TRACKING HUMAN MOVEMENT PATTERNS USING PARTICLE FILTERING

Richard D. Green
Dept of Electrical and Information Engineering
The University of Sydney, NSW 2006
Australia

Ling Guan
Dept of Electrical and Computer Engineering
Ryerson University, Canada M5B 2K3
e-mail: lguan@ee.ryerson.ca

ABSTRACT

At least 32 joint related degrees of freedom need to be estimated to reliably track the human body in 3D. The particle filter is robust to distracting clutter by maintaining multiple hypotheses for each of these joint angles. Real-time tracking is difficult however with the computational overhead of such a large search space. This paper optimizes this search space utilizing feedback from a Continuous Human Movement Recognition (CHMR) system and improves the robustness and efficiency of each particle calculation using a novel body model. The joint angles are estimated for the next frame using a Particle filter with forward smoothing. A new paradigm enables the temporal segmentation of continuous motion into *dynemes*. Using HMM, the CHMR system attempts to infer the human movement skill that could have produced the observed sequence of dynemes. Hundreds of movement skills, from gait to saltos, are successfully tracked and recognized.

1. INTRODUCTION

Research into tracking, recognizing and understanding full body human motion has so far been mainly limited to gait or frontal posing. This paper describes a framework for tracking, recognizing and quantifying full body human motion, free of joint markers, set-up procedures and hand-initialization, over a larger range of motion than previously attempted by considering hundreds of different movement skills.

Robust tracking of the full human body in 3D is enhanced by predicting the joint angles for the next frame to stabilize the tracking. This calculation of joint angles, for the next frame, was cast as an estimation problem, which was solved using a Particle filter.

The Particle Filter was developed to address the problem of tracking contour outlines through heavy image clutter [4, 5]. The filter's output at a given time-step, rather than being a single estimate of position and covariance as in a Kalman filter, is an approximation of an entire probability distribution of likely joint angles. This allows the filter to maintain multiple hypotheses and thus be robust to distracting clutter.

With about 32 degree of freedom (DOFs) to be determined for each frame, there is the potential of exponential complexity evaluating such a high dimensional search space. MacCormick [7] proposed Partitioned Sampling and Sullivan [11] proposed Layered Sampling to reduce the search space by partitioning it for more efficient particle filtering. Although Annealed Particle Filtering [2] is an even more general and robust solution, it struggles with efficiency which Deutscher [3] improves with Partitioned Annealed Particle Filtering. This paper optimizes the huge search space related to calculating many particles for over 32 DOFs by utilizing feedback from the CHMR system. A novel body model is also engaged to improve the robustness and efficiency of each calculation for the remaining particles.

Recognizing and quantifying human movement requires spatial segmentation followed by temporal segmentation (Fig. 1). The spatial segmentation is essentially a tracking process which determines a *motion vector* encapsulating a set of joint angles (and other biomechanical parameters) for each frame. The temporal segmentation is a CHMR system which attempts to infer the movement *skill* that could have produced the observed sequence of motion vectors.

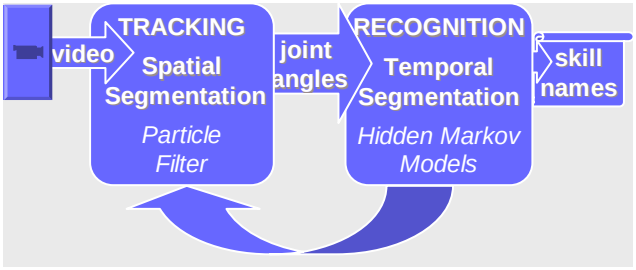

Fig. 1. Overview of segmentation of human motion.

Where the tracking process utilizes a body model and a kinematic model, the CHMR system draws on a *dyneme*-model, skill-model, and a semantic-model (Fig. 5). Where the tracking process stochastically enhances spatial segmentation with a particle filter, the CHMR system stochastically enhances temporal segmentation with a HMM. The tracking is further stabilized and optimized by feeding back information from the CHMR system (Fig. 1).

2. TRACKING

Various approaches for tracking the whole body have been proposed in the image processing literature. They can be distinguished by the representation of the body as a stick figure, 2D contour or volumetric model and by their dimensionality being 2D or 3D. Joint angles are able to be more directly estimated by mapping human body models directly onto a given image. Volumetric 3D models have the advantage of being more generally valid with self occlusions more easily resolved. Most volumetric approaches model body parts using generalized cylinders [8] or super-quadratics [9]. Some extract features [12] and others fit the projected model directly to the image [8].

2.1 Body Model

Cylindrical and quadratic body models, used in previous studies, do not contour accurately to the body, thus decreasing tracking stability. No study has yet utilized a color 3D texture

map of the entire body, which would enable body parts to be tracked more accurately to further stabilize tracking. In a novel approach to body part representation proposed in this paper, 3D regions are sized and texture mapped from each body part by extracting features during the initialization phase.

Anthropometric data [10] is used as a Gaussian prior for the initial body-part proportions with left-right symmetry of the body used as a stabilizing guide. Initially a low accuracy is set for each body-part with the accuracy increasing as structure from motion resolves the relative proportions. For example, a low color and high radius accuracy is initially set for pixels near the edge of a body part, high color and low radius accuracy for other near side pixels, and a low color and low radius accuracy is set for far side pixels. The ongoing temporal resolution following self occlusions enables increasing radius and color accuracy. Breathing, muscle flexion and other normal variations of body part radius are accounted for by the radius elasticity parameter.

2.2 Kinematic Model

The kinematic model tracking the position and orientation of a person relative to the camera, entails projecting 3D body model parts onto a 2D image with three chained homogeneous transformation matrices:

$$p(x, b) = I_i(x, C_i(x, B_i(x, b))) \quad (1)$$

where x is a parameter vector calculated for optimum alignment of the projected model with the image, B is the Body frame of reference transformation, C is the Camera frame of reference transformation, I is the Image frame of reference transformation, b is a body-part surface point, p is a pixel in 2D frame of video.

Joint angles are used to track the location and orientation of each body part, with the range of joint angles being constrained by limiting the degrees of freedom (DOF) associated with each joint. A simple motion model of constant angular velocity for joint angles is used in the kinematical model. Each DOF is constrained by anatomical joint-angle limits, body-part inter-penetration avoidance and joint-angle equilibrium positions modeled with Gaussian stabilizers around their equilibria. To stabilize tracking, the joint angles are estimated for the next frame. The calculation of joint angles, for the next frame, is cast as an estimation problem which is solved using a Particle filter (Condensation algorithm).

2.3 Particle Filter

The Particle Filter is a considerably simpler algorithm than the Kalman Filter. Moreover despite its use of random sampling, which is often thought to be computationally inefficient, the Particle Filter can run in real-time. This is because tracking over time maintains relatively tight distributions for shape at successive time steps and particularly so given the availability of accurate learned models of shape and motion from the human-movement-recognition (CHMR) system.

The particle filter has

- three probability distributions in problem specification:
  1. Prior density p(x) for the state x ⇨ joint angles in previous frame
  2. Process density $p(x_t|x_{t-1})$ ⇨ kinematic and body models
  3. Observation density p(z|x) ⇨ image in previous frame
- one probability distribution in the solution specification:
  1. State Density $p(x_t|Z_t)$ ⇨ joint angles in next frame

When tracking through background clutter or occlusion, a joint angle may have N alternate possible values (samples) ***s*** with respective weights ***w***, where

prior density $p(x) \approx S_{t-1} = \{(s^{(n)}, w^{(n)}), n=1..N\}$ = a sample set

For the next frame, a new sample is selected, $\acute{s}_t = s_{t-1}$ by finding the smallest i for which $c^{(i)} \geq r$, where $c^{(i)} = \sum_t w^{(i)}$ and ***r*** is a random number {0,1}.

A joint angle, $s_t^{(n)}$ in the next frame is predicted by sampling from the process density, $p(x_t|x_{t-1} = \acute{s}_t^{(n)})$ which encompasses the kinematic model, body model and cost function minimization. In this prediction step both edge and region information are used. The edge information is used to directly match the image gradients with the expected model edge gradients. The region information is also used to directly match the values of pixels in the image with those of the body model's 3D color texture map.

The prediction step involved minimizing the cost functions:

edge error E_e using edge information:

$$E_e(S_t) = \frac{1}{2n_e v_e} \sum_{x,y} (|\nabla i_t(x, y)| - m_t(x, y, S_t))^2$$

$$+0.5(S - S_t)^T C_t^{-1}(S - S_t) \rightarrow \min S_t \quad (2)$$

region error E_r using region information:

$$E_r(S_t) = \frac{1}{2n_r v_r} \sum_{j=1}^{n_r} (i_t[p_j(S_t)] - i_{t-1}[p_j(S_{t-1})])^2$$

$$+E_e(S_t) \rightarrow \min S_t \quad (3)$$

where i_t represents the image at time t, m_t the model gradients at time t, n_e is the number of edge values summed, v_e is the edge variance, n_r is the number of region values summed, v_r is the region variance, p_j is the image pixel coordinate of the jth surface point on a body part.

Performance is enhanced by minimizing the area of body part being tracked, based on angular speed and occlusion.

The new position in terms of the observation density, $p(z_t|x_t)$ is then measured and weighed with forward smoothing:

- Estimate weights $wt = p(z_t|x_t = s_t)$
- Normalize weights $\sum_n w^{(n)} = 1$
- Smooth weights wt over 1..t, for n trajectories

- Replace each sample set with its n trajectories $\{(s_t, w_t)\}$ for 1..t
- Re-weight all $w^{(n)}$ over 1..t

Trajectories tend to merge within 10 frames

⇨ $O(N_t)$ storage prunes down to O(N)

In this paper, feedback from the CHMR system utilizes the large training set of skills to achieve an even larger reduction of the search space. In practice, human movement is found to be most efficient, with minimal DOFs rotating at any one time. The equilibrium positions and physical limits of each DOF further stabilize and minimize the dimensional space. With so few DOFs to track at any one time, a minimal number of particles are required, significantly raising the efficiency of the tracking process. Such highly constrained movement results in a sparse domain of motion projected by each motion vector.

3. DYNEMES

A new paradigm has been developed for the temporal segmentation of continuous motion into dynemes. As the phoneme is a phonetic unit of human speech, so is the dyneme a dynamic unit of human motion. An alphabet of dynemes has been determined by deconstructing hundreds of movement skills into their correlated lowest common denominator of basic movement.

For example, a Centre of Mass (COM) category of dyneme is illustrated in Fig. 3a where each running step is delimited by a COM minima. A full 360° rotation of the principle axis during a cartwheel in Fig. 3b illustrates another dyneme category of rotation from the vertical.

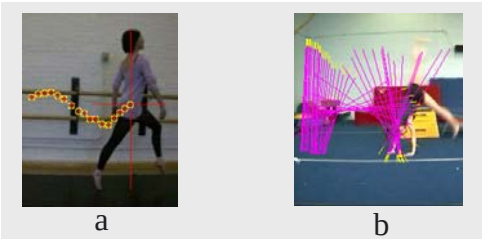

Fig. 3. A sequence of COM parameters during running and a sequence of principle-axis parameters thru a cartwheel.

The pronunciation of the English language is constructed from approximately 50 phonemes. This work has so far determined about 35 principle dynemes with the expectation of more dynemes being realized in future research.

4. SKILL RECOGNITION

To simplify the design, it is assumed that the CHMR system contains a limited set of possible human movement *skills*. This approach restricts the search for possible skill sequences to those skills listed in the *skill model*, which lists the candidate skills and provides *dynemes* – a set of basic units, individual granules of human movement – for the composition of each skill. The current skill model contains hundreds of skills where the length of the skill sequence being performed by a person is unknown. If M represents the number of human movement skills in the skill model, the CHMR system could hypothesize M^N possible skill sequences for a skill sequence of length N. However these skill sequences are not equally likely to occur due to the biomechanical constraints of human motion.

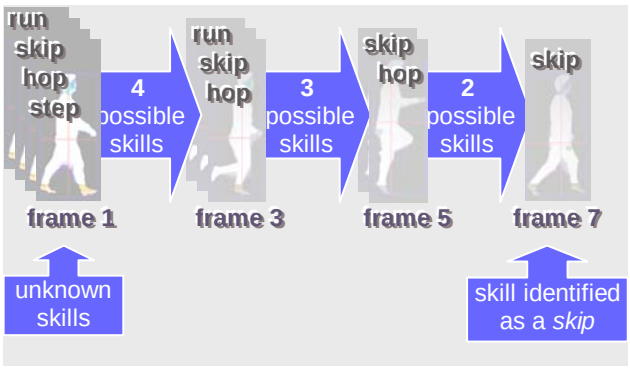

Fig. 4. Stochastic Prediction and Recognition of a Movement Skill given the Motion Vector for each Frame.

A generative probabilistic model that encapsulates this sequence of steps is used. Given an observed sequence of motion vectors y_1^T the recognition process attempts to find the skill sequence $\hat{s}_1^N$ that maximizes this skill sequence's probability:

$$\hat{s}_1^N = \arg\max_{s_1^N} p(s_1^N \mid y_1^T) \equiv \arg\max_{s_1^N} p(y_1^T \mid s_1^N)\, p(s_1^N) \quad (4)$$

This approach applies Bayes' law and ignores the denominator term to maximize the product of two terms: the probability of the motion vectors given the skill sequence and the probability of the skill sequence itself. The CHMR framework described by this equation is illustrated below in Figure 5 where, using motion vectors from the tracking process, the recognition process uses the dyneme, skill, semantic and activity models to construct a hypothesis for interpreting a video sequence.

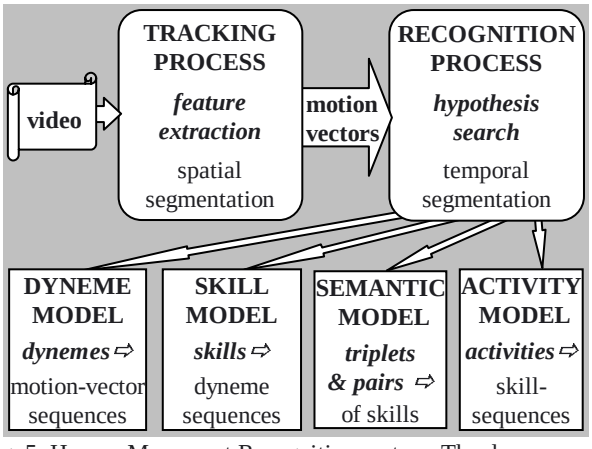

Fig. 5. Human Movement Recognition system. The dyneme, skill and semantic and activity models construct a hypothesis for interpreting a video sequence.

In the tracking process, motion vectors are extracted from the video stream. In the recognition process, the search hypothesizes a probable movement skill sequence using four models:

- the *dyneme model* models the relationship between the motion vectors and the dynemes.
- the *skill model* block defines the possible movement skills that the search can hypothesize, representing each movement skill as a linear sequence of dynemes;
- the *semantic model* models the semantic structure of movement by modeling the probability of sequences of skills simplified to triplets and pairs; and
- The *activity model* defines the possible human movement activities that the search can hypothesize, representing each activity as a linear sequence of skills.

5. PERFORMANCE

Hundreds of skills were tracked and classified using a 1.8GHz, 640MB RAM Pentium IV platform processing 24 bit color within the Microsoft DirectX 8.1 environment under Windows XP. The video sequences were captured with a JVC DVL-9800 digital video camera at 30 fps, 720 by 480 pixel resolution.

Each person moved in front of a stationary camera with a static background and static lighting conditions. Only one person was in frame at any one time. Tracking began when the whole body was visible which enables initialization of the body model. However, an elongated trunk with disproportionate short legs is the body-model consequence of the presence of a skirt – the body model failed to initialize for tracking due to the variance of body-part proportions exceeding an acceptable threshold.

Particle filter tracking also failed for loose clothing. Even with smoothing, joint angles surrounded by baggy clothes permutated thru unexpected angles within an envelope sufficiently large as to invalidate the tracking.

Motion blurring lasted about 10 frames on average with the effect of perturbing joint angles within the blur envelope. Forward smoothing of the particle filter did not produce an acceptable result throughout the blurring sequence. Given a reasonably accurate angular velocity, it was possible to de-blur the image sufficiently to alleviate this problem (Fig. 6).

A B C

Fig. 6. A: particle filter tracking through motion blur of right calf and foot segments during a flick-flack (back-handspring).
B: 3 alternative particles (knee angles) for the right calf location.
C: Motion-blur corrected particle filter tracked location.

The *skill error rate* quantifies CHMR system performance by expressing, as a percentage, the ratio of the number of skill errors to the number of skills in the reference training set. Depending on the task, CHMR system skill error rates can vary by an order or magnitude. The CHMR system was tested on a training set of 840 movement patterns, from walking to twisting saltos. An independent testing set of 200 skills were evaluated. Both the training and testing skill sets were performed by the same people. These were successfully tracked, recognized and evaluated with their respective biomechanical components quantified where a skill error rate of 4.5% was achieved.

6. CONCLUSIONS AND FUTURE RESEARCH

In this paper, it was demonstrated that this approach was able to successfully track and classify diverse motion patterns in real-time, free of joint markers, set-up procedures and hand-initialization, to detect human movement skills with a skill error rate of 4.5%. Hundreds of movement skills, from walking to twisting saltos, were successfully tracked, recognized and evaluated with their respective biomechanical components quantified. The results suggest that this approach has the potential to guide clinicians and coaches toward analyzing movement and quantifying improvement utilizing non-invasive biomechanical analysis.

Future studies aim to extend the dyneme, skill, semantic and activity models and also to improve the robustness and accuracy of the system, especially the poorly observable depth DOFs, by applying to the Particle filter, inflated posteriors and dynamics for sample generation and then reweighing the results.

7. REFERENCES

[1] N. I. Badler, C. B. Phillips, B. L. Webber, “Simulating humans”, Oxford University Press, New York, NY, 1993.

[2] J. Deutscher, A.. Blake and I. Reid. “Articulated body motion capture by annealed particle filtering”, Proc. Conf. Computer Vision and Pattern Recognition, vol. 2, 1144-1149, 2000.

[3] J. Deutscher, A. Davison, and I. Reid. “Automatic partitioning of high dimensional search spaces associated with articulated body motion capture”, Computer Vision and Pattern Recognition, volume 2, pages 669-676, 2001.

[4] M. A. Isard and A. Blake. “Visual tracking by stochastic propagation of conditional density” Proc. 4th European Conf. Computer Vision, 343-356, Cambridge, England, Apr 1996.

[5] M. A. Isard and A. Blake. “A mixed-state Condensation tracker with automatic model switching”, Proc. 6th Int. Conf. on Computer Vision, 107-112, 1998.

[6] F. Jelinek, Statistical Methods for Speech Recognition, MIT Press, Cambridge, Mass., 1999.

[7] J. MacCormick and M. Isard, “Partitioned sampling, articulated objects and interface-quality hand tracking”, Proc European Conf. Computer Vision, vol. 2, 3-19, 2000.

[8] J. M. Rehg, T. Kanade, “Model-based tracking of self-occluding articulated objects”, Fifth Int. Conf. on Computer Vision: 612-617, 1995.

[9] A. Pentland, B. Horowitz, “Recovery of nonrigid motion and structure”, IEEE Trans. on PAMI, 13: 730-742, 1991.

[10] S. Pheasant, Bodyspace. Anthropometry, Ergonomics and the Design of Work, Taylor & Francis, 1996.

[11] J. Sullivan, A. Blake, M. Isard, and J. MacCormick. “Object localization by bayessian correlation”, Proc. 7th Int. Conf. on Computer Vision, vol. 2, 1068-1075, 1999.

[12] C. Wren, A. Azarbayejani, T. Darrell, A. Pentland, “Pfinder: Real-time tracking of the human body”, IEEE Trans. on PAMI, 19(7): 780-785, 1997.