input image [2, 8, 18]. These methods are useful only for a small set of body postures due to the complexity of human postures. They can recover only the body postures that are similar to those in the training images and exemplars.

Model-based approach estimates body posture by synthesizing possible postures from a model and matching them to the input images. Within this approach, continuous methods use continuous optimization algorithms to efficiently find locally optimal posture estimates [22]. They cannot guarantee that the solutions are globally optimal.

Probabilistic methods, which include particle filtering (CONDENSATION), Markov Chain Monte Carlo, and Belief Propagation, use sampling techniques to estimate body postures [5, 12, 15, 26]. With enough samples, these methods can potentially obtain the globally optimal solution. The main difficulty of these methods is to search a very high-dimensional space for the globally optimal solution. To tackle this problem, Belief Propagation (BP) methods decompose the high-dimensional search problem into a set of low-dimensional problems by estimating the pose of each body part individually [15, 26]. BP method is adapted and extended in our algorithm framework.

3. Problem Formulation

To clearly describe the problem, it is necessary to first describe the inputs of the problem, which consist of 3D reference motion of the expert and 2D input video of the performer (Section 3.1), and the complex relationships (Section 3.2) between them.

The 3D reference motion of the expert includes:

1. Time-independent component: human body model
   The human body model H consists of a hierarchical skeleton model of bones and joints, and a triangular mesh model for the shapes of the body parts.

2. Time-dependent component: 3D motion data
   The 3D motion data comprises a temporal sequence of global positions $\mathbf{p}_t$ of human body in the world coordinate system, and joint angles $\boldsymbol{\theta}_t$ of the body parts. These data define the *reference posture* (Fig. 2(c)) at time t denoted as B_t, i.e., $\mathbf{p}_t, \boldsymbol{\theta}_t \in B_t$. The sequence M of B_t, $t = 0, \ldots, L$, together with the human body model H, defines the 3D reference motion, which is assumed to be retargetted to the performer's body in the input video using, e.g., the algorithm in [7].

The motion m' of a performer is captured in the input video, which consists of a sequence of image frames $I'_{t'}$ (Fig. 2(a)) over time $t' = 0, \ldots, L'$. Typically, $L' < L$ because video camera has a lower sampling rate than 3D motion capture system. Each input image $I'_{t'}$ contains the image of a performer generated by the projection of an unknown *performer's posture* $B'_{t'}$ onto the image plane. The human body region $S'_{t'}$ in image $I'_{t'}$ is separated from the background using automatic segmentation and skin color detection algorithms [11, 17] (Fig. 2(b)). Note that in a single camera view, depth ambiguity of body parts and self-occlusion can occur.

There are many complex spatiotemporal relationships between the 3D reference motion and the 2D input video. Three major relationships are highlighted below.

1. Temporal Difference: The performer's motion can differ from the expert's motion in terms of execution speed. So, a temporal correspondence C needs to be established from 2D video time t' to 3D motion time t, i.e., $C(t')$ is a particular t that corresponds to t'. C should satisfy the temporal order constraint: for any two postures in the performer's motion, the two corresponding postures in the reference motion have the same temporal order. Without loss of generality, it is assumed that $C(0) = 0$ and $C(L') = L$.

2. Spatial Difference: The performer's (unknown) posture $B'_{t'}$ can differ from the expert's posture $B_{C(t')}$ at the corresponding time frame by a global rigid transformation T and a joint articulation A, i.e., $B'_{t'} = A_{t'}(T_{t'}(B_{C(t')}))$. In the algorithm, $B'_{t'}$ is inferred by registering the projection P of $A_{t'}(T_{t'}(B_{C(t')}))$ to the input body region $S'_{t'}$ in image $I'_{t'}$. Then, the posture error $\varepsilon_{t'}$ is naturally captured in $A_{t'}$ and $T_{t'}$.

3. Smooth Motion: The posture error $\varepsilon_{t'}$ can be large when the performer's motion differs significantly from the reference motion. Nevertheless, the rate of change of posture errors should remain small because the motion of interest is smooth. That is, $\Delta\varepsilon_{t'}/\Delta t'$ is small.

Now, we can formulate the problem of spatiotemporal registration for sports motion analysis as follows:

> Given the reference motion $M = \{B_t\}$ and the input motion $m' = \{S'_{t'}\}$, determine the temporal correspondence C, projection P, rigid transformation $T_{t'}$, and join articulation $A_{t'}$ that minimize the errors E_S and E_D:

$$E_S = \frac{1}{L'+1} \sum_{t'} d_S(P(A_{t'}(T_{t'}(B_{C(t')}))), S'_{t'}), \quad (1)$$

$$E_D = \frac{1}{L'+1} \sum_{t'} \varepsilon_{t'}. \quad (2)$$

E_S is the registration error, where d_S is an appropriate difference measure. The total posture error E_D is minimized to capture the idea of computing the minimum correction required by the performer to match the expert's motion.

The minimization of E_S and E_D is subjected to the following constraints:

A. Joint angle limit. The valid angle between two connected body parts is physically limited to certain ranges.

B. Temporal order constraint. For any t'_1 and t'_2 such that $t'_1 < t'_2$, $C(t'_1) < C(t'_2)$.

C. Small rate of change of posture errors. For each t', $\Delta\varepsilon_{t'}/\Delta t'$ is small.

4. Spatiotemporal Registration Framework

It is infeasible to directly solve the proposed problem, which is a very complex high-dimensional optimization problem with long time sequence. So, it is decomposed into four subproblems and solved in the following stages:

1. Estimation of camera projection P.
   This stage can be performed using standard calibration algorithm. So, it is omitted in this paper.

2. Estimation of approximate temporal correspondence C and rigid transformation T.
   Determine initial estimates of C and $T_{t'}$ that minimize the error E_C subject to Constraint B:

$$E_C = \frac{1}{L'+1} \sum_{t'=0}^{L'} d_S(P(T_{t'}(B_{C(t')})), S'_{t'}), \quad (3)$$

   Joint articulation $A_{t'}$ is omitted in this stage.

3. Estimation of posture candidates.
   Due to depth ambiguity, multiple postures can match an input body region in the image. So, this stage determines, for each t', multiple $A_{t'l}$ and $T_{t'l}$ that minimize the error $E_{t'}$ subject to Constraint A:

$$E_{t'} = d_S(P(A_{t'l}(T_{t'l}(B_{C(t')}))), S'_{t'}). \quad (4)$$

   The approximate C estimated in the previous stage is used to identify approximate corresponding reference posture $B_{C(t')}$, which is transformed by $A_{t'l}$ and $T_{t'l}$ to match the input body region $S'_{t'}$. This approach avoids the accumulation of estimation error over time, which is present in many human body tracking methods. The resulting $\mathcal{B}_{t'} = \{B'_{t'l}\}$, where $B'_{t'l} = A_{t'l}(T_{t'l}(B_{C(t')}))$, is the set of posture candidates that match $S'_{t'}$ well.

4. Candidate selection and refinement of estimates.
   Select the best posture candidate $B'_{t'}$ from $\mathcal{B}_{t'}$ and determine the C that together minimize E_D subject to Constraints B and C. After finding the best $B'_{t'}$, posture error can be computed as the difference between $B'_{t'}$ and the corresponding $B_{C(t')}$.

The algorithms for Stages 2, 3, and 4 are discussed in the following sections.

5. Estimation of Temporal Correspondence

This stage estimates approximate temporal correspondence C and transformation T using dynamic programming (DP). Actually, DP is guaranteed to produce globally optimal solution. However, the optimal solution at this stage is not globally optimal for the whole problem because articulation is omitted. So, the temporal correspondence estimated at this stage is only an approximation.

Let $d(t', C(t'))$ denote $d_S(P(T_{t'}(B_{C(t')})), S'_{t'})$. The task is to determine C by minimizing E_C:

$$E_C = \frac{1}{L'+1} \sum_{t'=0}^{L'} d(t', C(t')) \quad (5)$$

subject to temporal order constraint. Given a particular C, $T_{t'}$ at each time t' is determined using sampling technique. The DP problem is formulated as follows.

Let $\mathbf{D}$ denote a $(L'+1)\times(L+1)$ correspondence matrix. Each matrix element at (t', t) represents the possible frame correspondence between t' and t, and the correspondence cost is $d(t', t)$. A path in $\mathbf{D}$ is a sequence of frame correspondences for $t' = 0, \ldots, L'$ such that each t' has a unique corresponding $t = C(t')$, with $C(0) = 0$ and $C(L') = L$. The cost of a path is the sum of the correspondence costs over all t', and the average path cost is E_C. The problem is to find the least cost path on which E_C is minimized.

The least cost path can be efficiently found by making use of the temporal order constraint. Suppose the frame pair

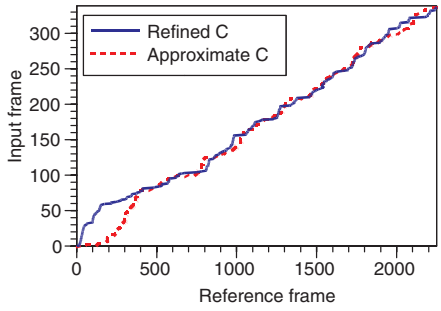


Figure 1. Approximate and refined temporal correspondence C.

(t', t) is on the least cost path. Then, the possible previous frame pair should be one of $(t'-1, t-1-i)$ for $i = 0, \ldots, w$. The temporal window size w is defined as kL/L' for a small $k \geq 1$. k is small because the change of posture error between the pair of corresponding frames over time is small (Section 3.2). The least cost path from the first frame pair $(0, 0)$ to the current pair (t', t) can be determined by recursively computing the least cost path from $(0, 0)$ to one of $(t'-1, t-1-i), i = 0, \ldots, w$.

Let $D(t', t)$ denote the least cost from frame pair $(0, 0)$ up to (t', t) on the least cost path, and $D(0, 0) = d(0, 0)$. Then $D(L', L)$ can be recursively computed as follows:

$$D(t', t) \quad = \quad d(t', t) + \min_{i=0}^{w} D(t'-1, t-1-i) \quad (6)$$

Once $D(L', L)$ is computed, the least cost path is obtained by tracing back the path from $D(L', L)$ to $D(0, 0)$. The least cost path gives the correspondence C (Fig. 1).

6. Estimation of Posture Candidates

Posture candidates are estimated using an extension of Belief Propagation (BP) [9, 10, 15, 25, 26]. The algorithm uses the approximate temporal correspondence C estimated in the previous stage to identify approximate corresponding reference posture $B_{C(t')}$ at time t' (Fig. 2(c)). Then, BP uses $B_{C(t')}$ as an initial estimate to search for the posture candidates that match the input body region $S'_{t'}$ (Fig. 2(b)), thereby determining the candidate articulations $A_{t'l}$ and rigid transformations $T_{t'l}$. First, let us briefly describe BP.

Let $p(B'|S')$ denote the probability that B' is a good posture candidate given input body region S'. Then, posture candidate estimation is to find B' with large $p(B'|S')$. Denote the pose of body part i as b_i, i.e., $B' = \{b_i\}$. Instead of computing $p(B'|S')$ directly, BP iteratively computes $p(b_i|S')$ for each body part i using these equations:

$$p(b_i|S') \propto \phi(b_i, S') \prod_{j \in \Gamma(i)} m_{ji}(b_i) \quad (7)$$

$$m_{ji}(b_i) \propto \int \phi(b_j, S')\ \psi(b_j, b_i) \prod_{k \in \Gamma(j) \setminus i} m_{kj}(b_j)\, db_j \quad (8)$$

where $\Gamma(i)$ is the set of body parts connected to body part i, and $m_{ji}(b_i)$ is the *contribution* of body part j to the pose b_i of body part i. To compute $m_{ji}(b_i)$ and $p(b_i|S')$, the functions $\phi(b_i, S')$ and $\psi(b_i, b_j)$ need to be defined.

The similarity function $\phi(b_i, S')$ measures the degree of match between S' and body part i at pose b_i. Each body part at pose b_i computed in the current iteration is projected and rendered, together with all other body parts whose poses are obtained in the previous iteration, to produce the projected body region S. Then, the similarity is computed as $\phi(b_i, S') = \exp(-d_S(S, S'))$. $d_S(S, S')$ is defined in terms of the amount of overlap between S and S', and the matching of their region edges, which allows the algorithm to handle partial self-occlusion of body parts. In comparison, the original BP [26] measures similarity using only region overlap between the projection of a single body part and the entire input body region. Therefore, it cannot handle partial self-occlusion of body parts.

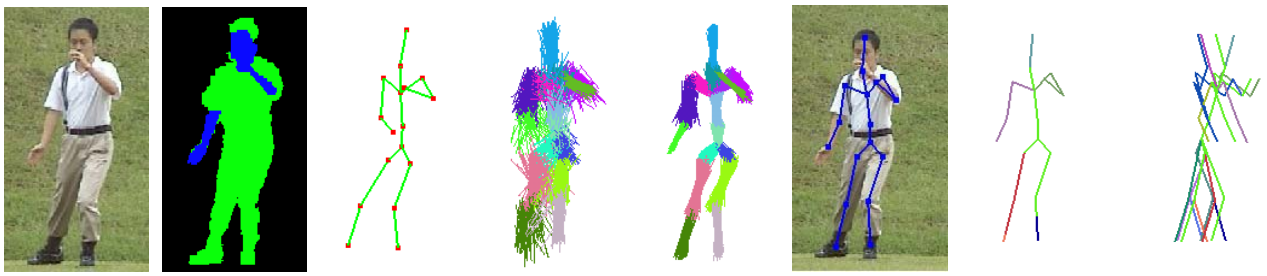
The joint constraint function $\psi(b_i, b_j)$ enforces joint constraint and joint angle constraint between two connected body parts i and j. The joint constraint states that two neighboring body parts should be connected at the joint. Let $\mathbf{x}_i$ and $\mathbf{x}_j$ denote the 3D positions of the points on body parts i and j that connect to form a joint. When body parts i and j adopt poses b_i and b_j, the degree of satisfaction of joint constraint is measured by $\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$, where σ is a positive parameter.

The joint angle constraint ensures that the angle between two connected body parts i and j falls within physical limit. The degree of satisfaction of joint angle constraint is measured by $J(b_i, b_j)$, which is 1 when the joint angle is within limit, and a smaller constant a otherwise. Combining the two constraints, we obtain $\psi(b_i, b_j) = J(b_i, b_j) \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\sigma^2)$.

The parameters σ and a decrease over iteration. At the first few iterations, the pose estimate of each body part may be far from the actual pose. So the constraints are loosely enforced initially to ensure that the correct poses can be included. Gradually, the pose estimate of each body part is expected to become more similar to the actual pose, and therefore the constraints should become more strict.

In practice, evaluation of the integral in Equ. 8 is often intractable with continuous state variable b_i. So, nonparametric sampling technique similar to Belief Propagation Monte Carlo [9] is adopted to compute $m_{ji}(b_i)$.

The BP algorithm described above estimates only the



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

Figure 2. Estimation of posture candidates. (a) Input image. (b) Input body regions. (c) Approximate corresponding reference posture. (d, e) The projections of pose samples of each body part after the 1st and 30th iterations. (f) Posture candidate overlapped onto input image. (g, h) Frontal and side views of all posture candidates. Different pose samples and posture candidates are colored with different colors.

pose samples of each body part (Fig. 2(d, e)). These pose samples are used to generate posture candidates as follows. The first posture candidate is computed such that each body part has the same depth orientation as that in the corresponding reference posture, and its projection matches the mean of its pose samples (Fig. 2(f)). Then, based on the first posture candidate, flip the depth orientation of n body parts about their parent joints, starting with $n = 1$, while keeping the body parts connected at the joints. This step is repeated for $n = 1, 2, \ldots$, until enough posture candidates are generated. These posture candidates have exactly the same frontal projection (Figure 2(g)) but different side projections (Figure 2(h)). Therefore, they capture all possible depth ambiguities in the image of a single camera view.

7. Refinement of Estimates

This stage selects the best posture candidate at each t' that minimize the error E_D (Eq. 2), and simultaneously refines the temporal correspondence C. Let $\ell(t')$ denote the index of the best posture candidate at t'. Then, the problem is to determine the ℓ and C that minimize E_D subject to Constraints B and C. Constraint C, i.e., small rate of change of posture errors, can be incorporated into E_D to obtain E_F:

$$E_F = \frac{1}{L'}\sum_{t'=0}^{L'}[d_c(t', C(t'), \ell(t')) + \lambda\, d_s(t', C(t'), C(t'-1), \ell(t'), \ell(t'-1))], \tag{9}$$

where λ is a weighting factor. The difference d_c is obtained from E_D, i.e., $d_c(t', t, l') = \varepsilon_{t'} = d_B(B_t, B'_{t'l'})$. d_B is the posture error between the posture candidate $B'_{t'l'}$ and the reference posture B_t, which is defined as the mean orientation difference of all the body parts in the postures. The difference $d_s(t', t, s, l', k')$ measures the change of posture errors between two pairs of corresponding postures $(B'_{t'l'}, B_t)$ and $(B'_{t'-1,k'}, B_s)$:

$$d_s(t', t, s, l', k') = [d_B(B_t, B'_{t'l'}) - d_B(B_s, B'_{t'-1,k'})]^2. \tag{10}$$

DP technique similar to that in Section 5 is developed to determine the optimal $\ell(t')$ and $C(t')$. In this case, the correspondence matrix $\mathbf{D}$ is a $(L'+1)\times(L+1)\times N_B$ matrix, where N_B is the maximum number of posture candidates at each t'. Each matrix element at (t', t, l') represents the possible correspondence between posture candidate $B'_{t'l'}$ and reference posture B_t. The correspondence cost consists of two terms: $d_c(t', t, l')$ and $d_s(t', t, s, l', k')$. A path in $\mathbf{D}$ is a sequence of correspondences for $t' = 0, \ldots, L'$ such that each t' has a unique corresponding $t = C(t')$ and $l' = \ell(t')$. The cost of a path is the sum of the correspondence costs over all t', and the average path cost is E_F. The problem is to find the least cost path on which E_F is minimized.

Let $D(t', t, l')$ denote the least cost from the triplet $(0, 0, \ell(0))$ up to (t', t, l') on the least cost path, and $D(0, 0, \ell(0)) = d_c(0, 0, \ell(0))$. Then, by a similar reasoning as illustrated in Section 5, $D(L', L, \ell(L'))$ can be computed recursively using the formulae

$$D(t', t, \ell(t')) = \min_{l'} D(t', t, l') \tag{11}$$

$$\ell(t') = \arg\min_{l'} D(t', t, l') \tag{12}$$

$$D(t', t, l') = d_c(t', t, l') + \min_{i,k'}\{D(t'-1, t-1-i, k') + d_s(t', t, t-1-i, l', k')\}. \tag{13}$$

Once $D(L', L, \ell(L'))$ is computed, the least cost path is obtained by tracing back the path from $D(L', L, \ell(L'))$ to $D(0, 0, \ell(0))$. Test result in Fig. 1 shows that the globally optimal C is not a linear function.

8. Experiments and Discussions

Due to the unavailability of ground truth data of the performer's motion, it is impossible to directly measure the accuracy of the whole spatiotemporal registration algorithm. Even if a 3D motion capture system is available, the acquisition of the performer's 3D motion still requires great skills and experience in order to minimize human error during the capture process. Instead, we split the test into two

phases. First, synthetic data were used to assess the accuracy of the posture candidate estimation algorithm. The test results gave an estimate of the algorithmic error in estimating the performer's actual 3D postures from 2D input images. Next, the whole spatiotemporal registration algorithm was tested on real data to measure the performer's posture error. As long as the measured error is significantly greater than the algorithmic error, we are confident that the measured error reliably reflects the actual posture error of the performer. Two sets of motion sequences were used for the tests: (1) 3D Taichi reference motion with 2250 reference postures and input video with 339 input images, and (2) 3D golf swing motion with 250 reference postures and input video with 51 input images.

In this test, synthetic test data were generated as follows. 110 reference postures were selected at regular intervals from the 3D Taichi sequence. Each selected 3D posture was mapped to an articulated 3D human model, which was projected to obtain a synthetic input image. The 3D reference posture served as the ground-truth of the input image. Next, the joint angles of the ground-truth posture were changed by random values in the range $[-20°, +20°]$ to generate a new posture to serve as the initial posture for the posture candidate estimation algorithm. This approach was adopted to emulate the real application situation that the actual performer's posture may differ from the initial posture estimate. Note that some of the synthetic input images generated contained self-occlusion and depth ambiguity.

The posture candidate estimation algorithm was executed to generate posture candidates that best match the synthetic input images. Among the posture candidates, there is one best candidate that is most similar to the ground truth. For the algorithm to be accurate, the posture error between the best candidate and the ground truth should be small.

Figure 3 shows that the algorithmic error ranges from $2°$ to $15°$, with a mean of $7°$ and a standard deviation of $2.6°$. The larger errors occur in the input images with total occlusion of some body parts. For the other input images, the errors are mainly due to depth ambiguity of body parts. In our test, a body part of length 30cm parallel to the image plane measures about 36 pixels in the image, and the length of the body part in the image changes by only one pixel when the body part is rotated by $14°$ in depth. Therefore, a mean error of $7°$ is reasonable and acceptable for an algorithm that uses a single camera view. The accuracy can be further improved using images with larger resolution or sub-pixel algorithm, which will take more time.

In this test, the spatiotemporal registration algorithm was executed on the Taichi sequence and the golf sequence.

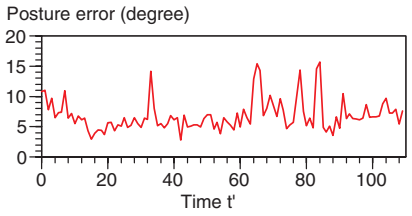
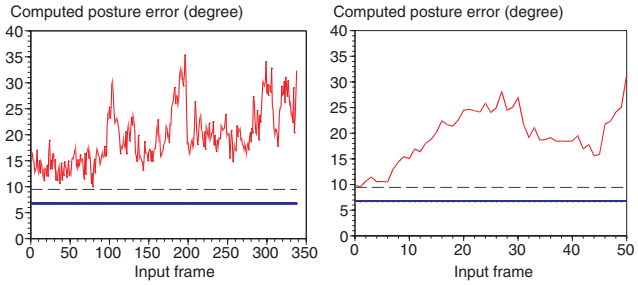


Figure 3. Algorithmic error in estimating performer's posture.



(a) (b)

Figure 4. Computed posture error for (a) Taichi motion and (b) golf swing. Dashed lines indicate the expected algorithmic error.

Then, the posture error between the selected best posture candidate and the corresponding reference posture was computed for each input image in the motion sequences.

Figure 4 illustrates the computed errors for the two sequences. As discussed in the previous section, the algorithm has a mean error of $7°$ (solid line in Fig. 4) in estimating postures in synthetic data. For real images, this algorithmic error is expected to be larger, say the mean error plus the standard deviation (dashed line in Fig. 4). The computed error includes both algorithmic error and performer's actual posture error. Since the algorithmic error is small compared to the computed error, there is high confidence that the computed error indeed reflects the performer's error.

Figure 4(a) shows that the computed posture errors are relatively small in most of the first 100 frames compared to the later frames. This is reasonable because the performer started from a standard standing posture which was easy to perform correctly. As the performer moved on to the more difficult postures, more error were made.

Figure 5 shows sample results of the Taichi sequence with small posture errors. The selected posture candidates are similar to the corresponding reference postures. The depth orientations of the body parts in the selected posture candidates are the same as those in the performer's postures in the input images. These results qualitatively verifies that the algorithm can select the best posture candidates.

Figure 6 shows sample results of the Taichi sequence with larger posture errors. Comparing the best posture candidates selected by the algorithm (blue) and the corresponding reference postures (green), there are large errors in the

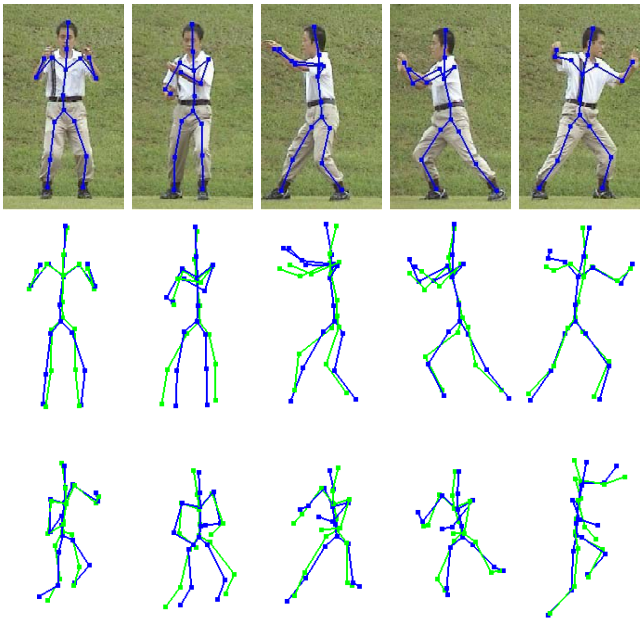


Figure 5. Sample postures in Taichi sequence with small errors. First row: input images with the selected posture candidates overlaid. Second and third rows: selected posture candidates (blue skeleton) overlapped with the corresponding reference postures (green skeleton) in the frontal and oblique views.

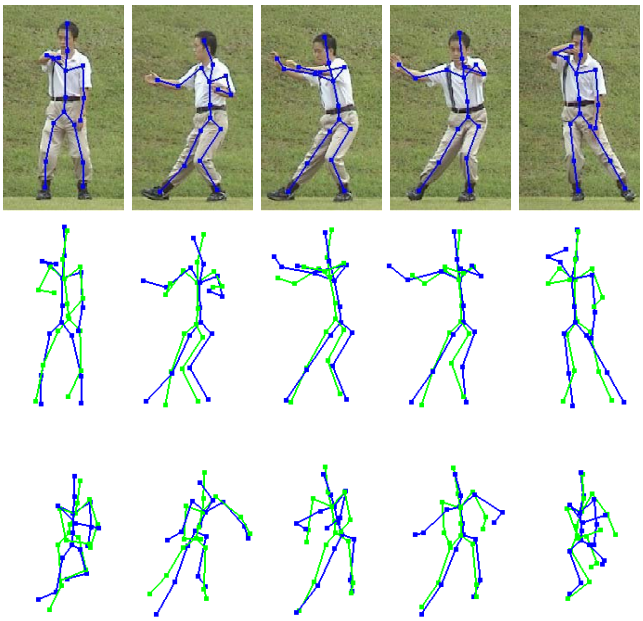


Figure 6. Sample postures in Taichi sequence with larger errors.

poses of the performer's arms. It shows that the algorithm can indeed identify errors in the performer's postures.

Figure 7 illustrates sample results of the golf swing sequence. Similar to the Taichi case, the performer made less error at the beginning of the swing and larger error later on

Figure 7. Sample postures in golf swing sequence.

in the swing, which is verified in Figure 7. The depth orientations of body parts in the selected posture candidates are the same as those in the performer's posture in the input images. These results verify that the algorithm be applied to the analysis of different types of sports motion.

Figure 8 shows sample test results of the Taichi sequence under ambiguous conditions. Depth ambiguity exists in all the images and self-occlusion of the right arm exists in the last three images. Nevertheless, the algorithm can still infer the pose of the occluded body part when the performer's posture does not differ greatly from the reference posture. That is, the algorithm is robust against depth ambiguity and self-occlusion. Of course, when the pose of the totally occluded body part differs significantly from that in the reference posture, no information will exist in a single camera view for the algorithm to infer the actual pose.

9. Conclusions

This paper proposes a novel and fundamental problem for sports motion analysis: 3D-2D spatiotemporal motion registration. Since it is infeasible to directly solve such a complex problem, this paper presents a framework that decomposes the problem into four subproblems, which are solved in stages. By using reference postures as initial postures to estimate possible posture candidates in the input images, the algorithm avoids the accumulation of estimation error over time. Moreover, the algorithm seeks to compute the smallest amount of correction required by the per-

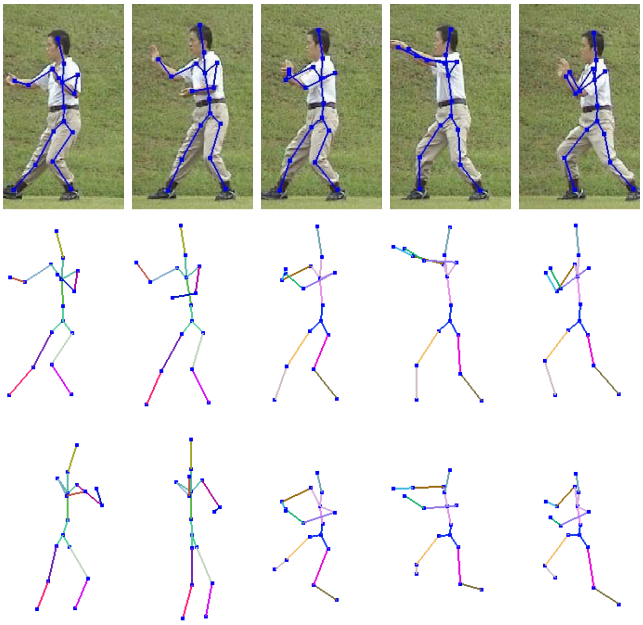


Figure 8. Selected best postures under ambiguous conditions. (Row 3) The first two images show the side views of selected posture candidates, and the last three show the frontal oblique views. Each bone is marked with a unique color for easy identification.

former to match the reference motion. Comprehensive tests were performed to evaluate the performance of the algorithms. Test results show that the computed errors are significantly larger than the expected algorithmic errors when performer's errors occur. This indicates that there is high confidence that the computed errors indeed reflect the performer's errors. In addition, the algorithm can handle depth ambiguity and partial self-occlusion of body parts. In the case of total self-occlusion, the algorithm can infer the pose of the occluded body part if the performer's posture does not differ greatly from the reference posture. The algorithm can also be applied to analyze different types of sports motion.

References

[1] A. Agarwal and B. Triggs. 3D human pose from silhouettes by relevance vector regression. In *Proc. CVPR*, pages 882–888, 2004. 2

[2] V. Athitsos, J. Alon, S. Sclaroff, and G. Kollios. Boostmap: A method for efficient approximate similarity rankings. In *Proc. CVPR*, pages 268–275, 2004. 2

[3] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. CVPR*, pages 8–15, 1998. 1

[4] Y. Capsi and M. Irani. Spatio-temporal alignment of sequences. *IEEE Trans. PAMI*, 24(11):1409–1424, 2002. 1

[5] T. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. In *Proc. CVPR*, pages 239–245, 1999. 2

[6] A. Elgammal and C. Lee. Inferring 3D body pose from silhouettes using activity manifold learning. In *Proc. CVPR*, pages 681–688, 2004. 2

[7] M. Gleicher. Retargeting motion to new characters. In *ACM SIGGRAPH*, pages 33–42, 1998. 2

[8] G. Hjaltason and H. Samet. Properties of embedding methods for similarity searching in metric spaces. *IEEE Trans. PAMI*, 25(5):530–549, 2003. 2

[9] G. Hua and Y. Wu. Multi-scale visual tracking by sequential belief propagation. In *Proc. CVPR*, pages 826–833, 2004. 4

[10] M. Isard. Pampas: Real-valued graphical models for computer vision. In *Proc. CVPR*, pages 613–620, 2003. 4

[11] M. Jones and J. Rehg. Statistical color models with application to skin detection. *IJCV*, 46(1):81–96, 2002. 2

[12] M. W. Lee and I. Cohen. Proposal maps driven mcmc for estimating human body pose in static images. In *Proc. CVPR*, pages 334–341, 2004. 2

[13] R. Li, M. Yang, S. Sclaroff, and T. Tian. Monocular tracking of 3d human motion with a coordinated mixture of factor analyzers. In *Proc. ECCV*, pages 137–150, 2006. 1, 2

[14] MotionCoach: Golf swing analysis. www.motioncoach.com. 1

[15] D. Ramanan, D. A. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. CVPR*, pages 271–278, 2005. 2, 4

[16] C. Rao, A. Gritai, M. Shah, and T. Mahmood. View-invariant alignment and matching of video sequences. In *Proc. ICCV*, pages 939–945, 2003. 1

[17] C. Rother, V. Kolmogorov, and A. Blake. Grabcut - interactive foreground extraction using iterated graph cuts. In *Proc. ACM SIGGRAPH*, pages 309–314, 2004. 2

[18] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter-sensitive hashing. In *Proc. ICCV*, pages 750–757, 2003. 2

[19] H. Sidenbladh, M. Black, and D. Fleet. Stochastic tracking of 3D human figures using 2D image motion. In *Proc. ECCV*, pages 702–718, 2000. 1

[20] Simi: 3D motion tracking system. www.simi.com. 1

[21] Simi: Video based motion analysis. www.simi.com. 1

[22] C. Sminchisescu and B. Triggs. Covariance scaled sampling for monocular 3D body tracking. In *Proc. CVPR*, pages 447–454, 2001. 1, 2

[23] C. Sminchisescu and B. Triggs. Kinematic jump processes for monocular 3D human tracking. In *Proc. CVPR*, pages 69–76, 2003. 1

[24] Sports Motion: 2D video-based motion analysis system. www.sports-motion.com. 1

[25] E. Sudderth, A. Ihler, W. Freeman, and A. Willsky. Nonparametric belief propagation. In *Proc. CVPR*, pages 605–612, 2003. 4

[26] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Visual hand tracking using nonparametric belief propagation. In *IEEE CVPR Workshop on Generative Model based Vision*, 2004. 2, 4

[27] R. Urtasun, D. J. Fleet, and P. Fua. 3D people tracking with gaussian process dynamical models. In *Proc. CVPR*, pages 238–245, 2006. 1, 2

[28] V1 Pro: Golf swing analysis software. www.ifrontiers.com. 1

[29] Vicon: Optical motion capture system. www.vicon.com. 1