Evaluating Video-Based Motion Capture

Michael Gleicher
Department of Computer Sciences

Nicola Ferrier
Department of Mechanical Engineering

University of Wisconsin-Madison, Madison, WI[*]

gleicher@cs.wisc.edu
ferrier@engr.wisc.edu

Abstract

Motion capture can be an effective method of creating realistic human motion for animation. Unfortunately, the quality demands for animation place challenging demands on a capture system. To date, capture solutions that meet these demands have required specialized hardware that is invasive and expensive. Computer vision could make animation data much easier to obtain. Unfortunately, current techniques fall short of the demands of animation applications. In this paper, we will explore why the demands of animation lead to a particularly difficult challenge for capture techniques. We present a constraint-based methodology for reconstructing the 3D motion given image observations, and use this as a tool for understanding the problem. Synthetic experiments confirm that these situations would arise in practice. The experiments show how even simple visual tracking information can be used to create human motion but even with perfect tracking, incorrect reconstructions are not only possible but inevitable.

1 Introduction

Motion capture is an attractive method for creating the movement for computer animation. It can provide motion that is realistic, and that contains the nuance and specific details of particular performers. It permits an actor and director to work together to create a specific desired performance, that may be difficult to describe with enough specificity to have an animator re-create manually.

Motion capture also has its share of weaknesses. In order to represent its detailed, nuanced results, motion capture data contains large quantities of unstructured data that is cumbersome to manipulate. The specificity of the data makes it difficult to alter, especially since the key “essence” of the motion is not distinguished from the large amount of potentially irrelevant details. The development of new and improved methods of editing and processing motion capture data has made great strides in making motion capture a more viable tool for animation production [12, 23].

Another weakness of motion capture has been the pragmatic challenge of acquiring data. While research has made progress on using the data, capture techniques have evolved slowly. Special purpose tracking technologies, based either on mechanical or magnetic sensors, or specially designed cameras viewing carefully lit markers, are required to create the observations that are processed into motion data. While these systems have improved in their reliability, precision, and range, they are still generally expensive and intrusive. This relegates motion capture to be performed by dedicated studios providing specific environments for the production.

Ideally, the capture of motion data should be easily available, inexpensive, and non-invasive. Any performer could be captured in any setting that is desired. Using standard video cameras is an enticing prospect as it could meet these goals. The use of a single camera is a particularly enticing. It offers the lowest cost, simplified setup, and the potential use of legacy sources such as films.

The creation of motion capture data from a single video stream seems like a plausible idea. People are able to watch a video and understand the motion, why can’t a computer do it? But clearly, the recreation of human motion from a video stream is a challenging task for computer vision. The computer vision community has been actively exploring the problem for several years, and has made great advances. However, the current results are disappointing for animation applications as will be discussed in section 6.1.

In this paper, we examine the disconnect between the computer vision research in human tracking and the needs for motion capture for animation applications. We will consider why animation is a particularly challenging application for video motion capture, and that it is different than the traditional vision applications. We will examine the mathematics of the single camera capture problem to understand why it is a fundamentally hard problem, seeming to require techniques that are at odds with the demands of the animation application. We will confirm these observations with synthetic experiments, and a review of the current state of the art.

This paper does **not** provide a solution to the problem of capturing animation data from video sources. Our goal in this paper is to reconcile the demands of computer animation with the progress that has been made in computer vision methods. The methods that we will introduce are a tool for helping us understand these challenges, not a solution to the capture problem. By examining the challenges that one faces in using perfect, synthetic data, we hope to better understand what must be achieved for a viable solution.

2 The Motion Capture Problem

The goal of motion capture is to record the movement of a performer (typically, but not always, human) in a compact and usable manner. Were those last attributes not required, simply creating a video would suffice. For this paper, we are concerned with the gross motion of the body. The specific capture of facial motion poses a different set of challenges.

Computer graphics and computer vision usually abstract the body into a small number of rigid segments that rotate relative to one another. This approximation is crude. Human knees, elbows, and ankles do not have a single pivot point. The true motion of more complex joints, such as shoulders, hips, or the neck, are even further from their kinematic approximations.

While the skeletal approximation is crude, it is required for tractability. Because some information is necessarily thrown away, animation techniques must take care to preserve the “essence” of the motion in doing any processing. What makes this particularly challenging is that the important properties of a motion are difficult to identify. Somewhere in the myriad of details properties such as mood, expression, and personality lie.

[*] This work was supported in part by NSF grants CCR-9984506, IIS-0097456, and IRI-9703352, Micrososft Research, equipment donations from IBM, NVidia and Intel, software donations from Discreet, Alias/Wavefront and Pixar, and data donations from House of Moves, Spectrum Studios, and Digital Domain.

The motion capture problem we consider, therefore, must have the following form: given a single stream of video observations of a performer, compute a 3D skeletal representation of the motion of sufficient quality to be useful for animation. This ill-defined last clause is the part unique to animation.

2.1 Challenges of Motion Capture for Animation

Animation may seem to be an easy application of motion capture as precision is unimportant. Animation rarely cares about exact positions, as applications such as medical analysis might. Animation is more concerned with seemingly less precise things such as emotion, style or intent. Unfortunately, these are difficult to quantify. Precisely capturing all details of the movement is a conservative way to insure that the critical, but intangible properties are retained.

Animation has direct needs for precision that stem from the sensitivities that viewers have in experiencing motion. For example, a viewer is likely to notice imprecision in a character's interaction with its world. A foot floating slightly above the floor, or sliding a small amount, or a hand not quite reaching the doorknob, are tiny imprecisions yet can completely destroy the illusion of realism.

Another place where perception makes small changes noticeable are high-frequencies. Jitters are extremely easy to notice, possibly because of the eye's sensitivity to high frequency. Such artifacts are often created by systematic noise in a capture system. Pops, where a body part moves impossibly far in a single frame, are similarly noticeable, and are created by poorly designed processing algorithms that do not guarantee temporal coherence. "Wobbles," or misplaced mid-frequencies, can also be obvious. For example, when processing errors are distributed amongst several frames (rather than introducing a pop), an uncharacteristic motion can be observed in a character's knees, causing them to move in a disconcerting way that stands out against common movements like walking.

In all of these cases, the measurable magnitude of the errors are small, but their visible effects are significant. The same quantity of error, occurring at a more fortunate time, or in a more fortunate manner, would not be a problem.

Since high-frequency noise is often a problem, a common approach to dealing with it is to low-pass filter the data. Unfortunately, this rarely has the desired effect: the same reasons that make unwanted high frequencies so obtrusive also make the removal of important high-frequencies a problem. Significant high-frequencies are often the result of important events in a motion, such as a contact, impact, or purposeful gesture. Removing the high-frequencies from a karate kick or a soldier's salute destroys the motion, just as adding high-frequencies to make a motion jittery would. The overuse of low-pass filtering is a common problem in motion capture processing [23], and leads to a damped look that lacks the crispness that makes for attractive motion. Over-filtering also causes other problems, such as destroying the relationships between the character and its environment (the all-too-well-known foot-slide problem).

To make matters worse, the nuance that is so hard to capture and preserve is critical to most animation. We rarely want "a" motion, but rather want "the" performance. The specifics of the performer and performance are important. Just as actors are trained to know that there is no such thing as a "generic" movement, the sense of character, context, and intent should be conveyed in movements. Just as with stage and film actors, motion capture directors work with performers to achieve what they want to see. Good motion capture must preserve this.

In fact, it is often these specific performances that are the reasons for wanting to perform motion capture in the first place. If we wanted a standard movement, we might use a recorded one from a library. More often, motion capture is required for recording a specific person, a specific situation, or something otherwise non-standard: we are trying to record something unusual in some way, surprising, athletic, or artistic, because predictable and standard movements do not need to be incessantly re-created.

2.2 Motion Capture in Practice

In order to meet the challenges of animation, motion capture practitioners use a combination of precise capture equipment, planning, and post-processing.

Motion capture for animation is almost always done with a precision in excess of what the application demands, in both space and time. Temporal sampling rates of 120Hz are common for animation production, even when the final product will only be created at 30 frames per second. By over-sampling, statistical and signal processing methods can be used to reduce noise. Requiring this excess precision means that the capture equipment is significantly more expensive and complex. Often this equipment requires specific environments to operate in, or at least has a cost that forces it to be used only at specific studios.

As with any media production, planning is an essential part of motion capture. Planning responds to the challenges of capture by insuring that the desired performance is created for the sensors to record, and to stage it in a way that makes it most likely that it will be recorded properly.

Once the performance is captured, the recorded data is processed into a useful form. Overall, the procedure must transform the data in a way that maintains the temporal continuity and spatial precision of the data, ideally suppressing noise and avoiding the addition of artifacts. Well designed processing pipelines use heuristics, such as end-effector tracking, to preserve the most important aspects of a motion, see Shin et. al. [20] for a discussion. Constraint-based [7, 8] and importance-based [20] methods attempt to vary the choice of what is preserved by identifying or predicting what is most important. Unfortunately, manual tweaking is often a necessary step for most capture processing operations.

2.3 Human Motion Tracking

In addition to computer animation, medical, surveillance, and recognition applications would also benefit from an inexpensive capture solution. The development of a video-based capture solution has, therefore, been an important topic in the computer vision community. Recent surveys of the computer vision literature on human motion capture are available [6, 13] and we give a brief one in section 6.1.

The capture problem is inherently difficult: the articulated model does not accurately reflect the real performer, articulations lead to self-occlusions, even the articulated models contain many degrees of freedom, the skeleton is internal and therefore cannot be observed directly. Our information sources are inherently 2 dimensional and occlusion is possible. In addition, the medium provides a finite resolution (spatially and temporally), and the parameters of real cameras are difficult to obtain precisely. Given these limitations, it is not surprising that the practical approach to motion capture involves using sensing modalities without these limitations.

Clearly, the amount of detail that can be recovered from a restricted set of observations is limited. For example, if we observe a point on the performer in an image created by a camera, we cannot determine the position of the point, only constrain its location to lie along a ray. For our discussion, we assume an idealized pinhole camera model such that the ray is defined by the camera's focal point and the point on the image plane [1]. In practice, a camera has a finite resolution so observations are only localized to a region of the image plane. This leads to a weaker constraint that places the point within a cone, rather than on a ray. Uncertainty in camera parameters further enlarges the set of feasible positions.

Additional information is needed to determine the position of a point in space. A variety of sources have been utilized in various computer vision techniques, and a few can be applied to motion reconstruction. Most methods assume strong models to place further restrictions on possible poses. Such assumptions are problematic for motion capture applications: because we are interested in capturing novel motions, we cannot count on previous motions for cues. For example, Shadow Puppets [1] would create 3D reconstructions that contain segments of (and therefore nuances of) the training motions, not necessarily the observed motion.

2.4 Animation vs. Vision

If the general challenges of human motion capture were not hard enough for computer vision, the specific challenges of animation make the problem even tougher.

- Unlike applications such as recognition and surveillance, animation does care about small details.
- Jitter and wobbles often come from uncertainty in computations, or the failure to account for interframe coherence properly. These small errors can be extremely noticeable.
- The importance of high frequencies means that filtering is not a viable tool for noise removal at video sampling rates (for example, using Kalman filters). It also is problematic for methods that use regularization or damping to achieve coherence, or that rely on a highly damped dynamic model.
- The unpredictability and unusual motions that we need to capture limit the strength of the models we can apply.

Clearly, achieving video-based motion capture for animation applications is challenging. The key to success seems to be augmenting the information that can be obtained from the single stream of video observations.

3 Constraint-Based Approach

The video motion capture problem demands that we compute a detailed 3D model of the performance based on a limited set of 2D observations. The observations are weak: they provide limited cues as to the pose of the performer. A set of observations limits the range of possible poses. Additional information is required to determine the specific pose.

Our approach to analyzing (and potentially implementing) video motion capture is based on finding restrictions on the pose caused by each piece of available information. Each piece of information serves as a constraint, limiting the potential space of possible poses[2]. We can assess how much each new type of information might be able to help sufficiently narrow the space of possible poses. For a given set of constraints, we can understand how insufficient they are for reconstructing motion.

Different types of information yield different types of constraints, each yielding equational relationships that must hold on the determined state. For video motion capture, we must consider:

Character Constraints provide limitations on possible poses based on the performer that we are tracking. Examples include the rigidity of distances between joints (when we assume a skeletal model), non self-intersection, and limits in the range of joint angles.

Observation constraints that limit poses to be consistent with what is observed.

World constraints that place geometric restrictions based on knowledge of the world, for example that the performer's feet cannot move through the floor. These are difficult to use with video motion capture because we typically use camera-relative coordinate systems in which world objects, such as the floor, are not in known positions.

Pose constraints that specify the configuration of the character at particular instances. Such true 3D information might be obtained for some small set of frames by some process that is too difficult to apply to a large number of frames, E.g. [3], [11] and [14] all require the user to specify the initial pose.

Dynamics constraints that place limitations on the relationships between frames. Smoothness is **not** an effective constraint: high-frequencies are typically very important in motions as they are noticeable and must be captured effectively.

For our initial analysis and experiments, we consider a limited set of constraints: limitations on the distances between joint positions (enforcing skeletal rigidity), and the pose at an initial frame.

3.1 Uncertainty Models

In practice, the constraints on a pose is imprecise and incomplete. We cannot be certain what the pose is. Much recent work uses Bayesian assumptions, such as [1, 16, 21, 9] to determine what the most likely pose is given the observations and some prior world model. For animation, we wish to avoid about what is *most likely* in order that the approach work for novel motions and characters. Often, it is the *unlikely* motions that are most interesting to capture.

The differences between motions and characters are subtle, biasing a solution towards some prior model may lose the detail we are most interested in. Similarly assuming Gaussian distributions or using least-squares from range centers introduces biases towards the center, in effect "inventing" information. Even if a Gaussian model of noise is appropriate for certain sensing operations, the non-linearity of the imaging operation yields distributions that are distinctly non-Gaussian.

In practice, some sort of bias must be introduced into systems in order to make selection from within the feasible range practical. However, for analysis, we examine realistic sets of constraints to see how broad a set of solutions they permit, noting that a system cannot be certain that it is doing any better.

4 Analysis of Constraints

Given a set of constraints, we can determine the possible poses. Here, we consider an idealized situation to see the challenges even in this simple situation. We assume that the object we are tracking is a known rigid kinematic tree, that an initial pose is known, and that image observations of points (at the joints) are available.

In practice, the point data could be obtained from a region detector, a corner detector, marker tracking or manual annotation of video sequences. We assume that we can extract image points with some uncertainty. For our applications we assume the character points

[1] This model is more realistic than the orthographic model often used in the vision literature

[2] Plankers and Fua [17] take a similar approach

form a skeleton, and that we have distance constraints on the points on the character. Even though we have this skeletal model, we do not use it in the tracking process. There are several benefits to the separation of the tracking and reconstruction processes. One is that we do not utilize strong models where they may not be needed and singularities are removed from the tracking phase [14]. Thus we assume a separate tracking process has recovered the image locations of corresponding points.

We model the camera using an ideal pin hole projection which is parameterized by the focal distance, f, and the principal point (c_u, c_v). In the following we will use units of length and hence the pixel scaling parameters are omitted from our discussion. We express the constraints of the noisy image observations, (u, v), of 3D point $X = (x, y, z)$ as

$$\left|\left(\frac{fx}{z} + c_u\right) - u\right| \leq d, \quad \left|\left(\frac{fy}{z} + c_v\right) - v\right| \leq d, \qquad (1)$$

where $d = 0$ is the ideal image model. The initial pose, $X_k^0, k = 1, \ldots, M$ is known for M points on the skeleton. The distance constraints between pairs of points are given by $D(X_m, X_n) = l_{mn}$ where

$$D(X_m, X_n) = \sqrt{(x_m - x_n)^2 + (y_m - y_n)^2 + (z_m - z_n)^2}$$

is the Euclidean distance between two points and for clarity we have dropped the superscript denoted the time (or image frame).

Using these definitions, the **constraint-based 3D reconstruction problem** can be expressed as follows: *Suppose we have a model with M points, a video sequence with N frames, and we are given*

- *at least one known pose X_k^i (usually assumed to be the initial pose i=0), $k = 1, \ldots M$*
- *$D(X_m, X_n) = l_{mn}$ for some pairs mn*
- *image observations points (u_k^i, v_k^i) for $k = 1, \ldots M$, points with uncertainty d (i.e. eq. 1), for frames $i = 0, \ldots, N$*

Find the best pose (3D position of X_k^i) in each frame i, satisfying these constraints.

In this form it is clear that the image observations constrain the 3D point to lie within a region (a tetrahedron defined by the optical center and the image pixel region) rather than along a ray. In solving the constrained optimization problem as stated there is a *space* of possible solutions. While other assumptions (e.g. orthographic projection and precise image information) can lead to parameterized families of solutions, such a succinct description of the solution space is not possible with our constraints. The distance constraint is a quadratic constraint on a 3D position thus describing a surface and the image uncertainty inequalities are planes that bound portions of the distance constraint surface. Depending on the particular geometry there can be zero, one, or two connected surface regions[3].

Consider points A and B, with $D(A, B) = l$. If the coordinates of A are fixed then the rotations of the link $\overline{AB}$ that can occur while still satisfying the image observations for B are bounded by

$$\theta = \cos^{-1}\left(\frac{\overline{AB_1} \cdot \overline{AB_2}}{\|\overline{AB_1}\| \, \|\overline{AB_2}\|}\right)$$

where $B_1 (B_2)$ is the intersection of the image uncertainty plane and the quadratic surface (distance constraint) with the maximum (minimum) Z value (fig. 1). If all four image constraint planes intersect the distance constraint surface then there are two disconnected regions of possible points and θ can be computed for *each* component. Only motion in the plane parallel to the image plane is fully observable.

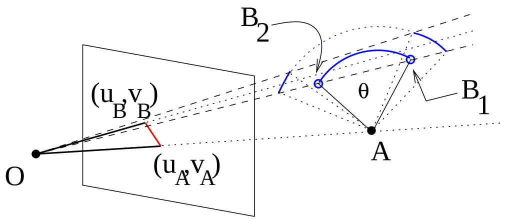

Figure 1: Rotation within the plane defined by the image constraints and the optical center is not observable. The dashed lines represent the image observation uncertainty for B. B_1 and B_2 are at the extremes of a region that satisfies the constraints. Also shown, in dotted lines, are two disconnected regions resulting from a larger distance constraint.

If an additional link (with endpoint C) is connected to B, each possible position of B describes a surface of points for C. The union of these surfaces form a volume (again, possibly disconnected) in the space of possible positions for point C. The additional constraint may eliminate spurious solutions (i.e. it may resolve the ambiguity for the position of B), or it may multiply the number of possible solutions.

In practice we find that for reasonable image noise models (1-2 pixel errors) there are two possible solutions for the single limb case (i.e. the situation depicted by the dotted lines in figure 1). Under reasonable viewing assumptions (viewing a 25cm link at a distance of 1.5m – e.g. a forearm) the range of values of θ depends on the orientation of the link with respect to the camera. Assuming 1-2 pixel error in image observations, for fronto-parallel views the angular variation tends to be on the order of 5 degrees, for views oriented roughly parallel to the optical axes, angular variation tends to be on the order of 1 degree for each component, however the two components are typically far apart (upwards of 120 degrees).

Here we have considered only the uncertainty in the image observations. If one considers uncertainty in the optical parameters and in the link lengths the region of uncertainty grows.

The advantage of using the constraint based method is that we can easily add additional constraints (e.g. lines, other known poses) if they are available. For example, the above ambiguity in joint position may be resolved using additional constraints on the pose (e.g. a knee joint motion can be constrained to allow only a physically plausible range of motion).

5 Empirical Assessment

To verify our analyses, we have implemented the constraints of the last section within *Timelines,* our constraint-based animation testbed. This system provided facilities for many aspects of character animation, and allows us to examine our results in the context of the application. Timelines provides a non-linear constraint solver that can handle a wide array of constraint types, such as self-intersection and footplants, however for these discussions we consider only the constraints discussed in Section 4. To test the image constraints, various simple predictors (0th and 1st order) were used on a per-frame basis.

To create synthetic examples, we began with high-quality motion capture data created using an optical motion capture system by a premier motion capture service provider. We specifically chose ex-

[3] If the region of uncertainty of two image observations overlap on the image plane then there are infinite possible positions for the two points (i.e. the distance constraint does not help)

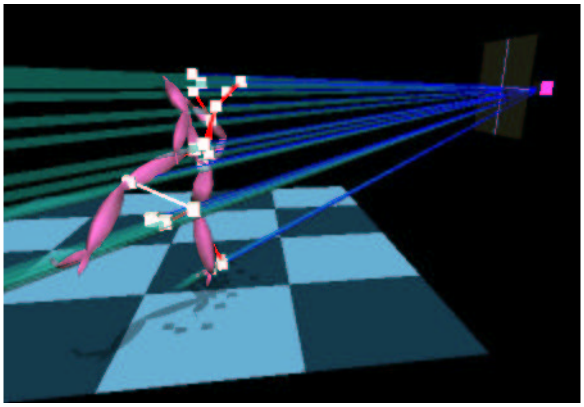

Figure 2: A frame in the reconstruction of a karate kick sequence. The ellipsoidal character is the original "ground truth" motion capture data. The simulated camera's image plane and focal point are shown in the upper right, the image constraint cones are also shown. The reconstructed motion is shown as bright cubes. Notice that the left lower leg has entered the "wrong" solution. This solution is a valid reconstruction, and in this case is as plausible a pose as the original.

amples that were of sufficient quality for animation, and would be too difficult for even the state-of-the-art computer vision methods. The motion capture data contained location of the skeleton joints at discret time steps This motion capture data was "filmed" by a virtual pinhole camera, and the locations of the joints on the virtual camera's image plane recorded as simulated tracker output. This simulated data is "ideal:" our virtual camera is a pinhole camera for which we know the exact parameters (and parameters were chosen to be consistent with a standard digital video camera), the model is exactly a rigid skeleton with fixed link lengths, and the observations exactly record image coordinates without noise.

The synthetic data allows us to consider ideal circumstances. We have a precise camera calibration, initial pose, and image measurements (we used 1 pixel boxes). Arguably, we can never expect to do better than under these ideal situations. As pointed out in section 4, the constraints can lead to multiple solutions. Without additional constraints the solver "gets stuck" in the wrong local minima. In practice, we find this case occurs frequently.

While the constraint-based algorithm succeeds in many cases, it fails surprisingly often. Figure 2 shows a particularly illustrative example. In this motion, a martial artist performs three swift kicks. The depicted frame shows a case where the motion matches the original as far as the constraints are concerned, but has some substantial differences. Were the ground-truth motion not shown, the pose of the left leg would have seemed perfectly fine.

The obvious improvements to the simple algorithm would not solve this example problem. For example, a first-order predictor or continuity-based method would most certainly be thrown off by the substantial high-frequency "snaps" that occur during the kicks. The failure of prediction to apply might be expected: the kicks were meant to surprise opponents, why should we expect our algorithms to fare better?

The synthetic framework also allows us to explore how the constraints degrade with noise. From a practical viewpoint, the addition of our uncertainty model is trivial: the equality constraints on image positions are replaced by inequality bounds. Experimental results show that the performance of the algorithm degrades gracefully as the bounds are enlarged. Figure 3 shows an example result. Not surprisingly, as the tolerance is enlarged, the small details of the motion are removed, however, the overall content of the motion is maintained.

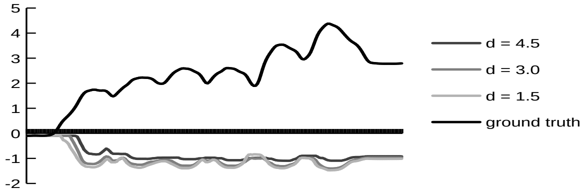

Figure 3: A graph of reconstructed position vs. time vs. the tightness of the bounds on image observations (d in equation 1). The graph of the Z coordinate shows the situation where, because of the depth ambiguity, the solver takes a different "fork" of the potential solution than the ground truth.

In a similar vein we can perturb other parameters such as the optical parameters. Again we find that the algorithm performs on-par with the noise free case for small errors, and degrades as the error in the assumed optical parameters rises.

6 Discussion

Vision-based motion capture is attractive yet challenging. A solution that meets the needs of animation seems even more so: not only is animation a particularly demanding application of motion capture, its demands seem directly at odds with many of the approaches one might take towards building a video-based capture system.

The constraint-based approach provides a tool for understanding the limitations of processing observations to reconstruct motion. Our analysis, and experiments, show that a single video stream coupled only with weak constraints does not come close to providing sufficient capture performance for animation applications. Clearly, additional information is required to provide more constraints that will better narrow the potential solutions.

More information can have the form of additional constraints, or better predictive models (used in the prediction phase). The latter can be difficult for animation applications as we hope to capture unusual and unpredictable motions. The former may come from a variety of sources, such as additional observations, more detailed models of the limitations of human movement, and detected interactions between the character and its environment.

The constraint-based approach allows us to understand the effectiveness of additional information as we find effective ways of determining it, and to implement capture systems based on these formalisms. For example, commercial optical motion capture system, such as those sold by Vicon [22] or Motion Analysis [15] are distinguished from video-based techniques as they use special hardware and a controlled environment to make observations precisely. Such systems engineer away the vision problem by using controlled lighting, camera optics, and markers to facilitate identification and localization. Because of the known shape and small size of the markers, and the careful understanding of the camera optics, motion blur and partial occlusion can be accounted for. This allows very precise 2D localization. Extremely careful calibration amongst multiple cameras provides precision in three dimensions. Such systems presently use very weak motion models, typically viewing the points individually, and using continuity only for determining correspondence. Very high sampling rates allow for continuity and filtering to be effective tools for noise reduction.

Within the constraint-based framework, we can see that optical tracking systems provide their better results by using much stronger constraint information. Sub-pixel marker localization and multiple cameras provide considerable cues as to the location of individual markers. This allows avoidance of stronger predictors and world models. The markers can often be tracked to a degree of precision that the tractable, skeletal approximations do not afford, and can handle the novelty and nuance of animation applications. Unfortunately, this requires expensive equipment and a controlled environment to achieve.

6.1 Taxonomy of Computer Vision Approaches

A video motion capture solution, based on computer vision techniques, must somehow create motion data using the limited information provided by video. The considerable activity in the vision community over the past several years have focused on a few key strategies:

1. Better localization of features by use of regions.
2. Stronger geometric models to restrict possible poses.
3. Stronger motion models to predict likely poses.
4. Novel applications that accept coarser pose results.

As discussed in Section 2, strategies 3 and 4 are unlikely to be of much assistance for animation applications.

The published results of current vision research do not meet the quality demands of motion capture for animation. At best, these techniques operate on simple motions (e.g. walking), and produce a level of fidelity where these motions are recognizable but details are absent. Here, we consider some of the more recent results.

Our discussion of section 4 assumed point features. Larger features may offer better opportunities for precise localization by incorporating more information. For example Ju et. al. [11] track all pixels associated with each body segment. This is extended by Bregler and Malik [3], and Yamamoto et. al. [24] to get 3D information by interpreting the optical flow within these segments. Edge based information is used in [4, 5, 17], and template matching is used in [10] to recognize body segments.

Strong motion models are often used to facilitate tracking by focusing the search to poses that are likely to follow from the previous motion. For example, Rehg [18] and Bregler and Malik[3] use a dynamic model within a Kalman Filter framework. More recent work uses more sophisticated motion models. An example of a strong model is the work of Brand [1] which generates motion by driving a trained Hidden Markov model from observations. The system is restricted to playing clips from the training corpus, limiting the variability of its output. In between are systems that use statistical models to train predictors that provide likelihood estimates for poses [21, 4, 5].

A problem with strong models is that they limit the system to those motions described by the model. For example Rohr [19] explicitly assumes a specific motion (walking), whereas Sidenbladh and Black [21] implicitly make this assumption by training their statistics only on walking motions. Bregler [2] builds a dynamic model of human running.

Geometric models (typically a kinematic chain for the skeleton) are used in many approaches (e.g. [10, 5, 4, 3, 17]). Probabilistic frameworks are incorporated to determine which poses are more likely.

Geometric and biomechanical constraints are less restrictive than motion models. Geometric models (typically a kinematic chain for the skeleton) are used in many approaches (e.g. [10, 5, 4, 3, 17]). This is attractive because it improves a system's ability to handle novel motions, but it is difficult to find models that are both mathematically tractable and sufficiently restictive. To date, the models employed in the video capture literature are merely kinematic chains [10, 5, 4, 3, 17]. More sophisticated geometric and biomechanical models seem to be a promising, but under-explored, direction for future video motion capture systems.

References

[1] M. Brand. Shadow puppetry. In *ICCV99*, pages 1237–1244, 1999.

[2] Christoph Bregler. Learning and recognizing human dynamics in video sequences. In *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, San Juan, Puerto Rico, June 1997.

[3] Christoph Bregler and Jitendra Malik. Tracking people with twists and exponential maps. In *Proc. IEEE Conf. Comp. Vision and Pattern Recognition*, 1998.

[4] J. Deutscher, Andrew Blake, and Ian Reid. Articulated body motion capture by annealed particle filtering. In *Proceedings of the Conference on Vision and Pattern Recognition*, volume 2, pages 126–133, Hilton Head, USA, June 2000.

[5] T. Drummond and R. Cipolla. Real-time tracking of highly articulated structures in the presence of noisy measurements. In *Proceedings of the International Conference on Computer Vision*, volume II, pages 315–320, Vancouver, Canada, July 2001.

[6] D.M. Gavrila. The visual analysis of human movement: A survey. *Computer Vision and Image Understanding*, 3(1), January 1999.

[7] Michael Gleicher. Retargeting motion to new characters. In Michael Cohen, editor, *SIGGRAPH 98 Conference Proceedings*, Annual Conference Series, pages 33–42. ACM SIGGRAPH, Addison Wesley, July 1998.

[8] Michael Gleicher. Comparing constraint-based motion editing methods. *Graphical Models*, 63, 2001.

[9] Nicholas R. Howe, Michael E. Leventon, and William T. Freeman. Bayesian reconstruction of 3d human motion from single-camera video. *Advances in Neural Information Processing Systems*, 12, 2000 (to appear).

[10] Sergey Ioffe and David Forsyth. Human tracking with mixtures of trees. In *Proceedings of the International Conference on Computer Vision*, volume I, pages 690–695, Vancouver, Canada, July 2001.

[11] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In *2nd Int. Conf. on Automatic Face- and Gesture-Recognition*, pages 38–44, Killington, Vermont, Oct 1996.

[12] A. Menache. *Understanding motion capture for computer animation and video games*. Morgan Kaufman (Academic Press), 2000.

[13] Thomas Moeslund. Summaries of 107 computer vision-based human motion capture papers. Technical Report LIA 99-01, Lab. for Image Analysis, University of Aalborg, 1999.

[14] D.D. Morris and J.M. Rehg. Singularity analysis for articulated object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 289–296, 1998.

[15] Motion Analysis Corporation, 2001. http://www.motionanalysis.com.

[16] V. Pavlovic, J.M. Rehg, T.J. Cham, and X. Murphy. A dynamic bayesian network approach to figure tracking using learned dynamic models. In *ICCV99*, pages 94–101, 1999.

[17] Ralf Plänkers and Pascal Fua. Articulated soft objects for video-based body modeling. In *Proceedings of the International Conference on Computer Vision*, volume I, pages 394–401, Vancouver, Canada, July 2001.

[18] J.M. Rehg. *Visual Analysis of High DOF Articulated Objects with Application to Hand Tracking*. PhD thesis, Dept of Elec. Eng., April 1995.

[19] K. Rohr. *Human movement analysis based on explicit motion models*, chapter 8, pages 171–198. Kluwer Academic Publisher, Dordrecht/Boston, 1997.

[20] Hyun-Joon Shin, Jehee Lee, Michael Gleicher, and Sung-Yong Shin. Computer puppetry: an importance-based approach. *ACM Transactions on Graphics*, 20(2), april 2001.

[21] Hedvig Sidenbladh and Michael J. Black. Learning image statistics for bayesian tracking. In *Proceedings of the International Conference on Computer Vision*, volume II, pages 709–716, Vancouver, Canada, July 2001.

[22] Vicon Motion Systems, 2001. http://www.vicon.com.

[23] David Washburn. The quest for pure motion capture. *Game Developer*, 8(12):24–31, December 2001.

[24] M. Yamamoto, A. Sato, S. Kawada, T. Kondo, and Y. Osaki. Incremental tracking of human actions from multiple views. In *CVPR98*, pages 2–7, 1998.