Human Motion Analysis & Synthesis using Computer Vision and Graphics Techniques. State of Art and Applications.*

Ph. D. F.J. Perales
Computer Graphics and Vision Group
Department of Computer Science
Universitat de les Illes Balears (UIB)
e-mail: paco.perales@uib.es

ABSTRACT

In the late 1800s, Marey and Muybridge conducted independent studies of human an animal motion by shooting multiple photographs of moving subjects over a short period of time. Actually the study of human movement using a computer is very useful and can apply to many areas. One such application is the three-dimensional reconstruction of the structure of the human body and its movement using sequences of images and graphic models. For this reconstruction to be accurate and precise the person analysed and the virtual human (humanoid) must have similar anthropometric characteristics. In this paper presents a brief synthesis of actual systems from different viewpoints. First introduce the understanding motion capture for computer animation and video games applications. Second the analytical perspective is considered. So the system try to track, reconstruct and understand the movement that a person is doing in a sequence of images by computer vision techniques. The computer graphics and computer vision are inverse problems, but now we can combine techniques from one way to the other way and to solve the original problem more efficiently. Then ,in our particular sub-area, the articulated human motion, we can combine capturing techniques and animation systems to reach a realistic representation of motion and geometry of articulated complex systems. This problem is a challenging topic in both areas. In computer graphics, is very complex to animate realistic articulated figures like humans or animals. In computer vision, is very interesting design new algorithms to detect, track and recover articulate motion by example an avatar walking or jumping front a video camera. I try to introduce only, by space requirements, the principal ideas and the best actual works or systems. In the bibliography you can complete this topic with more detailed papers or books.

Keywords: motion capture, rotoscoping, humanoid, real and synthetic images, matching, calibration, graphic model, computer vision techniques.

1. INTRODUCTION

Human motion analysis and synthesis is receiving increasing attention from several communities of researches. In our particular scientific environment the computer vision and graphics researches are developing a lot of new theories and mathematical models to manipulate the human structure inside the computer world. This interest is motivated by applications over a wide spectrum of topics. In computer vision, segmenting the parts of the human body in a image, tracking the movement of joints over an image sequence, and recovering the 3D body structure are useful for precise analysis of athletic performance or medical diagnostics. In the aspect related to security, the police and military are interested to the capability to automatically monitor human activities using computers in airport, borders, and building lobbies. We can no forget the computer animation and video games industry, because the human avatars are very common in commercial films and console video games with a high degree of realism. The Internet network also introduce the possibility to develop video conference but using synthetic human faces as avatar or clone from the original video sequence. So, the potential number of applications are very high and in the next sections we introduce from a very synthetic form the main trends in computer animations systems and the computer vision approaches.

2. MOTION CAPTURE FOR COMPUTER ANIMATION

In general sense, motion capture is the process of recording a live motion event and translating it into usable mathematical terms by tracking a number of key points or regions / segments in space over time and combining them to obtain a three-dimensional representation of the performance. Then, using this technology we can translate a live movement or performance into a digital performance. The captured object could be anything that exist in the real world and make some motion. In the case that we use makers; these key points are the areas that

* This work is subsidized by the CICYT through the project TIC98-0302-C02-01 and is partially funded by the European TMR (training and Mobility of Research) project PARV (Platform for Animation and Virtual Reality), theEuropean Fund for regional Development and The Flemish Government.

best represent the motion of the subject's different moving part of the person.

Performance animation is not the same as motion capture, although many people use the terms interchangeably. Whereas motion capture pertains to the technology used to collect motion, the term performance animation means to the actual performance that the animation designers use to bring a character to life, independently of the technology used. The performance animation is the final product of a character driven by a performer, an motion capture is only the collection of data that represent the motion.

There are different ways of capturing motion. Some systems use cameras that digitize different views of the movement, which are then used to put together the position of key points or markers normally reflective. Others use electromagnetic fields or ultrasound to track a group of sensors. Also mechanical systems based on linked structures or armatures that use potentiometers to determine the rotation of each link are also available. Finally we can design combinations of two or more of these technologies to reduce the inherent limitations. But news technologies are also being, all aiming the possibility to make a real-time tracking of an unlimited number of key points or all the segments of the person with no space limitations at the highest frequency possible with the smallest margin of error. We are thinking in a non invasive systems that use computer vision procedures and can recover the movement from a non invasive optical data using sophisticate motion models. These systems are at the moment in early state but in near future will be use by commercial purposes. For a more precise historical and technical information you can read [1,2].

Briefly, we would like to present the types of motion capture systems using a classification criteria as outside-in, inside-out, and inside-in depending of where the captures sources and sensors are placed.

1) An outside-in system uses external sensors to collect data from sources placed on the body.
2) Inside-out systems have sensors placed on the body that collect external sources
3) Inside-in systems have their sources and sensors placed on the body

Examples of the first kind of systems are camera-based systems, which the cameras are the sensors and the reflective markers are the sources. Electromagnetic systems, are examples of inside-out systems, whose sensors move in an externally generated electromagnetic field. Finally the inside-in systems are electromechanical suits, in which the sensors are potentiometers and the sources are the actual joints inside the body. So the principal technologies used today that represent these categories are optical, electromagnetic, and electromechanical human tracking systems.

In the next few lines, I resume the advantages and disadvantages of the different human tracking systems.

Advantages of Optical Systems:

Optical data is extremely accurate in most cases.

A larger number of markers can be used.

It is easy to change marker configurations.

It is possible to obtain approximations to internal skeletons by using groups of markers.

Performers are not constrained by cables.

Optical systems allow for a larger performance area than most other systems.

Optical systems have a higher frequency of capture, resulting in more samples per second.

Disadvantages of Optical Systems:

Optical data requires extensive post-processing.

The hardware is expensive, costing between $100,000 and $250,000.

Optical systems cannot capture motions when markers are occluded for a long period of time.

Capture must be carried out in a controlled environment, away from yellow light and reflective noise.

Advantages of Magnetic Trackers

Real-time data output can provide immediate feedback

Position and orientation data are available without post-processing.

Magnetic trackers are less expensive than optical systems, costing between $5000 and $150,000.

The sensors are never occluded.

It is possible to capture multiple performers interacting simultaneously with multiple setups.

Disadvantages of Magnetic Trackers

The tracker's sensitivity to metal can result in irregular output.

Performers are constrained by cables in most cases.

Magnetic trackers have a lower sampling rate than some optical systems.

The capture area is smaller than is possible with optical systems.

It is difficult to change marker configurations.

Advantages of Electromechanical Body Suits

The range of capture can be very large.

Electromechanical suits are less expensive than optical and magnetic systems.

The suit is portable.

Real-time data collection is possible.

Data is inexpensive to capture.

The sensors are never occluded.

It is possible to capture multiple performers simultaneously with multiple setups.

Disadvantages of Electromechanical Body Suits

The systems have a low sampling rate.

They are obtrusive due to the amount of hardware.

The systems apply constraints on human joints.

The configuration of sensors is fixed.

Most systems do not calculate global translations without a magnetic sensor.

Also we can consider the digital armatures that like the mechanical suits consist of a series of rigid modules connected by joints whose rotations are measured by potentiometers or angular sensors. For more details please see [1], [2] and web pages of commercial systems.

Unfortunately all the above systems are invasive with the person that are doing the movement. The optical systems include minimal reflective markers but must use these key points to recover the global motion. The electromechanical suits are very invasive and are not practical for real competitions or specific event that require minimal interference.

Then, in my opinion, the future will be in human motion tracking and recognition systems that are based on optical devices but don't include any marker on the body. So the person can move freely around an indoor environment. Of course, this kind of system have some minimal requirements and the algorithms must be very robust to recover the all motion parameters under changes of illuminations and occlusions. In the next section I present a synthetic classification of this new perspective of human motion analysis and understanding.

3. MOTION TRACKING AND RECOGNITION USING COMPUER VISION TECHNIQUES

The study of motion in image sequences is a typical topic research area in computer vision. Motion is a powerful feature of image sequences, revealing the dynamics of scenes by relating spatial image features to temporal changes. The task of motion analysis remains a challenging and fundamental problem of computer vision. The are several approximations of this problem.

From sequences of 2D images, the only accessible motion parameter is the optical flow **f**, an approximation of the 2D motion field **u**, on the image sensor. Other possibility is the correspondence based techniques that try to estimate a best match of features within consecutive frames [17].

In particular case o human motion, the previous work was oriented to motion estimation of a rigid body. But the human body is a non rigid form. Then the new approach is consider the human body as a articulate chain and also elastic objects That's means that we need more sophisticate models and more complex algorithms.

One possible classification of motion approaches is depending if a priori shape models are used or the motion recovered is without any a priori model. In the next table we can see this classification.

Motion without a priori Shape Models
Input Description (points, relations, heuristics)
Feature Correspondence (points, lines, 2D contours, Blobs)
Motion Analysis (motion vectors, additional constrains)
3D Structure Recovery (depth information, non-linear eq.)

Model-Based Approaches
Model Definition (Stick to Volumetric model)
Modelling Motion (Kinetics, Kinematics)
Part Location (Region Tracking, Body labelling)
3D Structure Determination (2D features to 3D structure)

Also we can consider tree main areas of human analysis: 1) motion analysis of the human body structure, 2) tracking human motion without using the body parts from a single view or multiple perspectives, and 3) recognising human activities from image sequences.

For motion analysis of the human body structure, in general Non Model Based and Model Based methodologies follow the general framework that include a: 1) Feature extraction process, 2) Feature correspondence and 3) High-level processing. Then the global process include a computational high cost. The major difference between the two methodologies is in establishing feature correspondence between consecutive frames. Methods which assume a priori shape models match the real images to a predefined model. When a non a priori shape model are available, correspondence between successive frames is based upon prediction or estimation of features related to position, velocity, shape, texture, and colour.

In the case of human tracking, we don't need to recover the segments of the persons, then the principal characteristics of this approach are:

- Computational more efficient
- Track or recognise moving humans by uninterpreted low-level visual features
- Matching between images using pixels, points, lines, blobs
- Criteria: Motion, shape and other visual information
- Single or Multiple Camera Tracking

Finally we are interested to recognise the activity that the person is doing in the scene . A large portion of literature in this area is devoted to human facial motion recognition and emotion detection, which fall into the category of elastic non-rigid motion. We are interested in this paper only in human body activity recognition (HAR). In a synthetic way:

- Usually HAR is based on successfully tracking the human through images sequences
- High level task
  - State-space approaches (HMM, events, etc..)
  - Template Matching (Optical flow between frames, MEI/MHI, etc..)

A lot of literature exist about the different approaches presented. I introduce only a short selection of the more recently uses or proposed systems and for a full description of other systems you can refer to [3, 4, 5, 6, 7, 8, 9, 13, 14, 16]

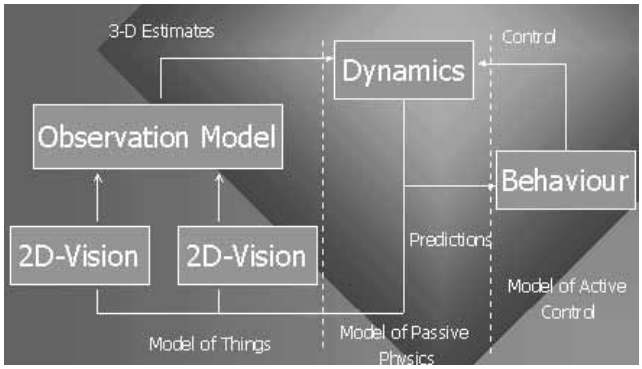
The Actual Methods (Non Exhaustive Survey) considered are:

1) Bregler & Malik Model [5]
2) P-finder (C. Wren), [3]
3) Understanding Purposeful Human Motion (A. Pentland, C. Wren), [6]
4) UIB Model, Computer Graphics and Vision Group, (F. J. Perales & J.M. Buades) [4], [13].

In the Bregler & Malik model is a new vision based motion capture technique that is able to recover high degree-of freedom articulated human body configurations in complex video sequences. In this model every segment is approximate by an rigid object. Grater proportion between numbers of measurements and DOF of the system improve the results and the global system will be more robust. The mathematical representation is based on "twist and product of exponential map formalism". The Initialisation is by user (1er frame) . The process of matching include an actualisation frame by frame of the cinematic chain. This product of exponential map and twists motions, and its integration into a differential motion estimation, solving simple linear systems enables to recover robustly the kinematics chain in noise and complex self occlude configurations. Unfortunately exist some limitations: Initial selection must be good, the shape of region object is assumed well defined, Orthographic scale projection (translation in Z axis is not computable), Human segments body approximated by 3D ellipsoids and only use one camera.

The systems proposed by C. Wren [3], named P-finder is a real-time system for tracking people and interpreting their behaviour. Uses a multiclass statistical model of color and shape Also the MAP (Maximun A Posteriori Probability) approach to detect and tracking of human body is used. The image primitive are the blobs and they use the blobs as clustered pixels that have similar image properties (colour and spatial). The 2-D region is considered as low-order statistics Gaussian Model blob. The systems incorporates a priori knowledge about people primary to bootstrap itself and recover form errors. Also I observe some limitations: P-finder expects the scene to be significantly less dynamic than the user. The systems can compensate small changes in illumination but no large changes. At the moment, only expects only one user in scene. In some cases has difficulty integration between Blobs and Contours features. Finally some errors in classification and feature tracking can lead to instability in the model

Remembering the classification of vision systems we can consider the last two systems as based on high level characteristics (HAR). The human motion understanding is a high challenging process that include low level primitives but also high level relations. The system proposed by A. Pentland [6] is based on the contextual knowledge encoded in the higher level models. Human body is a complex dynamic system, whose visual features are time-varying, nosy signals. So the Model of human body proposed goes from kinematics to dynamic perspective. Consequently a dynamic model is considered and the state vector includes velocity as well as position The state evolves according to Newton´s First Law. The system consider two types of constrains: Hard and Soft. The hard represents additional absolute limitations imposed on a system. The soft are probabilistic in nature The Observation Model is based on blobs with a Gaussian model. The Inverse Observation Model is a Maximum Likelihood (ML). The models of Purposeful Motion are based on Kalman gain matrix and Hidden Markov Models. Finally a behaviour Alphabet Auto-Selection is considered. The color is expressed in the YUV color space. In the image we can see a global diagram of the system.



Understanding Purposeful Human Motion [6]

Finally the UIB model [4][13]. is based on a biomechanical graphical model and matching process between primitives from the model and from the sequences of images. Our idea is that the system must be non- invasive and use multiple color cameras calibrate and synchronised. In [4] the system include a volume intersection process from the 2-D projections. Consider the large regions and try to represent this information with a minimum description. We try to reach a real and efficient integration of structural features extracted from two views using few assumptions on the human proportions and few assumptions on the type of movement. After a binarization process of the input images, the silhouette skeleton is found using an algorithm which decomposes the shape into regular and singular regions. Epipolar geometry is used to relate points from the two views. Nodes and segments are inserted or removed to obtain two graphs in alignment. We obtain a ·3D skeleton graph. The 3D Graph G is matched with a human model H, which doesn't need to have exactly the same proportions as the subject. Approximately collinear segments are grouped into paths. An interpretation is a partial one-to-one mapping from the set of paths of the model to the set of paths of the candidate graph G. An Algorithm A* performs a search starting with the trunk and following with its adjacent parts.

The criteria for the search find the minimum of the following: Geometrical difference between model and graph, Movement. Multiple matching of 2D segments (avoid phantom volumes), Parts non-matched in the 3D graph, Parts non-matched in the model. The heuristic used is admissible, so algorithm A* finds the optimal solution.

Unfortunately, the system proposed has some important limitations: In the actual implementation, the movement is unstable for several reasons: Regular regions of the same part are variable, The matching fails in some frames, because it takes too much time and we must prune the search The H-Anim model has not exactly the same proportions as the person. It is difficult to choose a point for the global position.

Consequently we propose a new model based on a color segmentation and a mutliview model matching [13, 18]. We do capturing the motion from different calibrated cameras. The matching process between humanoid and real person (semiautomatic and automatic) is proposed. The mains objetives are: accuracy in matching (depending applications), maximum automation and interactivity, non invasive and no use markers, extraction numeric values, view the movement from different points of view. We use a H.anim model of the person. To the

matching process we offer two methods: 1) Manually, the user do the main job. 2) Automatically, guided by the compute.

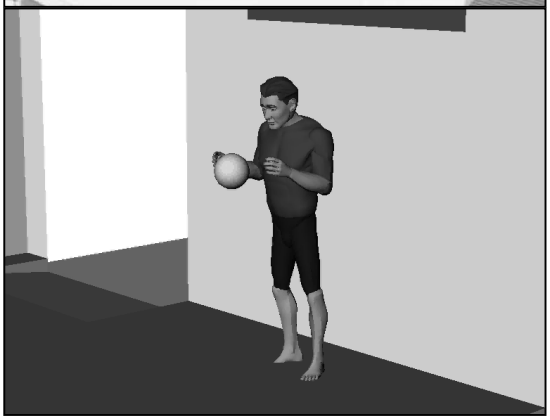
The automatic process can be corrected in all time. The humanoid modelled must be an exact replica of the human in antropometric proportions. From computer vision perspective is more interesting the second approximation the we create a database with the characteristics of the movement, the rotation of every joint in relation with all joints, getting from manual method. Estimate the position from a set of conditions: actual state depends of a old state, movement's speed, time coherence, etc.. We analyse the image and try to get the location of every joint and/or segment then we search the end-effector with the restriction to logic postures and continuos movements. Also is possible a tuning the process. The general matching criteria is "maximum overlapping function between projection of human graphical model and segmented images". Obviously we have the restriction to logic postures and continuos movements (Conditions) [13, 18]. At the moment we also working with a volume reconstruction of the person to fit the space of matching. In the next figures we can see some examples about our system.

4. EXAMPLES AND APPLICATIONS

In this section, we present some examples of original sequences colour images, segmented images, and overlapped model-person images. In the next table include a general classification of domains and specific areas considered. For our purpose the motion analysis and synthesis are priority but the number of possible applications is very high. Results are illustrated in consecutive the figures.

Applications include human images and syntehis interaction	
General domain	Specific Domain
Virtual reality	-Interactive virtual worlds -Games -Virtual studios -Character anjmation -Teleconferencing (e.g., film. advertising, home-use)
"Smart" surveillance systems	-Access control -Parking lots -Supermarkets, department stores -Vending machines, ATMs -Traffic
Advanced user interfaces	-Social intert.aces -Sign-Ianguage translation -Gesture driven control -Signaling in high-noise environments (airports, factories)
Motion analysis and sythesis	-Content-based indexing of sports video footage -Personalized training in golf. tennis, etc - Choreography of dance and ballet -Clinical studies of orthopedic patients - **Computer Vision Systems**
Model-based coding	-Very low bit-rate video compression

The images presented are the human model , the VRML synthetic model and the matched model with the original person.



5. CONCLUSIONS

The paper presented an short overview of actual human motion capture systems. The main idea was to introduce a computer animation and computer vision approaches. So the optical, magnetic and electromechanical technology is acceptable for computer animations purposes, but the computer vision researches are more interested in combine computer graphics and computer vision techniques to design a general tracking an recovering motion system without any kind of sensors in an minimum controlled environment. I know that is a very changeling problem but using robust high level models about the person structure and motion we believe that in a near future we can obtain goods results.

6. REFERENCES

/1/ Alberto Menache. “Understanding Motion Capture for Computer Animation and Video Games”, Morgan-Kaufman, 2000.

/2/ N. Badler, C. Phillips, B. Webber. *Simulating Humans. Computer Graphics Animation and Control*. Oxford University Press, 1993.

/3/ C. Wren, A. Azarbayejani, T. Darrell, A. Pentland. “Pfinder: Real-Time Tracking of the Human Body”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 7, pp 780-785

/4/ C. Yáñiz, J. Rocha, F. Perales. “3D Part Recognition Method for Human Motion Analysis”. CAPTECH ’98 Modelling and Motion Capture Techniques for Virtual Environments. pp 41-55.

/5/ C. Begler & J. Malik. “Vidoe Motion Capture”. Computer Science Division, Univ. Of California, Berkeley, Berkeley, CA 94720-1776.

/6/ Christopher R. Wren, Alex P. Pentland. “Understanding Purposeful Human Motion”, Submitted to ICCV 1999

/7/ J.K. Aggarwal, Q. Cai, W. Liao and B. Sabata. Nonrigid Motion Analysis: Articulated and Elastic Motion., Computer Vision and Image Understanding, Vol. 70, no. 2, May, pp. 142-156, 1998.

/8/ . J.K. Aggarwal, Q. Cai, Human Motion Analysis: A reviewComputer Vision and Image Understanding, Vol. 73, no. 3, March, pp. 428-440, 1999.

/9/ . D. M. Gravila, The Visual Analysis of Human Movement: A survey., Computer Vision and Image Understanding, Vol. 73, no. 1, January pp. 82-98, 1999.

/10/ . WonSook Lee, Jin Gu and Nadia Magnenat-Thalmann, Generating Animatable 3D Virtual humans from Photographs, EUROGRAPHICS2000 / M. Gross and F.R.A: Hopgood, , Vol. 19 (2000), no. 3.

/11/ .H. H. Nagel, F. J. Perales Ed., Articulated Motion and Deformable Objects, First International Workshop, AMDO2000, Palma de Mallorca, Spain, 2000, LNSC 1899.

/12/ Magnenat-Thalmann, D. Thalman, B. Arnaldi, Computer Animation and Simulation 2000. N., Sringer Computer Science,2000.

/13/ F.J. Perales, J. Torres. “A system for human motion matching between synthetic and real images based on a biomechanical graphical model”, IEEE Computer Society. Workshop on Motion of Non-Rigid and Articulated Objects, November 11-12, 1994, Austin Texas.

/14/ D.M. Gravila, L.S. Davis. “3-D model-based tracking of humans in action: a multi-view approach”, Computer Vision Laboratory, Proc. CVPR, pag 73-80, IEEE, 1996.

/15/] F. Perez, C. Koch. “Toward Color Image Segmentation in Analog VLSI: Algorithm and Hardware”, International Journal of Computer Vision, 12:1, pp 17-42, 1994

/16/ T. Nunomaki, S. Yonemoto, D. Arita, R. Taniguchi, “Multipart NoN-Rigid Object Tracking Based on Time Model-Space Gradients”, AMDO 2000, First International Workshop. Palma de Mallorca, September 2000. pp 72-82.

/17/] B. Jähne, H. Haubecker, P. Geibler. Handbook of Computer Vision and Applications, Vol. 2 Academic Press, 1999

/18/ J.M. Buades, A. Igelmo, F.J. Perales. "Modelos antropométricos a partir de secuencias de imágenes". CEIG 2000 X Congreso Español de Informática Gráfica. pp. 395-396, 2000.

7. WEB REFERENCES COMERCIAL SYSTEMS

Fifth Dimension Technologies. http://www.5dt.com/

Ariel Dynamics World-wide. http://www.apas.com/default.html

Animazoo Motion Capture.
http://www.animazoo.com/html2/anima2.html

Audio Motion
http://www.audiomotion.com/site/amframeset.htm

Motion Engineering

http://www.motionengineering.com/contactmec.html

Credo Interactive

http://www.credo-interactive.com/main_HTML.htm

House of Moves. http://www.moves.com/

MetaMotion. http://www.metamotion.com/contact.htm

Motek http://www.e-motek.com/newsroom/press/companypress/optical_dev.html

Pacific Data Images. http://www.pdi.com/corp/corporate.htm

Peak Technologies. http://www.peakperform.com/

Qualysis Precision Motion Capture. http://www.qualisys.com/

Vicon Motion Capture. http://www.vicon.com/animation/