

Human Body Segmentation and Matching using Biomechanics 3D models

J.M. Buades, F.J. Perales, M. Gonzalez
Computer Graphics and Vision Group
Department of Computer Science (UIB)
e-mail: paco.perales@uib.es

A.Aguiló, P.Martinez,
Nursing and Physical Therapy Department
Universitat de les Illes Balears (UIB)
e-mail: aaguilo@uib.es

Abstract

In many applications the study of human movement using a computer vision and graphics techniques is very useful. One of these applications is the three-dimensional reconstruction of the structure of the human body and its movement using sequences of images and biomechanical graphics models. We present a whole and general system to study the human motion without markers but, in this case, we apply it to high-level competition disabled swimmers. This kind of study needs accuracy on the analysis and reconstruction of the person's body, therefore the virtual human (avatar) must have similar anthropometrical characteristics than the person who is doing the movement.

We define a process to adjust the humanoid to the morphology of the person. It could be very laborious and subjective if done manually or by selection of points, but in this article we present a global human motion system capturing, modeling and matching a semiautomatic process between the real person and the modeled humanoid or synthetic avatar. Semiautomatic means that the computers propose the best matching from previous frames and the user can accept it or not.

Once this process is carried out we are ready to analyze and represent the movements under study. It must be defined a specific set of parameter for every kind of sport. The study is adapted to specific sport activities, in our case swimming activities with an additional complexity: water occlusion and distortion. In these cases the adjustment process must be assisted by the computer using rules and models to help the expert user in correspondence tasks.

1. Introduction

We will divide the general process of the system into four stages: in the first one, we capture images of the person from different points of view and of the background, in particular, up to four IEEE 1934 colour cameras are used, that means that a initialization process is needed to know the exact anthropometrical data of the person that is moving in front of the cameras. In the second stage, we select the humanoid with similar characteristics to the

original individual, so the result of this initialization process is an avatar with the same segments length of the real person. This process is need because in high-level sport activities the accuracy is very critical (in computer vision tracking the precision or accuracy requirements can be reduced). In the following stage we apply a semiautomatic process to obtain the humanoid adjusted to the person's measurements. The final goal is to reach an automatic recognition process, but this is a challenging task, that we are solving with less accuracy and under controlled environments. In the case of sport activities, a semiautomatic system is a good trade-off solution between full automatic systems and commercial standard systems using markers. The matching criteria propose a future pose of human segments based in previous frames and the user select the best one. Manual joints matching are not being done. The last stage combines the captured images and the generated humanoid to verify the result of the process and study the values we are interested in. We can use numerical data to make a deep study of performance of movements or paint special lines or points in real and synthetic images to help the trainer in the process of coaching to reach the perfect performance of the disabled sport people.

In particular the actual system is being used in indoor spaces and controlled environments. Running with a high degree of accuracy, in not controlled environments can lead to a less accurate matching. The system requires no markers or special clothing worn by the swimmer, therefore its range of application is very wide. In other cases, special clothing might help very much in segmentation process. It is also very important its portability into domestic environments using VRML 2.0 and the H-anim standard for specification of virtual humanoids. This is a great advantage for experts to visualize movements on any personal computer with a commercial Internet browser. The system is also defined using new IEEE 1934 digital standards with portable and low cost systems. In the following sections we explain the capturing, modeling, and segmentation processes and also the matching criteria. Finally we conclude with an explanation about the biomechanical parameters studied, some interesting results and future work.

2. The capturing process and modeling humanoids

Here we introduce the basic ideas of capturing system because digital cameras and IEEE 1394 become a standard and low cost solutions to implement a multiple camera color systems in real time. The capture software has been implemented with the API designed by Microsoft for Windows platform, this API called DirectShow allows you to use any camera (IEEE 1394, USB Web Cam, parallel port scanner, video file...) as long as you have drivers for Windows. Any kind of these input devices is programmed in a transparently and independent hardware way, without the need to modify our application. This API has been chosen with the intention to cover the most number of end users at a low cost without changing the capturing system. The video acquisition hardware used to test our algorithm is an IEEE 1394 controller Texas Instrument with three ports, which are connected to two Sony VFW-V500 color cameras that are able to synchronize with an external trigger. Therefore the results of this implementation have the following properties: versatility, independence of hardware, oriented object programming, efficiency and easily parallelizable.

In the module of human definition, there are various specifications [2] for humanoids; we have chosen the one created by the group h-anim¹ in VRML format for its portability and adaptability to different applications. The humanoid is composed of a collection of joints and segments structured in the form of a tree. Each joint corresponds to an anatomical joint (knee, shoulder, vertebrae), for example with each vertebra hanging from the one above, and the wrist joint hanging from the elbow joint. Each joint has associated with it a segment, it represents, for example, the elbow and the upper arm as its segment.

Each segment may have sites and displacers. A site corresponds to an endpoint, such as the fingertips, while a displacer corresponds to an effect applied to the segment. In particular we would like to define a multilevel system. That means that depending on the accuracy of applications, the system may include more degrees of freedom. For human motion recognition task particularly, the model is simpler and for human graphical simulations the system is more complex. Also, a modeling editor has been developed to adapt the human limitations of disabled people in this application. Figure 1 show a snapshot of the biomechanical editor. A simple animation tool is also included to generate simple kind of

movements. The final expected result is a virtual model of the real person compliant with H-anim standards.

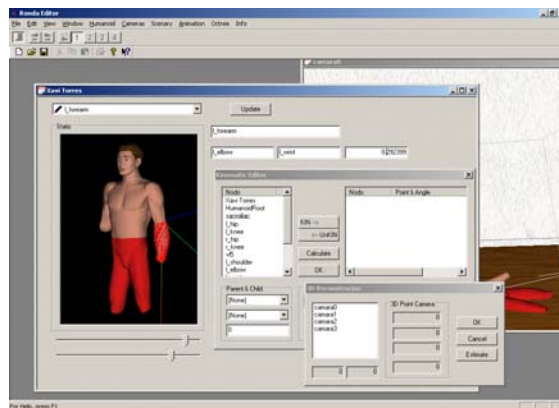
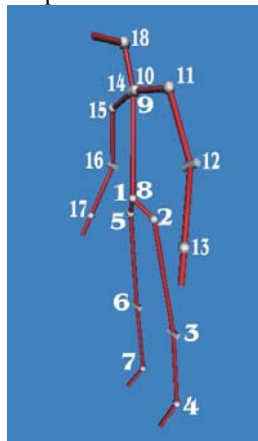


Figure 1. H-anim biomechanical editor (by CG&VG-UIB)

Another property of the system is the independence between hierarchical structure and graphical representation. That means, that we can change the graphical primitives and represent the same biomechanical model with wireframe, shapes or volumetric primitives. Of course a more simple biomechanical structure is used in the matching process. In particular, the body structure used has one guided joint, 13 spherical joints and 4 cylindrical joints. That means 43 degrees of freedom in total, but in special cases the joints are simplified. In figure 2 we can see the simplified model.



- Joints nomenclature:**
- 1: HumanoidRoot/sacroiliac (*guided*)
 - 2: l_hip (3 DOF)
 - 3: l_knee (1 DOF)
 - 4: l_ankle (3 DOF)
 - 5: r_hip (3 DOF)
 - 6: r_knee (1 DOF)
 - 7: r_ankle (3 DOF)
 - 8: vl5 (3 DOF)
 - 9: vc7 (3 DOF)
 - 10: l_sternoclavicular (3 DOF)
 - 11: l_shoulder (3 DOF)
 - 12: l_elbow (1 DOF)
 - 13: l_wrist (3 DOF)
 - 14: r_sternoclavicular (3 DOF)
 - 15: r_shoulder (3 DOF)
 - 16: r_elbow (1 DOF)
 - 17: r_wrist (3 DOF)
 - 18: skullbase (3 DOF)

Figure 2. Human biomechanical structure

3. Image color segmentation

Image segmentation is the first step in data extraction for computer vision systems. Achieving good segmentation

¹ www.hanim.org

has turned out to be extremely difficult, and it is a complex process. Moreover, it depends on the technique used to detect the uniformity of the characteristics founded between image pixels and to isolate regions of the image that have that uniformity. Multiple techniques have been developed to achieve this goal, such as contour detection, split and merging regions, histogram thresholding, clustering, etc. A Survey can be found in [7]. In color image processing, pixel color is usually determined by three values corresponding to R (red), G (green) and B (blue). The distinctive color sets have been used with different goals, and specific sets have even been designed to be used with specific segmentation techniques.

We define a color image as a scalar function $g = (g^1, g^2, g^3)$, defined over image domain $\Omega \subseteq \mathfrak{R}^2$ (normally a rectangle), in such a way that $g: \Omega \rightarrow \mathfrak{R}^3$. The image will be defined for three channels, under the hypothesis that they are good indicators of autosimilarity of regions. A segmentation of image g will be a partition of the rectangle in a finite number of regions; each one corresponding to a region of the image where components of g are approximately constant. As we will try to explicitly compute the region boundaries and of course control both their regularity and localization, we will use the principles established in [1, 8] to define a good segmentation. To achieve our goals we consider the functional defined by Mumford-Shah in [6] (to segment gray level images) which is expressed as:

$$E(u, B) = \int_{\Omega} \sum_{i=1}^3 (u^i - g^i)^2 d\mu + \lambda \ell(B) \quad (1)$$

where B is the set of boundaries of homogenous regions that define a segmentation and u (each u^k) is a mean value, or more generally a regularized version of g (of each g^k) in the interior of such areas. The scale parameter λ in the functional (1) can be interpreted as a measure of the amount of boundary contained in the final segmentation B : if λ is small, we allow for many boundaries in B , if λ is large we allow for few boundaries. A segmentation B of a color image g will be a finite set of piecewise affine curves - that is, finite length curves - in such a way that for each set of curves B , we are going to consider the corresponding u to be completely defined because the value of each u^i coordinate over each connected component of $\Omega \setminus B$ is equal to the mean value of g^i in this connected component. Unless stated otherwise, we shall assume that only one u is associated with each B . Therefore, we shall write in this case $E(B)$ instead of $E(u, B)$. A segmentation concept, which is easier to compute, is defined as follows:

Definition 1. A segmentation B is called 2-normal if, for every pair of neighboring regions O_i y O_j , the new

segmentation B' obtained by merging these regions satisfies $E(B') > E(B)$.

We shall consider only segmentations where the number of regions is finite, in other words $\Omega \setminus B$ has a finite number of connected components and the regions do not have internal boundaries.

A more detailed explanation of the concepts and their mathematical properties can be consulted in [1, 8] and we can see the properties of the functional in [1, 6]. The use of these theoretical concepts in the case of multi-channel images (e.g. color images) can be seen in [1]. We shall use a variation of segmentation algorithm by region merging described in [8] adapted to color images.

The concept of 2-normal segmentations synthesizes the concept of optimal segmentation we are looking for, and it lays on the basis of the computational method we use. The whole process is presented in a previous work [4]. For further explanation please refer to this paper.

The algorithm uses the RGB components because the segmentations obtained are very accurate to our goal. But the system is able to use other color space or color descriptor as we can see in [7]. Moreover, if it is needed it can weigh the channels used in order to obtain the segmentation. After the segmentation process we test every region to detect skin regions. Skin color is a characteristic color that is very different to other colors. Skin color test is applied, but underwater skin regions are not detected due to water distortion in light spectrum. In any case the results cases are not bad, and we are using only standard images from commercial video cameras. In the future we plan to use special light conditions and they might be infrared images. Some special clothing at end effectors can also help us in the segmentation process. In the Figures 3 and 4 we can see some colour segmentation results.

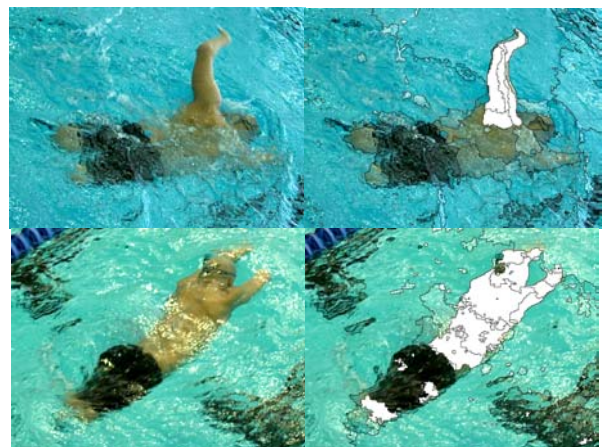


Figure 3. Arm and skin region segmentation



Figure 4. Whole human body segmentation

4. The matching criteria

In the proposed system, Figure 5 display a diagram of the task framework, the matching process is the kernel. Our main objective is to find a one-to-one correspondence between real and synthetic human body segments and joints, in every frame, in 3D space. The humanoid matching is currently done semiautomatically; we are working on automatic process for a general indoor environment and, for the moment we have reached promising results in cases where occlusions are not so long in time. A trade off solution is a semi-automatic process that means, a human interaction in matching process, but supervised by computer that avoids wrong biomechanical postures. As we have mentioned above, the system can estimate a posture from previous frames matching in time t_{i-1} . This estimation is based on a function that evaluates visual parameters (contours, regions, colour, etc.) and biomechanical conditions. The general method of this problem is known as *analysis-by-synthesis*. As we know, to search quickly in a 43-dimensional state is extremely difficult; therefore we propose some set of conditions to reduce our space search.

The information captured comes from the colour cameras in three positions. They follow specific anthropometric criteria, from which we obtain the desired parameters. We use a database of predefined movements and models to help the matching process. We can also use this knowledge to estimate new movements that are not previously recorded. The semiautomatic matching process has certain limitations of precision according to the models and the calibration of the cameras. This process is based on a set of well-defined conditions.

The matching process associates each joint with a 2D point in the image. This process consists on analyzing each image obtained from the cameras in an instant of time t . Once we have the articulation located in

two or more cameras we estimate the most accurate 3D point; a joint may only be detected in one or none image, therefore the process will have to be completed with contextual information from a higher level. Analyzing the sequence, we are able to apply physical and temporal restrictions, which help us to carry out the matching and to reduce the errors and the search space. This adjustment process is conditioned by a set of conditions, which are optimized in each case and type of movement. The restrictions are:

- Angle and distance limitations of the joints
- Temporal continuity of the movement in speed or acceleration
- Prediction of the movement based on a database of known movements, in the case of having non-visible joints.
- Collision of entities.

This matching process works currently in an automatic way for simple movement parallel to the camera plane and in a semiautomatic way in complex environments, thus making that it is possible to work at different levels of precision depending on the application the results are required for.

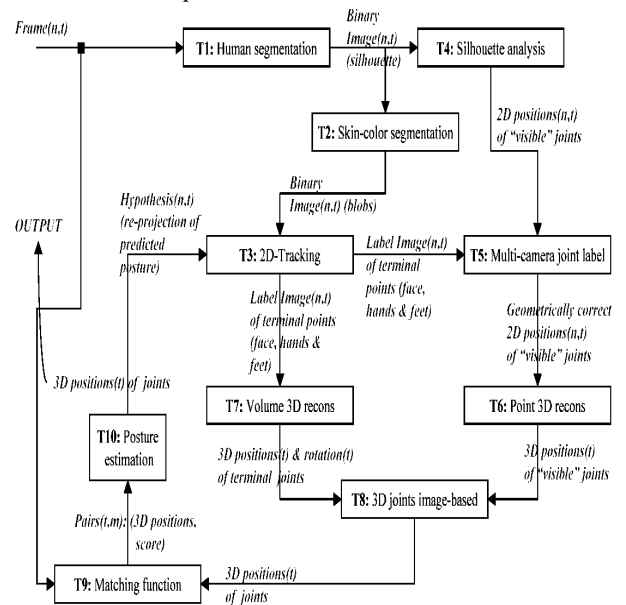


Figure 5. Our proposed system and task framework

To reduce the volumetric space search we detects the occupied volume by the person, computed from the n different cameras, this is done in real time. To compute it we take the following steps:

1. The study volume is divided in equal volume voxels.
2. For each voxel, compute if it is occupied or not.

This process is carried out for each captured frame, and for the total volume, but normally the volume that the person occupies is lower than the total volume, therefore we can decrease computing time if we limit the studied volume to a reduced volume. For this reason, if in the previous frame any voxel has been detected occupied, the algorithm computes the bounding box that contains the occupied volume, and only computes for this subset volume, the bounding box.

This allows us to restrict the studied volume to the zone of interest and modifying and moving the bounding box according to the movements of the subject. To achieve the best results we smooth previously the captured and background image. Smoothing gives as a result the stabilization of the H-component and therefore it divides by two the threshold of such component. For the I-component, we can allocate a high threshold and thus eliminate shadows without erasing parts of interest of the person. Finally, the S-component provides us little useful information. At the moment, no high level information is used in this last presented approach. We are working to combine the process presented in the first part of this section with the last segmentation and voxel occupancy criteria.

In Figure 6 we display some matching results. In this case, the matching is done semi-automatically and without calibration, so the results can include some errors. Besides, the end effectors are not considered (hand, etc.).

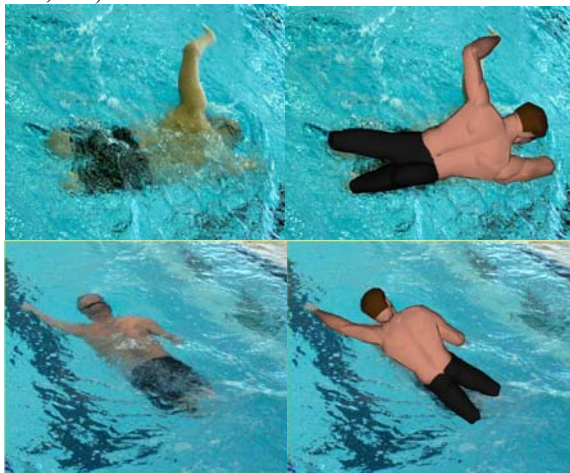


Figure 6. Matching examples with non-calibrated cameras.

5. Biomechanical parameters studied

To study a specific swimming human model we need a set of well-defined parameters. The following list includes the most important ones. The final aim of this work is to obtain all these parameters in an automatic

way, but at the moment only a subset are reached with computer vision segmentation and analysis techniques:

- Defining the skew of upper limbs in order to make a difference in the biomechanics performance between swimmers.

- Calculating and observing the movement of the gravity center

- Defining the anthropometrics parameters of several segments: a) Length of both lower limbs, b) Length of both upper limbs, c) Total length of each half-body, d) Distance between both acromioclavicular articulations, e) Distance from each lower limbs to the vertex, f) Length of trunk: distance between the suprasternal point and the middle point between both hips, g) Length of the three segments of the left upper limb: g.1) From the acromioclavicular to the olecranon, g.2) From olecranon to the styloid apophysis and g.3) From the styloid apophysis to the distal phalanx of the finger

At the moment, all these parameter are computed in a manual or semiautomatic way. Only a subset is computed automatically (for example: center mass point). In Figure 6 we can see a HIS segmentation and centroid computation. At left we can see the original image, next the segmented image and the centroid marked as a dot, finally we can found the H and S components.

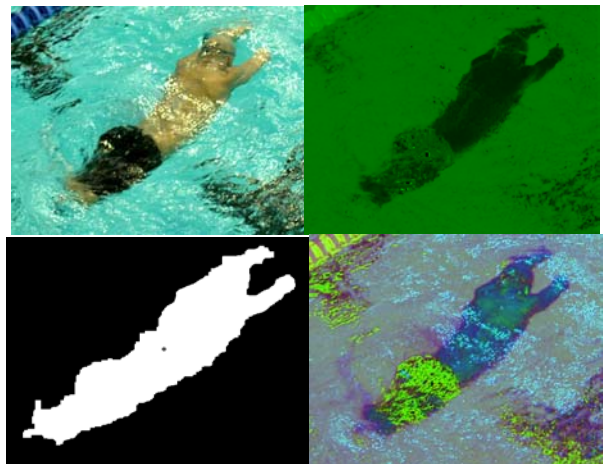


Figure 7. Colour segmentation HSL and centroid computation.

6. Conclusions

In this paper we have presented a whole system to analyzes and synthesise human movements. The main outstanding research and contributions are: a) Digital standard and portable low cost capturing system, b) Biomechanical compliant human modeling, c) Robust colour segmentation process using solid mathematical background theory, d) Semiautomatic matching process

based on rules from images features and biomechanical conditions e) High level accuracy application in real sport activities and disabled people.

In general, the final system should allow us to detect, follow, and recognize the activities of one or more people in a scene [4]. The part presented allows editing the humanoid only under supervision, but a possible extension would be obtaining cinematic and dynamic means with a view to more sophisticated applications. In fact, a determined level of precision will be defined according to the application.

This work is framed within a more general project of analysis and synthesis of human movement using techniques of digital image processing and computer vision. Is very important to remark that the system proposed is non invasive and does not use markers on people. The biomechanical model is also overlapped on real images and we do not need individual manual digitalization on every frame and an interactive feedback with the model is possible in real time processing.

The last part presents the adaptation of the proposed system to the analysis of sports motion in all the possible variations, in particular for people with several physical limitations. Obviously, each discipline has its parameters of interest, which will be defined by the user or specialized technician. The precision of the system is adequate for indoor sequences. We have problems with water interferences in the segmentation process and we plan to include underwater images analysis in near future. We are working to improve the segmentation process, and reach the biomechanical data in a full automatic way. The manual interaction must be avoided to reach more precision in human matching. We are working with filtering estimation using predictive methods and dynamic and kinetic control of biomechanical solution proposed and we plan to define a pose evaluation function (PEF) that combine region, contour and texture information.

7. Acknowledges

This work is subsidized by the European project HUMODAN IST-2001-32202 and Spanish Contract TIC2003-0931. Also we would like to express ours thanks to all members of Computer Graphics and Vision group by their explicit and implicit support to develop this work. We also want to make a special mention to Mr. Xavi Torres, a para-olimpic swimmer who has collaborate with us in the experiments (<http://www.xavitorres.com/>).

8. Bibliography

- [1] J.M. Morel and S. Solimini. "Variational Methods for Image Segmentation", Birkhauser Verlag. 1995
- [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, 1997, pp 780-785.
- [4] J.M. Buades, M. Gonzalez, F.J. Perales. "Face and Hands Segmentation in Color Images and Initial Matching" International Workshop on Computer Vision and Image Analysis. Palmas de Gran Canaria. Dec. 2003. pp 43-48
- [5] Jessica K.Hodgins, Nancy S. Pollard. "Adapting Simulated Behaviors For New Characters". Computer Graphics Proceedings, 1997 pp. 153-162
- [6] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and variational problems", Communications on Pure and Applied Mathematics, XLII(4), 1989
- [7] H.D. Cheng, X.H. Jiang, Y. Sun, JinGli Wang "Color Image Segmentation: Advances and Prospects", Journal of Pattern Recognition 34, (2001), pp. 2259-2281
- [8] G. Koepfler, J.M. Morel, and S. Solimini, "Segmentation by minimizing a functional and the merging methods", SIAM J. on Numerical Analysis, Vol 31, No 1, Feb. 1994