Sensing Human Shape and Motion

James Davis
University of California, Santa Cruz

Proposal Summary: The research objective of this proposal is to measure human body *shape* and *motion* without augmenting the subject. The hypothesis is that replacing traditional cameras with high accuracy 3D shape measurement devices and utilizing a carefully constructed prior model of human surface shape are the critical factors that have been missing from prior attempts to meet this goal. The long term accuracy targets are shape to 1mm and motion to 1deg.

This research transforms human shape and motion tracking from a problem of optimization, to a problem of *human shape priors* and *real time 3D shape measurement*. Inventing these tools and evaluating the hypothesis that they provide more robust motion estimation comprises the intellectual merit of this work.

- Data measurement – This proposal requires 3D shape sensors which are 1mm accurate, work with moving subjects, and can support multiple simultaneous viewpoints. No existing sensors support these requirements. Structured light triangulation and time-of-flight (lidar) sensors have complementary qualities and deficiencies. We consider novel designs combining these modalities to achieve our requirements.
- Shape prior – This work seeks to build a very detailed and accurate model of human surface shape as a function of pose and identity. This model will be constructed initially from full body laser scans of a variety of people with different body types in a variety of different poses. When a sufficiently accurate realtime 3D data measurement system is available, a much larger training data set can be captured.

Technology for reliable and accurate measurement of humans will ultimately enable new applications in ergonomics, smart spaces, fashion, surveillance, surgery, security, health, user interfaces, and art. Examples include :

- *User interfaces*: The first likely user interfaces are in video games, perhaps a more important change will be interfaces for the disabled which will interpret their motions.
- *Smart spaces*: Elevators that know about people will minimize average wait time, and car airbags that know whether the passenger is an adult or infant will save lives.
- *Ergonomics*: Monitoring factory workers performing repetitive actions would allow both analysis and early warning of workplace injuries.

Collaboration with LANL: In addition to the overall goal of human shape and motion estimation, the PI would welcome collaborations on sub-tasks that may be separately of interest to LANL. Improving the accuracy of lidar and triangulation 3D sensors is one area. Building accurate prior models of humans is another.

Description: Human motion estimation has been investigated under a wide range of conditions using a wide range of methods. Nearly all of these can be discussed as a process of acquiring measured data and then fitting a model of what humans look like to this data. This is solved by minimizing a high dimensional objective function defined by the measured data and the prior model. Due to many local minima in this objective function, a non-linear optimization is typically required to estimate the global minima. Figure 1 shows an overview of this process.

The accuracy of human motion estimation can be improved by increasing the complexity of any of the three fundamental building blocks: measured data, prior model, or optimization method. The vast majority of existing literature on human motion estimation has focused on just one of these building blocks: improving the methods for optimization (also called inference) in order to avoid local minima in the objective function's complex noisy high dimensional space. In contrast, this proposal seeks to explore the hypothesis that the other two building blocks are critical components, using *extremely high accuracy measured data* and *surface shape priors* so that the objective function is more precise and less noisy, resulting in an easier solution.

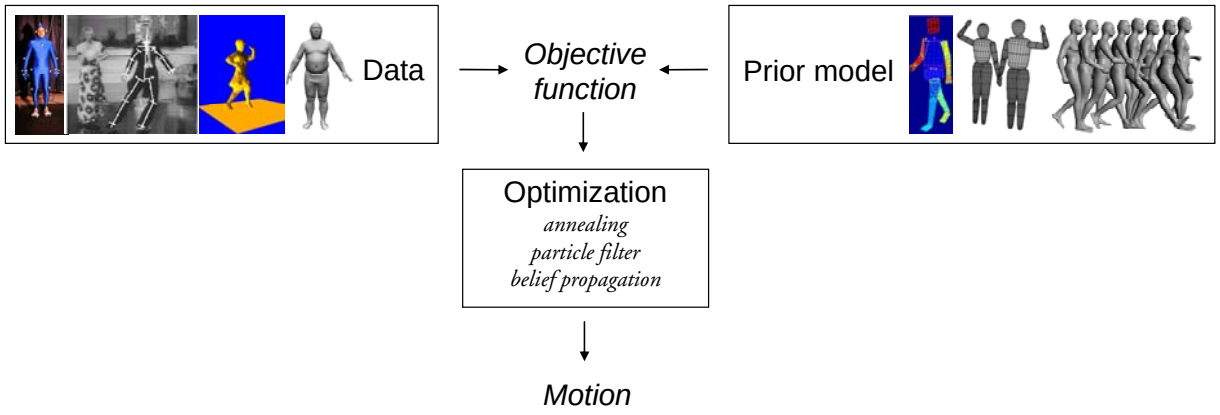

Figure 1 General overview of human motion estimation. Improvements in measured data, prior model, or optimization method can each lead to improved results.

3D Data Measurement: This proposal requires 3D shape sensors which are 1mm accurate, work with moving subjects, and can support multiple simultaneous viewpoints. Unfortunately no existing devices support these requirements.

Triangulation structured light sensors can be extremely accurate at close ranges, often a fraction of a millimeter for working volumes on the order of 1 meter. Unfortunately their accuracy falls when fewer patterns of light are possible due to quickly moving objects.

In contrast, time-of-flight sensors have a relatively poor accuracy of 10 to 50 mm for working volumes of several meters. However they have noticeably less degradation at a distance, have truly independent depth estimates at each pixel, and no difficulty with moving objects.

By noting that triangulation and time-of-flight range imagers have complementary qualities and deficiencies, we will design a combined sensor which performs better than either individually. The relatively low accuracy of a time-of-flight sensor can be used as an estimate of

depth which will alleviate the need to project low frequency patterns in time coded structured light. By projecting only the higher frequency patterns, faster moving objects can be accommodated.

Supposing 180 fps cameras and projectors are available and range samples are to be obtained at 60Hz, three time coded patterns can be supported. Fully disambiguated depth requires approximately ten patterns. If the highest frequency information is accurate to 1mm, three patterns will leave a phase ambiguity of 4mm. That is, we know the objects surface, but can't figure out which 4mm depth range it belongs in. Phase unrolling is one method to address this issue, but makes problematic surface continuity assumptions.

Although time-of-flight sensors are not accurate to 4mm at every pixel, they can be statistically this accurate if multiple pixels are aggregated. Combining this estimate with the detailed information from time coded structured light will allow a detailed surface shape to be reconstructed. The resulting combined system should be capable of real-time millimeter accurate range sensing.

High quality prior models of human shape: Most of the motion tracking literature has assumed a rigidly articulated model represented by simple shapes such as cylinders or truncated cones. Unfortunately, rigid articulation fails to capture important nuances of human motion. We define a *deformable model* in terms of deformations to a template mesh. Given a set of 3×3 deformation matrices: pose induced deformation Q, rigid rotation R, body shape variation S, and pose-shape cross correlation C, associated with a particular individual in a particular pose, the model can be used to synthesize a mesh. For each individual triangle a prediction is made for the edges. The predictions of neighboring triangles are not necessarily consistent. Thus to construct a complete mesh, $\mathbf{E}^i$, we solve for the location of each vertex $y_1,\ldots,y_M$ such that prediction errors are minimized in a least squares sense.

$$\mathbf{E}^i = \arg\min_{y_1,\ldots,y_M} \sum_k \sum_{j=2,3} \left\| R^i_{I[k]} S^i_k Q^i_k C^i_k \hat{v}_{k,j} - \left(y_{j,k} - y_{1,k} \right) \right\|^2 \tag{1}$$

Preparing shape training data: Preliminary raw 3D shape measurements have been acquired using a Cyberware full body scanner (Cyberware). The model of shape described by equation (3) requires raw measurements to lie in correspondence, accomplished using a sequence of existing methods: Scan merging (Curless and Levoy 1996), hole filling (Davis, Marschner et al. 2002), simplification (Garland and Heckbert 1997), establishing rough correspondence (Anguelov, Srinivasan et al. 2004), refining correspondence (Hahnel, Thrun et al. 2003), body segment labeling (Anguelov, Koller et al. 2004). This provides a set of training examples in exact correspondence. Each surface mesh has an identical number of faces and vertices, with identical connectivity. In addition each face is labeled with respect to our articulated skeleton template.

Initial experiments: Our preliminary experiments were reported in (Anguelov, Srinivasan et al. 2005). Figure 2 shows a sequence of meshes generated while changing the pose and identity

parameters. As an initial verification that this model is suitable for motion estimation, an objective function was used to estimate R^i directly from a marker sequence, as shown in Figure 3. We have also used the model for tracking image data, as shown in Figure 4, and reported in (Balan, Sigal et al. 2007). Estimating motion from high accuracy real-time 3D data will be critical to achieving the level of accuracy and robustness required.

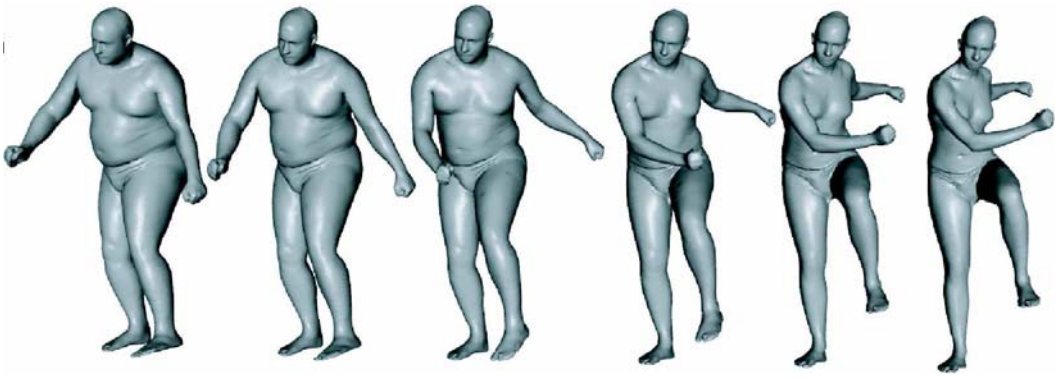

Figure 2 Sequence of meshes generated by the model while changing the pose and identity parameters on each frame.

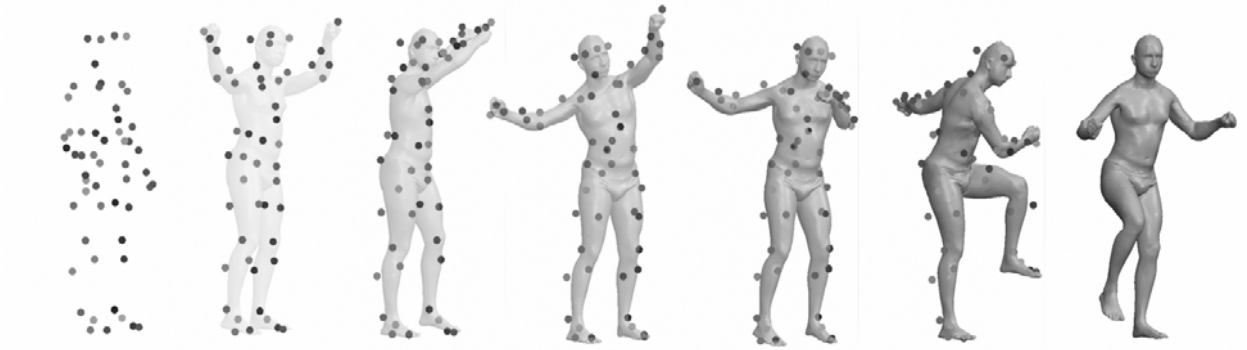

Figure 3 In preliminary work we have used the non-rigid body shape model for estimating motion from measured 3D marker positions. This proposal will replace the markers with real-time 3D surface shape measurement.

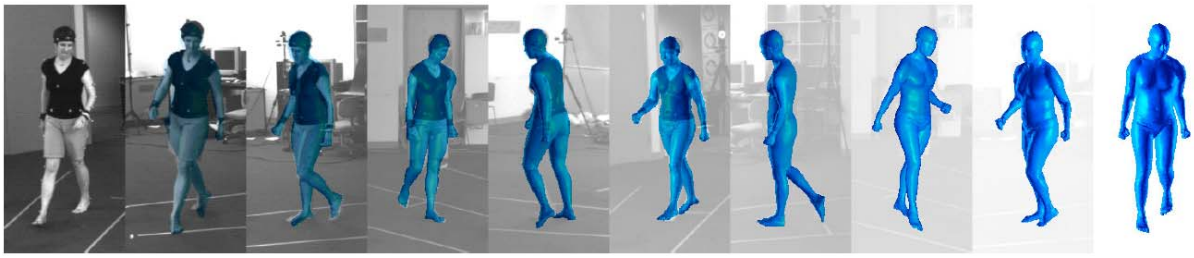

Figure 4 The model proposed in this document has been used to estimate motion directly from multi-view image sequences. This proposal will replace ambiguous image data with 3D surface shape measurement, leading to more robust tracking.