Human Motion Capture in Natural Environments

Zhiqing Cheng and Anthony Ligouri
Infoscitex Corporation
Dayton, Ohio, USA
zcheng@infoscitex.com, aligouri@infoscitex.com

Timothy Webb and Huaining Cheng
Air Force Research Laboratory
Dayton, Ohio, USA
timothy.webb.8@us.af.mil, huaining.cheng@us.af.mil

ABSTRACT

In this paper, the problem of capturing human motion in a natural environment is discussed from the perspective of needs, significance, scenarios, and technical challenges. The technologies that can be potentially used to capture human motion and activity in a natural environment are discussed, which include electromagnetic sensors, LED lights, inertial measurement units, range sensors, and computer vision-based markerless motion capture technology.

Two markerless motion capture methods for capturing human motion from video imagery are investigated and implemented in this paper. The first method uses a silhouette shape descriptor to describe silhouette shape and maps the silhouette shape descriptor (input vector) to joint angles (output vector) through a mapping matrix which is determined using *relevance vector machine*. The second method performs pose estimation by fitting a 3D human model to the silhouette through an iterative optimization. By minimizing the distance between the silhouette and the template skeleton-surface model that is embedded inside the silhouette, joint angles are estimated and thus pose is identified. The silhouettes extracted from human animation data are used for training the methods. The initial results of the two methods are presented and analyzed.

ABOUT THE AUTHORS

Dr. Zhiqing Cheng is a principal engineer and program manager for Infoscitex (IST) Corporation, providing support to the Air Force Research Laboratory (AFRL) for the programs of human identification and activity recognition based on human biosignatures. He has vast research experience in the areas of human modeling and simulation, artificial intelligence and machine learning, computer vision, and optimization. He has assumed many R&D projects as the principal investigator or as a major investigator and has published over 80 technical papers as the lead author.

Mr. Anthony Ligouri is a biomechanical engineer for IST, with undergraduate and graduate education and training in biomedical engineering and engineering mechanics from The Pennsylvania State University. He joined IST at the beginning of 2013 and has been working on the markerless motion capture technology development since then. His work included implementation of methods for estimating joint angles from silhouettes and using inertial measurement units for motion capture.

Dr. Timothy Webb has a Ph.D. in Analytic Health Sciences (Biostatistics) from the University of Colorado, Health Sciences Center. He has worked for the Air Force Research Laboratory as a Senior Research Statistician since 2007. His initial efforts in AFRL were identifying career fields within the Air Force that were at higher risk for injury and disability. Currently, his interests have transitioned to modeling human size, shape, and motion.

Mr. Huaining Cheng received the M.S. degree in flight dynamics from Beijing University of Aeronautics and Astronautics, Beijing, China. He also holds two M.S. degrees in mechanical engineering and computer science from Wright State University, Dayton, Ohio. He is currently a Research Computer Scientist with the Human Effectiveness Directorate, U.S. Air Force Research Laboratory (AFRL), Wright-Patterson Air Force Base. Prior to joining AFRL, he was a Principal Mechanical Engineer with General Dynamics Corporation. He has been involved in research on database and web design and implementation, engineering software development, computer simulation of human biomechanics systems, and human characteristics classification.

Human Motion Capture in Natural Environments

Zhiqing Cheng and Anthony Ligouri
Infoscitex Corporation
Dayton, Ohio, USA
zcheng@infoscitex.com, aligouri@infoscitex.com

Timothy Webb and Huaining Cheng
Air Force Research Laboratory
Dayton, Ohio, USA
timothy.webb.8@us.af.mil, huaining.cheng@us.af.mil

INTRODUCTION

The problem to be addressed in this paper is human motion capture in natural environments and settings, which may refer to any of the following conditions:

- Natural light, shadow, and occlusion, natural terrain and background, and natural scenarios;
- Humans in natural appearance with street clothing carrying some objects;
- Humans in natural states such that he/she moves or performs actions/activities freely;
- Humans in a natural context interacting with other humans or their surroundings.

Compared to traditional motion capture that is often conducted in a laboratory environment under controlled conditions, human motion capture in natural environments can potentially provide greater benefits and more unique features including:

- High biofidelity. Since human motion will no longer be restricted by suits, markers, or other sensors being placed on the body, the captured motion could provide higher fidelic representation of true human motion.
- True realism. Since a human moves or acts naturally, the captured motion is more realistic and natural.
- Large variety of motion that is otherwise impossible to capture. For instance, when a human subject wears loose clothing, it is almost impossible for marker based methods to capture the true body motion.
- Minimum pre-setting and no need for subject cooperation. These are two unique features provided by markerless motion capture technology.

There are various industry needs and commercial scenarios where capturing human motion in natural environments becomes necessary, such as athletics and sports, health care, human machine interface, and the entertainment industry. Human motion capture in natural environments has a variety of important applications within the United States (US) Department of Defense (DoD). For instance, within the modeling and simulation (M&S) community, human activity M&S plays an important role in simulation-based training and virtual reality (VR). However, human motion/activity simulation provided by current human modeling tools/technologies is either artificially synthesized or based on data collected in a laboratory environment, which lacks sufficient biofidelity and realism. In order to describe and simulate human motion/activity in the real world, it is necessary to capture human motion in a natural or real world environment. For homeland security purposes, human motion capture and analysis from video streams recoded in natural settings (e.g., airports and security check points) can be used to recognize human intent and to identify human borne threats.

Depending on specific application scenarios, the requirements on the motion capture technology (MCT) to be used in natural environments may vary. However, important, common requirements are as follows:

- *Accuracy.* With respect to different applications or scenarios, the accuracy of joint angle estimation provided by a MCT can be defined at three different levels.
    - Low level: The focus is on pose identification where the joint angles associated with a particular pose can vary in a range. The applications or scenarios include machine-human interface, human intent prediction, and human activity recognition.
    - Medium level: The emphasis is on pose identification as well as joint angle estimation. The problems can be, for example, human activity replication/animation in M&S based training and serious games where high accuracy of joint angle estimation is required for the biofidelic replication of motion, but it could be lowered as soon as the motion looks real and natural.
    - High level: The focus is on joint angle estimation and gait/motion analysis. The applications involve biomechanical issues which require precise estimation of joint angles so that the relationship between

force and motion can be accurately determined. The problems include sports (e.g., athletic training), health (e.g., prosthetic rehabilitation, and gait and balance training), and the extraction of spatial-temporal biosignatures.

- *Efficiency*. Motion capture discussed in this paper includes recording motion data and processing motion data. Recording data uses sensors and relevant hardware whereas processing data mainly implements software to derive the required information, such as pose identification and joint angle estimation. Since extensive computation is usually needed in processing data, efficiency refers to the computational speed at which a designated task (e.g., pose identification) can be accomplished. Depending on specific tasks of motion data processing and applications, efficiency can be considered at two different levels.
  - Real time: Ideally, the processing of captured motion data can be done in real time or nearly real-time so that the desired data or information can be provided for use in a timely manner. For many application scenarios, such as human-machine interface, immersive and interactive training, and security surveillance (e.g., human intent prediction and human borne threat detection), it is necessary to achieve real-time processing.
  - Off line: For many applications, such as those related to gait/motion analysis, activity replication, and biosignature extraction, real time processing is not necessary; instead, off line processing is acceptable.
- *Robustness.* It has two-fold implications: sensors and hardware systems can reliably acquire motion data under specified conditions, and the meaningful or desired data/information can be derived from the data collected. While it is desirable for a technology to perform robustly for every frame, given the complexity of human motion under various natural conditions, it is almost inevitable that a system or technology will fail for some ill-conditioned frames. However, it is necessary for a system to capture reliable motion data for key frames that depend on particular problems or applications.
- *Minimum setting/interference*. Capturing human motion in natural environments often requires minimum system setting (e.g., setting lights and placing markers or sensors on a subject) or even prohibits pre-setting. In many application scenarios, such as security surveillance, it is impossible to have a subject's cooperation, and it is preferable to avoid subject awareness.

Due to the complexity of human motion and the variety of natural environments, capturing human motion in natural environments has many hurdles to overcome. With respect to data collection, the major technical challenge is acquiring reliable, useful, and complete data under various conditions. With respect to data processing, the main technical difficulty is quickly analyzing the data to derive the desired motion data or information.

THE STATE-OF-THE-ART OF MCT

Retro-reflective optical motion capture technology, as a gold standard in accuracy, is widely used and commercially provided by many vendors using various optical systems. The technology relies on line of sight between multiple cameras with light emitting strobes and retro-reflective markers placed on the subject. The optical systems are, however, cumbersome to move and cannot be used with common attire or street clothing, which inhibit its use in natural or real-world settings. Besides, using a marker tracking system is sometimes disadvantageous because of the mere fact that markers must be placed on the body. Placing markers on the body not only introduces error in the skeletal position due to soft tissue artifacts but also changes the way subjects move, although the change is slight in most cases. For example, when markers are placed on the medial aspects of the arms and legs, some subjects will tend to walk bow-legged and with their arms out to avoid knocking off markers as the legs pass each other and the arms pass the torso.

The technologies that can be potentially used to capture human motion and activity in a natural environment include electromagnetic sensors, LED lights, inertial measurement units, range sensors (e.g., Microsoft Kinect), and computer vision-based markerless motion capture technology. Electromagnetic sensors provide accurate orientation and position, but are greatly limited by the range of the generated magnetic field. In recent years, depth cameras such as the Microsoft Kinect have become available for full body motion tracking at reasonable prices. However, like optical motion capture systems, depth cameras have relatively narrow fields of view which lead to very limited workspaces. Because these depth cameras are designed to use infrared sensors, their performance reduces significantly in outdoor environments under direct sunlight. Inertial measurement units (IMUs) could be used for human motion capture with great portability and flexibility. They can work almost anywhere, but are unable to maintain long-term stability and accuracy. Computer vision based markerless motion capture (MMC) approaches rely on image streams from one or multiple cameras for human motion analysis. It could maintain long-term tracking with a certain degree of accuracy, but may often produce inaccurate results due to occlusion. Based on their application potentials, IMU and MMC technologies will be discussed in more detail below.

Inertial Measurement Unit (IMU)

An IMU is a device that uses a 3-axis micro-electro-mechanical system (MEMS) gyroscope, 3-axis MEMS accelerometer, and 3-axis MEMS magnetometer to track the device's orientation in a global reference frame via sensor fusion of all 3 sensor outputs. A constant global reference frame is established by finding the downward vector of gravity via the accelerometer and an orthogonal vector pointing north using the magnetometer. A third vector orthogonal to the others to complete the 3D reference frame is calculated by taking the cross product of the first two. Orientation of the device is then tracked with respect to this reference frame. When an IMU is attached to each body segment, the pose of the subject can be defined in terms of Euler rotation angles between IMUs attached to articulated body segments. There are a number of commercially available IMU systems that can be used for full body motion capture with a wide range of applications. Xsens (http://www.xsens.com), for example, is a fully developed system that includes a suit with IMUs imbedded in the material, and proprietary software for creating a subject specific model and recording of data. On the other hand, IMU's from companies like APDM (http://apdm.com) do not come with ready-to-use software, but with a software development kit that allows users to create their own model and software that is specific to their applications and completely customizable. All IMUs work in the same general way. The factors that determine their accuracy include the quality of the sensors and on board software that corrects or compensates for sensor drift and other disturbances. Magnetic interference from power lines, machinery, ferromagnetic materials, etc. is a problem inherent to all IMUs because the interference throws off the magnetometer that is essential to defining the global reference frame.

Extensive research has been performed on using IMUs in motion capture applications and on increasing accuracy and performance of IMUs. For instance, Cutti et al. (2010) developed a protocol to measure the torso and lower limb kinematics of children with cerebral palsy and amputees during gait in free-living conditions. Their protocol consisted of 3 steps, placing IMUs on the body following some simple rules, computing the rotation axis of the knees, and measuring the IMU's orientation in a pre-defined body position. Using this protocol, the authors report root of mean square (RMS) error of 1.4 and 1.8 degrees and a standard error of 2.0 and 2.5 degrees for hip and knee angles, respectfully. However, creating a robust system is very difficult due to the disturbances from the factors described above. As noted by Favre et al. (2008) based on the ambulatory measurement of 3D knee joint angle, IMUs are often only reliable for a few minutes of recording. Various signal analysis methods were utilized to improve the performance of IMUs for motion capture applications. Among them, the Kalman filter is a common method being used by many researchers. For instance, by using the Kalman filter, Mazza et al. (2010) attained significant accuracy improvements of up to 11 degrees and RMS errors of less than 1 degree.

Markerless Motion Capture Technology

Markerless Motion Capture (MMC) is motivated by the need in many situations to capture the motion of people or objects outside of a motion capture lab in a natural environment. It has great potential applications including interactive training, clinical gait analysis, surveillance, and vision for autonomous systems, among many others. Using computer vision methods to perform MMC from a sequence of video images has been a central topic for computer vision research in the last two decades. Virtually all computer vision MMC methods can be grouped into two general categories: generative methods and discriminative methods.

Generative Methods

Generative methods typically use an optimization algorithm to minimize the difference between a template model and a measurement taken from the images. The variables (usually some measurement of subject pose and orientation) that produce the minimum difference are taken to be the true pose and position/orientation of the subject being recorded. They are called generative because they use a model to generate motion that matches the subject. They have the advantages of being general and robust to virtually any kind of movement or subject, but if the model or problem is formulated without appropriate constraints, they can produce unrealistic results. Various optimization algorithms also present a challenge to using generative methods, as each algorithm performs differently from the others. Some have parameters that need tedious tuning, and some algorithms may be more inherently suited to a type of problem than others.

The basic approach for generative methods involves the following steps:

a. A deformable 3D human model (e.g., a skeleton model or 3D shape model) is created. The model parameters include, but are not limited to position, orientation, joint angles, and shape parameters.
b. Silhouettes are used to compare the input images to the model. Segmentation algorithms are used to produce the silhouettes of the subject from input images, and a virtual camera is used to take projections of the model from the same perspective as the input images.
c. The silhouettes are compared in terms of certain metrics with a cost function. Regardless of metrics variation, the cost function output is a scalar value that quantifies the difference.
d. An optimization algorithm is used to find the optimum values of the model parameters for each frame that minimize the cost function.

There are many variations to this basic approach. One method by Kohli et al. (2008) called pose-cut, seeks to tackle the problems of segmentation and pose estimation together. Based on the representation of an image as a Markov Random Field, the method utilizes optimization to minimize a cost function that includes a term for segmentation and a term for pose estimation. Segmentation is performed using dynamic graph cuts where a stick man model is used as a shape prior. Pose estimation is achieved by fitting to the segmented image a shape prior that is deformed from the articulated stick man model. They showed that the rough, pose-specific shape prior provided by the model significantly improved segmentation results.

Another paper by Saito et al. (2014) tries to estimate body trunk shape and pose from silhouettes using a homologous human body model that treats human functional joints as the implanted vertices within body surface meshes. They create their deformable trunk model by analyzing a homologous model database using principal component analysis (PCA). The model then uses principal component projection coefficients (with 95% contribution ratio) and four joints with three degrees of freedom (DoF) in the spine to determine shape and pose of the model. The difference between the input images (torso silhouettes from the front and side views) and the projections of the model from the same perspective is represented by a cost function which is defined as follows. The input silhouettes are used to create a contour distance image which combines the distance transforms of the silhouette and its reverse image such that every pixel in the image has a value equal to the minimum distance to the silhouette contour. The projected silhouette outline of the model is then used as a mask and overlaid on the contour distance image. The cost function includes the pixel values of the contour distance image summed over each pixel of the mask and divided by the number of pixels in the mask. The cost function is then minimized using the covariance matrix adaptation evolution strategy (CMA-ES). They reported a mean error of only 6.67mm over 100 reconstructed torso models.

Discriminative Methods

Discriminative methods typically seek to describe the shape of a silhouette and then use a machine learning algorithm to determine the relationship between the silhouette shape descriptor and the joint angles as well as other model parameters from the training data sets. While these methods can be accurate, they are limited to the motions they expect to see based on the training set. Discriminative methods often have a hard time classifying movements that were not part of their training set. A method that has shown to be successful in tracking several motions from silhouettes was developed by Agarwal and Triggs (2008). In this method, a shape context is calculated for each silhouette image in the training set, which is a histogram of the angles and distances to the other points on the contour using 12 angular and 5 radial log-polar bins, creating a 60D vector. A k-means algorithm is then used to cluster the shape contexts into 100 clusters, the centers of which form a vector of 100 dimensions. The vector for each point on the silhouette is summed to create a 100D histogram for each silhouette. This is what they call the observational model, and the basis functions for this model are selected via relevance vector machine. A dynamic model (an auto regression model) based on the joint angles of previous frames constitutes the second part of the statistical model used in this method. The two models together form a robust statistical model that takes similar measures from input silhouettes and estimates joint angels and body gross motion based on the shape context and historical motion of the model.

A paper by Toshev et al. (2009) uses similar methods to track the motion of vehicles. While vehicles are rigid objects that generally lack separate segments connected by movable joints, the problem of fitting a model to a silhouette is the same. Toshev et al. creates a statistical model they call a model view graph. It consists of 500 silhouettes of each model taken from a uniform distribution of viewing angles. The edges connecting neighboring silhouettes represent view transitions that can be induced by motion of the model. A codebook is then created using shape descriptors of each silhouette, forming a 200D histogram for each model. Input images are then classified based on the best fit to the silhouette model histograms using the parameters of shape, orientation about the silhouette centroid, scale, and viewpoint. A score is assigned to each fitting and a maximum score is found using a backward-forward algorithm.

Silhouette Extraction

One common necessity shared by both categories of computer vision methods is the need for segmentation of the images before position and orientation of the subject can be estimated, meaning the subject in the image must be separated from the background in order to make an accurate estimate. This is often done by silhouette extraction, which generates an image that is just a white silhouette of the subject on a black background, but it is not the only method. Segmentation and silhouette extraction are separate, challenging computer vision problems in their own right. Some MMC methods tackle the problem of segmentation and pose estimation simultaneously, as the method described in Kohli et al (2008).

MARKERLESS MOTION CAPTURE TECHNOLOGY DEVELOPMENT

In a Small Business Innovative Research (SBIR) project sponsored by the US Air Force, efforts were made to develop a technology for markerless motion capture. One discriminative method and one generative method were implemented. Both require data for training and testing.

Training Data Generation

Using animated motion allows high quality silhouettes to be extracted thus making it easier to train the model. Therefore, the data (silhouettes) required for training both methods were obtained from human activity modeling and simulation through the following procedures (Cheng et al., 2011).

1. Use OpenSim (http://www.opensim.stanford.edu) to compute joint angles from the motion capture data of a human subject performing different activities. Create a biovision hierarchy (BVH) file for each activity based on the joint angles. The OpenSim animation of walking is illustrated in Figure 1.
2. Use Blender (http://www.blender.org) to create the shape model from the subject scan data and animate the model with the BVH file for each activity, as shown in Figure 2.
3. Create videos of the motion from a selected camera view (side view), as shown in Figure 3 as an example. The videos display a white image of the human subject on a dark background in order to make it easier to extract silhouettes.
4. Use OpenCV (http://www.opencv.org) to find the silhouette contour for each frame, as shown in Figure 4.

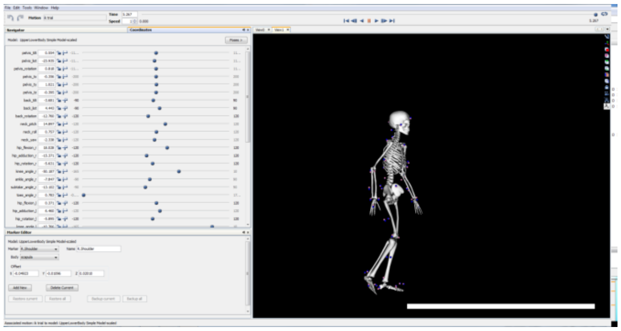

Figure 1.OpenSim animation of walking

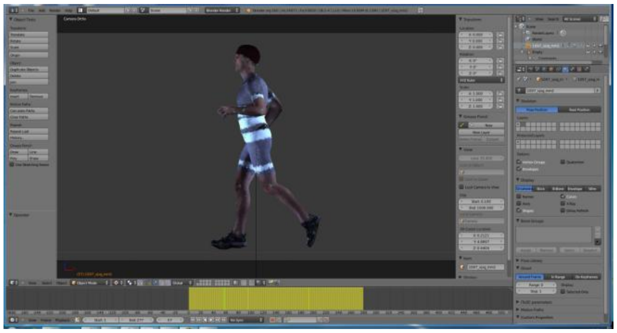

Figure 2. Blender modeling and animation

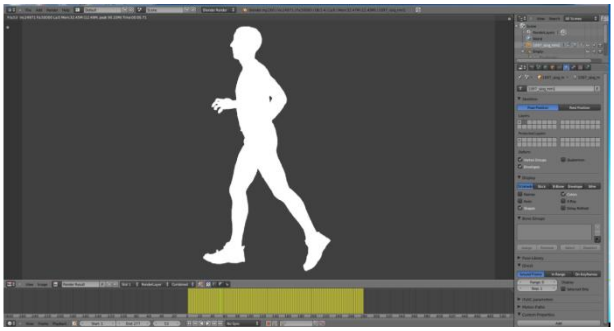

Figure 3. Image from side view of 3-D animation

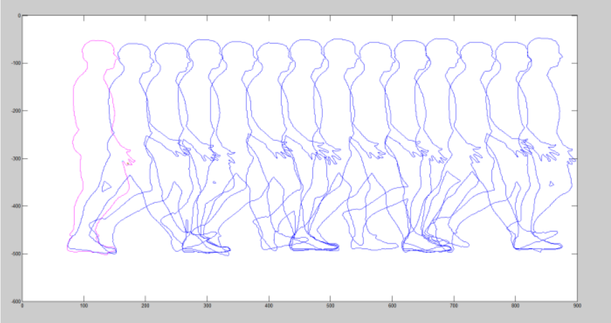

Figure 4. Silhouette contour for each frame

Method-1: Discriminative Method

The discriminative method implemented in this paper is based on the method proposed in Agarwal and Triggs (2008) with two major modifications. One modification is using the discrete cosine transform (DCT) to replace shape context (SC) as the silhouette shape descriptor. While SC is an effective silhouette shape descriptor, it is calculated for every point of the silhouette contour for each frame, thus taking excessive computer time and memory when a large number of silhouettes are analyzed. The DCT is computed on the entire silhouette contour for each frame, thus taking much less computer time and memory. All DCT coefficients or a truncation of its first part (the first 400 coefficients, for example) can be used to form a DCT coefficient vector for the silhouette shape description. The other modification is that a PCA is used to characterize the space formed by DCT vectors. Then each DCT vector was projected onto the eigenspace formed by the principal components. Instead of directly using DCT coefficient vector, the first 64 projection coefficients were used to describe the silhouette shape for each frame.

The silhouettes from nine subjects performing five activities (Cheng et al., 2012) were used to train the method. Another subject (subject 1100) performing the same five activities was used as the test case. By comparing the estimated joint angles with their true values (which were from the BVH files used in the simulation) for subject 1100, the RMS error for each joint angle is calculated and shown in Figure 5 for walking and jogging. Examples of the original motion of subject 1100 paired with a skeleton with the joint angles that were computed by the method are shown in Figures 6 and 7 for walking and digging, respectively. It can be seen that for simple periodic motions like walking the RMS error of most joint angles over the entire motion is less than 1 degree. However, as the motion to be tracked becomes more complex and dynamic, RMS error increases quickly, as seen in the jogging motion.

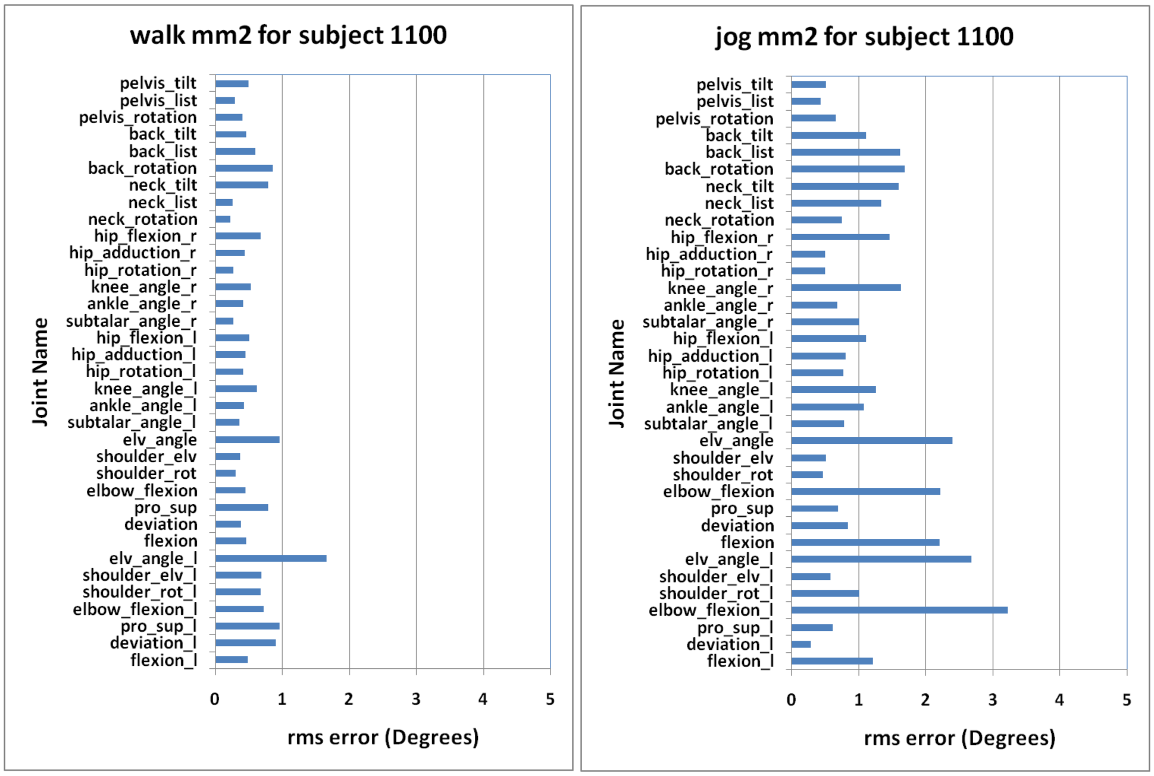

Figure 5. RMS errors for walk and jog activities

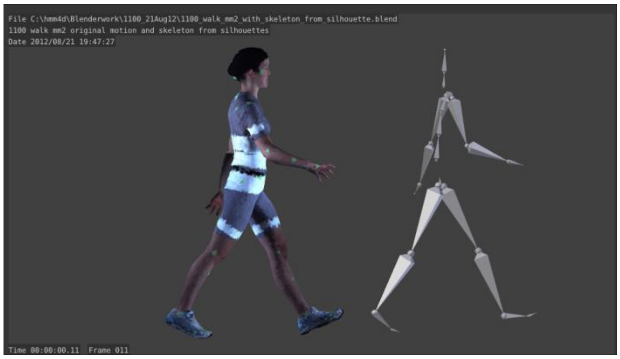
File C:\hmm4d\Blenderwork\1100_21Aug12\1100_walk_mm2_with_skeleton_from_silhouette.blend
1100 walk mm2 original motion and skeleton from silhouettes
Date 2012/08/21 19:47:27
Time 00:00:00.11 Frame 011

Figure 6. Subject 1100 walking

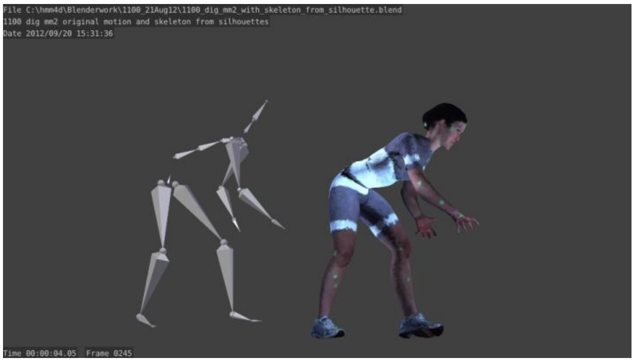
File C:\hmm4d\Blenderwork\1100_21Aug12\1100_dig_mm2_with_skeleton_from_silhouette.blend
1100 dig mm2 original motion and skeleton from silhouettes
Date 2012/09/20 15:31:36
Time 00:00:04.05 Frame 0245

Figure 7. Subject 1100 digging

Method-2: Generative Method

The generative method is based on the methods in Saito et al. (2014) described previously with several modifications. One is that a full body model with a fixed shape rather than a torso model described by PCA coefficients is used. The cost function was also modified to reflect the fact that only the pose and position of the model (not shape) is considered in the optimization and a penalty factor evaluation was added to help overcome problems with limb occlusion.

The animation videos of subject 1100 described above were used to create a contour distance image for each frame. It creates a pixel depth map with the lowest value (0) along the outline of the silhouette, and larger positive values farther away from the outline, both inside and outside the silhouette figure. For each frame of video, there are two views: 1) a head-on view aimed at the front of the subject and 2) an orthogonal view 90 degrees to the subject's right. A 3D deformable human model is then created. In order to create the best possible fit, the model reflects the exact shape of subject 1100. The model has 15 bones and 36 DoF that define the model's 3D pose, position, and orientation in space.

Using the CMA-ES algorithm (Hansen, 2009), the model parameters, i.e., joint angles and global 3D position, are optimized to best fit the silhouette. During the optimization iteration, the model is positioned according to the guess for the present function iteration. Front and side projections of the positioned model are then captured, creating similar silhouettes to those in the input video frame, which are then reduced to an image of the silhouette outline only using an edge finding algorithm. The length (number of pixels) in the model contour is calculated, and a penalty factor of 2 is applied if the contour lengths of the input image and model image are not similar. Otherwise, the penalty factor equals 1. The cost function is then calculated. Within the cost function, the front and side model outline images are overlaid on their respective contour distance images. For each view, the pixel values of the contour distance image are summed over the white pixels of the model contour and the sum is normalized by the number of pixels in the contour. The normalized sums of both views are added to calculate a total cost for the present iteration, which is multiplied by the penalty factor. The cost function is expressed as:

$$cost = \sum_{views} PF \times [(\int_{mask} CDI)/numPixels]$$

where PF is the penalty factor, views are the front and side view, CDI is the contour distance image, mask is the white pixels of the model contour image, and numPixels are the number of white pixels in the model contour image. This cost function is minimized to find the pose for each frame in the video. The cost represents the average distance in pixels between the outline of the model and the outline of the input silhouette, which for a perfect fit, would be 0.

The method was tested with walking, jogging, and throwing motions performed by subject 1100. The time histories of optimal joint angles were smoothed using a second order Butterworth filter with a cutoff frequency of 6 Hz. Animations of the model using the filtered joint angles look smooth and quite realistic. An example of the input silhouettes, a contour distance image, and the model in the optimized position for a frame of walking motion is shown in Figure 8.

The results of optimization were compared to the original BVH files used to create the test data described above to determine accuracy. Only major sagittal plane joint angles were compared, as they are the significant ones that are commonly analyzed in activity recognition and basic gait analysis. The RMS error for each joint angle for walking, jogging, and throwing was calculated. Of the three motions, walking had the smallest RMS errors. Over the length of the video, RMS error for the lower limb joints and lumbar joint, i.e. lumbar flexion, hip flexion, knee flexion, and ankle plantar/dorsiflexion ranged from 4.2-14.2 degrees. The upper limbs experience much more occlusion from the torso; consequently, upper limb joint angle errors, i.e. shoulder and elbow flexion, were significantly larger, ranging from 22.9-31.0 degrees. However, over the length of the video and multiple strides, limbs were occasionally confused by the model, which led to a solution with good fit but poor accuracy and increased RMS errors. Error calculations for one complete gait cycle with no limb confusion showed smaller RMS errors. Figure 9 shows a comparison of RMS errors for each joint angle between full video duration and one complete gait cycle of walking.

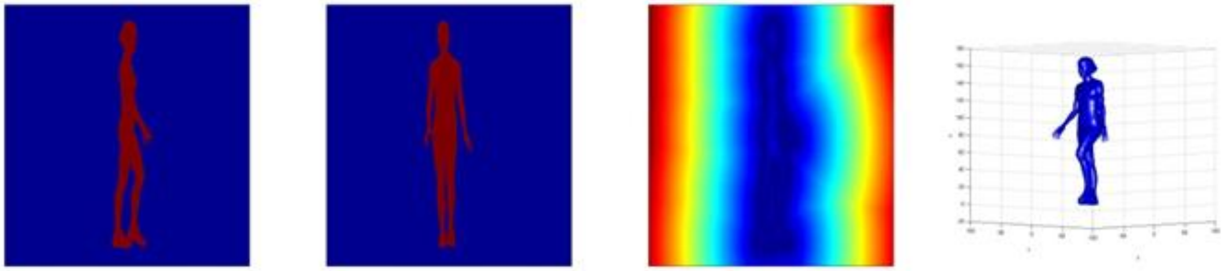

Figure 8. Input silhouette, contour distance image, and optimized model in 3D pose

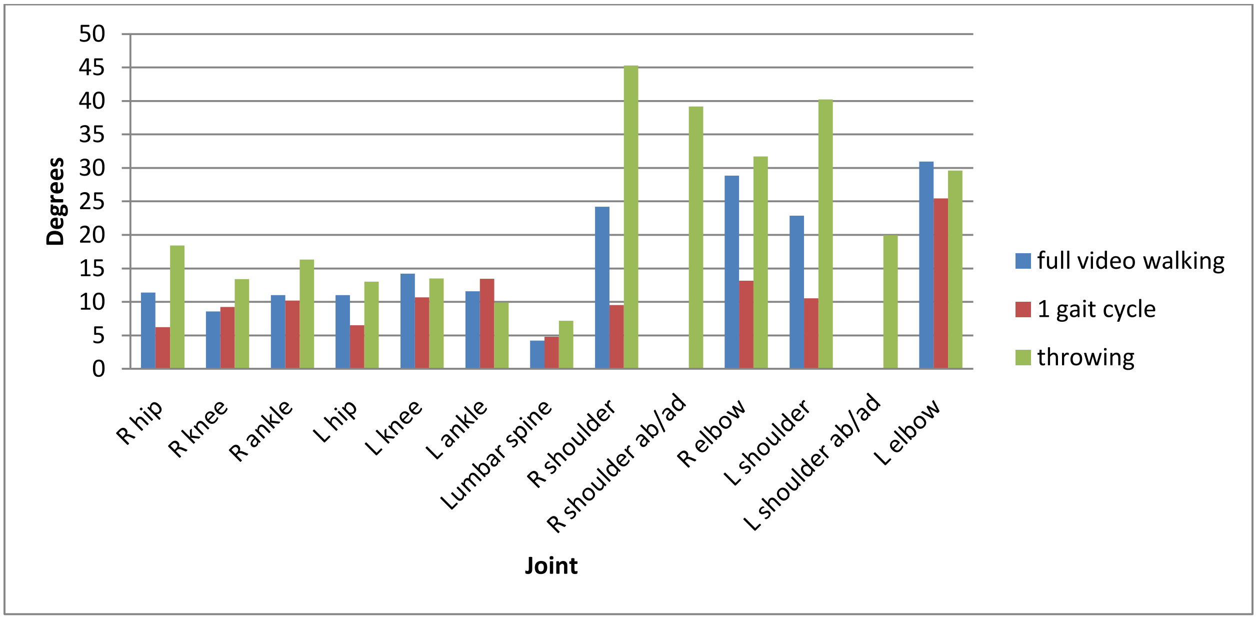

Figure 9. RMS error of major joint angles during walking and throwing

RMS errors for a throwing motion are also presented in Figure 9. Shoulder ab/adduction angles were also included in the error analysis because a large amount of motion happens about that joint axis during this motion. The throwing motion is much more dynamic than walking. Consequently we see higher RMS error across all joint angles. Like the walking motion, RMS error of the lower limb joints are much lower than the upper limb joints, with lower limb joint error ranging from 7.2-18.4 degrees, and upper limb joint angles ranging from 20 to 45.3 degrees. However, there is very little, if any, limb confusion.

Discussion

Both discriminative method and generative method implemented in this paper were able to track various human motions with varying degrees of accuracy. Both methods have advantages and disadvantages compared to the other. The discriminative method is faster and more accurate, but requires training for the motions it expects to capture.

The method also had difficulty with motions that were not periodic. On the other hand, the generative method does not require any training but captures motion with less accuracy. It can be seen from the results of both methods that the upper limb joint angles are estimated with greater error than the lower limbs. This is due to the limb occlusion by the torso. Whereas the lower limbs only occlude each other for a brief period of time, the upper limbs are often occluded by the torso for much longer time, especially in the side view. When occluded, there is no information in the silhouette describing the position of the upper limbs. Therefore, many positions can be accepted by the optimizer. Constraints must be chosen carefully to minimize these occlusion errors without constricting the model to the point where it cannot follow complex motions.

The error caused by limb occlusion could possibly be addressed by using multiple cameras to view the subject from different perspective angles. Additional views would provide extra information that would reduce the ambiguity of the pose when limbs are occluded in other views. The estimation error of the upper limb joints for the throwing motion can also be caused by the constraints applied to the model during optimization. The arm achieves very fast joint rotational velocity when throwing an object. If the joint velocity constraints on the model are too tight, fast movement of the arm will not be captured appropriately. This could possibly remedied by loosening the constraints; however, doing so would expand the solution space, leading to the increased solution times. It was also noted that there was far less limb confusion during the throwing motion than in walking. This is likely a result of the high degree of silhouette asymmetry in the frontal view of the throwing motion as opposed to the walking motion.

An interesting note is that despite the differences in accuracy between the discriminative method and the generative method, the replicated human motion from both methods for all activities still looks smooth and natural to the human observer. Therefore, we believe both methods could be used for pose/activity recognition and human motion replication/animation, since when different human subjects perform the same motion, their joint angels can have large variations.

CONCLUSIONS

There are various industry needs and commercial scenarios where capturing human motion in natural environments becomes necessary. There are various requirements on the motion capture technology to be used in natural environments. Due to the complexity of human motion and the variety of natural environments, capturing human motion in natural environments has great hurdles to overcome. Among the motion capture technologies that can be potentially used in natural environments, the computer vision based markerless motion capture technology has the largest potential, because it requires minimum pre-setting and has no need for subject cooperation.

ACKNOWLEDGEMENTS

This study was carried out under the support of SBIR Phase II funding (FA8650-11-C-6226) provided by the US Air Force. Thanks and appreciations are due to Mr. Stephen Mosher (retired) for his contributions to the markerless motion capture technology development while he was working at Infoscitex.

REFERENCES

Agarwal, A. and Triggs., B. (2006). Recovering 3D Human Pose from Monocular Images, IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(1):44-58.

Cheng, Z., Mosher, S., Camp, J., and Lochtefeld, D. (2012). Human Activity Modeling and Simulation with High Bio-fidelity, *Proceedings of I/ITSEC 2012*, Orlando, Fl.

Cheng, Z., Mosher, S., MtCastle, T., Smith, T., Parakkat, J., and Robinette, K. (2011). Biofidelic Virtual Terrorist—A Modeling and Simulation Tool for Human Threat Recognition Training, *Proceedings of I/ITSEC 2011*, Orlando, Fl.

Cutti, A., Ferrari, A., Garofalo, P., Raggi, M., Cappello, A., and Ferrari, A. (2010). ‘Outwalk’: a protocol for clinical gait analysis based on inertial and magnetic sensors, *Medical & Biological Engineering & Computing*, 48:17-25.

Farve, J., Jolles, B., Aissoui, R., and Aminian, K. (2008). Ambulatory measurement of 3D knee joint angle, *Journal of Biomechanics*, 41:1029-1035.

Kohli, P., Rihan, J., Bray, M., and Torr, P. (2008). Simultaneous Segmentation and Pose Estimation of Humans Using Dynamic Graph Cuts, Int. J Comput Vis. 79: 285-298.

Mazza, C., Donati, M., McCamley, J., Picerno, P., and Cappozzo, A. (2012). An optimized Kalman filter for the estimate of trunk orientation from inertial sensor data during treadmill walking, *Gait and Posture*, 35:138-142.

Hansen, N. (2009). Cmaes.m, *https://www.lri.fr/~hansen/cmaes.m*. Accessed March, 2014.

Saito, S., Kouchi, M., Mochimaru, M., Aoki, Y. (2014). Model-based 3D human shape estimation from silhouettes for virtual fitting, *Proceedings of SPIEI,* 9013.

Toshev, A., Makadia, A., Daniilidis, K. (2009). Shape-based Object Recognition in Videos Using 3D Synthetic Object Models, *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, FL.