Real-Time Marker-free Motion Capture from multiple cameras

Brice Michoud, Erwan Guillou, Hector Briceño and Saïda Bouakaz
LIRIS - UMR 5205 CNRS
University of Lyon, France
{bmichoud, eguillou, hbriceno, sbouakaz}@liris.cnrs.fr

Abstract

We present a fully-automated method for real-time and marker-free 3D human motion capture. The system computes the 3D shape of the person filmed from a synchronized camera set. We obtain a robust and real-time system by using both a fast 3D shape analysis and a skin segmentation algorithm for human tracking. A skeleton-based approach facilitates the shape analysis. We are able to track fast and complex human motion in very difficult cases, like self-occlusion. Results on long video sequences with rapid and complex movements, demonstrate our approach robustness.

1. Introduction

Marker-free human body motion capture is a promising technique developed in computer vision. Its goal is to find the main joints positions of the human body across time. It has a large range of applications, from movie special effects to human machine interaction systems like next-generation video game consoles. Human motion tracking is a difficult problem because of the complexity of human body kinematics.

Our goal is to provide motion capture for home applications. As our system targets the general public, it has to be user-friendly, fully automated, markerless and inexpensive. Because of the interaction constraint, the system needs to work in real-time (at least 30 frames processed per second). To provide a user friendly system, it should work without markers (active or passive) and without any particular sensors. Indeed, they are generally invasive and difficult to place correctly by non-professional people. In this paper, we propose a fully automated system, which provides motion capture data from a set of calibrated cameras under real-time constraints.

Several methods have been proposed for acquiring three-dimensional human motion. Broadly speaking, the methods proposed can be divided into different categories depending on the number of cameras or the data processed. The first set of methods work with a single camera [1, 6]. The results are generally ambiguous, particularly when the method is based on the object's silhouette. These methods are subject to be stuck in local minima as different positions can yield the same silhouette. To circumvent this, other methods use multiple cameras. Some of them work only on a 3D human shape estimation analysis [5, 12, 17]. These techniques provide good results when the 3D shape topology correspond to the filmed human topology. In other words each body parts have to be clearly identifiable in the 3D shape estimation. In self-occlusion cases or when there are large contacts between limbs and body, it's very difficult to make a distinction between rigid body parts and reconstruction imprecision. For this reason some teams, like Caillette *et al.* [2], have proposed methods based on shape and color analysis. They link blobs to a kinematic model in order to reliably track individual body parts with both volume and color information. This technique requires a contrasted clothing between each body parts for tracking. Thus it adds a usability constraint.

We propose a motion-tracking algorithm that uses skin color segmentation to guide body parts labeling. The skin colored visible parts correspond to the undressed body parts, such as the face, the hands and possibly the legs. The skin segmentation allows us to compute the subsets of the 3D shape that contain skin parts. It allows robust tracking of challenging human motions in real-time. Few methods provide real-time motion capture. Some of them run only with interactive frame rate (10 fps [2]). Our method runs at 30 fps, based on simple heuristics driven by shape topology analysis, skin segmentation and temporal coherence. This provides acceptable latency time for human machine interactions.

After an overview of the system, we will introduce the two main stages of the method: the automatic initialization pose and the body parts motion-tracking. Then, we will discuss on the results obtained from real and complex data, which are failure cases for most methods. Finally, we will conclude about our contributions, and we will present some perspectives for this work.

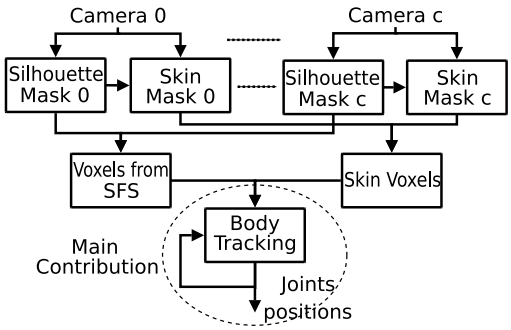

Figure 1. System overview: our main contribution is fast and robust motion capture system.

2. System Overview

The Figure 1 outlines the system structure. As a preprocess, the cameras are fully calibrated using Zhang method [18]. We aim at achieving real-time, hence fast background subtraction [4] and skin segmentation [15] methods are used.

At each frame, these images are computed and used as inputs for a Shape-From-Silhouette based method [3, 9, 11] to compute the 3D shape estimation of the whole model and the skinned parts. This results in two sets of voxels: $\mathcal{V}_{\overline{\text{skin}}}$ and $\mathcal{V}_{\text{skin}}$. We refer to their union as $\mathcal{V}_{\text{all}}$.

General algorithm idea : The goal is to classify each voxel as belonging to one of the body parts. As we try to resolve the location of body parts other than limbs, we start from an active set of voxels $\mathcal{V}_{\text{act}}$. Using heuristics, we determine at least one voxel that belongs to the body part $\mathcal{V}_{\text{part}}$ and try to fit a shape (*e.g.* a cylinder or a sphere) to the related voxels. While a criteria has not been met, we iteratively add voxels to $\mathcal{V}_{\text{part}}$ and remove them from $\mathcal{V}_{\text{act}}$. This decreases the computation time for the resolution of the next parts.

Joints labeling is presented in Figure 2. Our system is based on simple and fast heuristics. Less accurate than the registration-based methods, this approach provides a real-time system. Robustness is increased by using a multi-modal scheme composed of shape analysis, skin-colored segmentation, temporal coherence and human anthropometric constraints. For each body part to track, we estimate a subset of solutions by both the registration of simple geometric objects on the 3D shape and the registration driven by skin-colored parts. If there are some ambiguities, then we use temporal coherence and human anthropometric constraints to find the right solution. To speed up the process and increase its robustness, we remove the voxels used for each body part recognition. $\mathcal{V}_{\text{act}}$ denotes the set of voxels considered for the resolution of the remaining body parts.

Our system has two steps: initialization and tracking.

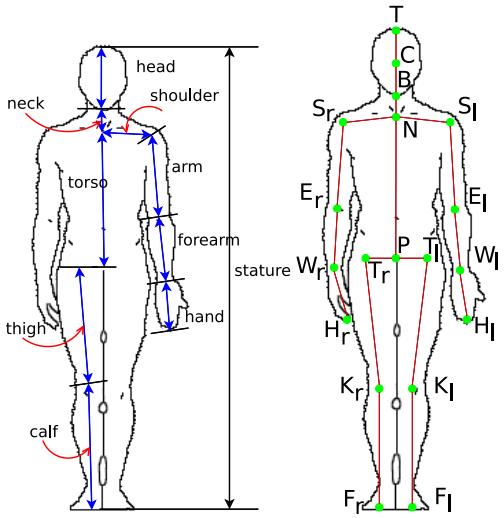

Figure 2. Left: body parts labeling, right: joint naming.

Both use the same algorithm with different initial conditions. The initialization step estimates anthropometric values and initial pose, then, using this information, the second step tracks joint positions over time. The current frame pose is estimated and its computation is facilitated by the current skin voxels, the current 3D shape estimation, and the joints position estimated at the previous frame. To ensure the robustness of the method, we suppose that both hands and person's face are partially uncovered. In the same way, we suppose that the torso is dressed, and that the clothing has a non-skin color.

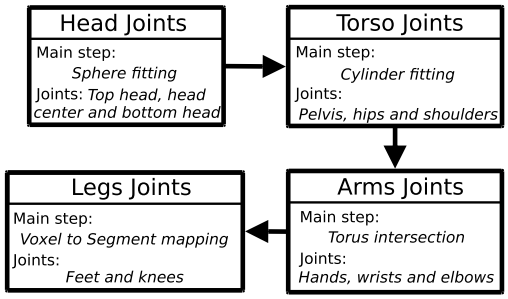

Figure 3. Our method structure: head is estimated first, then the torso, finishing with limbs (legs and arms).

The algorithm works in 4 stages (Figure 3) and will be described in the next two sections.

For reading convenience, we use the following notations:
L_{x} denotes the length of body part x (see Figure 2),
D_{x} its orientation
R_{x} its radius (of sphere or cylinder, *e.g.* head, forearm)
joint notation used is given in Figure 2 (right side).
J^n denotes the value of a quantity J (joint position, voxel set...) at frame n

J^0 its initial value
when dealing with sided joints (like knee or wrist), indices l and r denote respectively left and right side
$\mathcal{V}_x$ denotes a set of voxel and
$\mathcal{E}_{\mathcal{V}_x}$ its inertia ellipsoid
when dealing with iterative algorithms
$J(i)$ denotes the J quantity value at step i, $J(0)$ its initial value and $J(k)$ its final value.

3. Body Parts Initialization

We present in this section our techniques to estimate the anthropometric measures and the initial body pose. Our method is fully automated, based on anthropometric ratios and shape and color analysis techniques. The literature in connection with this step can be classified in three categories. In the first one [10], the dimensions and initial pose are manually specified. Second-category methods need that the person filmed takes an initialization pose like T-pose [8]. These methods are generally real-time friendly. The third class is composed of the fully automated methods [12] which are non real-time processes. Our approach is real-time and fully automated for any kind of movements as long as the person filmed is standing up, hands below the level of head, and feet not joined.

Our method computes each body part parameters sequentially. The order is presented in Figure 3. This process will be repeated until convergence or a timeout. This initial pose will be kept only from the last processed frame of initialization.

3.1. Anthropometric Measures

Several studies that include the anthropometric data [7, 13, 14] are used to develop ratio estimations. Statistical analysis is performed, including fitting to normal distribution. We propose simplified anthropometric ratios, whose accuracy is sufficient for human-machine interactions. Let L_{stat} be the acquired human body length, estimated as the maximum distance from foreground voxels to floor plane. To increase robustness, the element of maximum altitude is taken among voxels in the major connex component of foreground voxels

Hence, knowing L_{stat}, guesses for anthropometric measures are given by these ratios:

$$
\begin{array}{ll}
L_{\text{arm}} \approx L_{\text{stat}}/6 & L_{\text{farm}} \approx L_{\text{stat}}/6 \\
L_{\text{hand}} \approx L_{\text{stat}}/10 & R_{\text{head}} \approx L_{\text{stat}}/16 \\
L_{\text{torso}} \approx 3L_{\text{stat}}/8 & L_{\text{neck}} \approx L_{\text{torso}}/10 \\
L_{\text{thigh}} \approx L_{\text{stat}}/4 & L_{\text{calf}} \approx L_{\text{stat}}/4 \\
L_{\text{shld}} \approx L_{\text{stat}}/8 &
\end{array}
$$

The initialization process works on unused voxels $\mathcal{V}_{\text{act}}$. At the beginning , this set of voxels corresponds to all voxels $\mathcal{V}_{\text{all}}$ and is updated at each step by removing voxels used to estimate body parts.

3.2. Head Initialization

This step aims at finding T^0 and B^0, the positions of the top of the head and the connection point between head and neck at frame 0. From our hypothesis, the face's voxels (further-noted $\mathcal{V}^0_{\text{face}}$) of the subject acquired define the topmost connex component among $\mathcal{V}^0_{\text{skin}}$. Head position C^0 (position of the head center) is computed by fitting a sphere $S(i)$ in $\mathcal{V}^0_{\text{act}}$ (see figure 4). $S(i)$ is defined by its center $\text{C}^0(i)$ and radius R_{head}.

Head fitting algorithm : $\text{C}^0(0)$ is initialized as the centroid of $\mathcal{V}^0_{\text{face}}$.

At step i of the algorithm, $\text{C}^0(i)$ is the centroid of the set $\mathcal{V}^0_{\text{head}}(i)$ of unused voxels that lie in a sphere $\mathcal{S}(i-1)$ defined by its center $\text{C}^0(i-1)$ and its radius R_{head}.

The algorithm iterates until step k when the position of C^0 stabilizes, *i.e.* the distance between $\text{C}^0(k-1)$ and $\text{C}^0(k)$ falls below a threshold ϵ_{head}.

Head joints estimation : Knowing C^0 position, B^0 (respectively T^0) is computed as the lowest (resp. upper) intersection between $\mathcal{S}(k)$ and the principal axis of $\mathcal{E}_{\mathcal{V}^0_{\text{head}}}$. The back-to-front direction D^0_{b2f} is defined as the direction

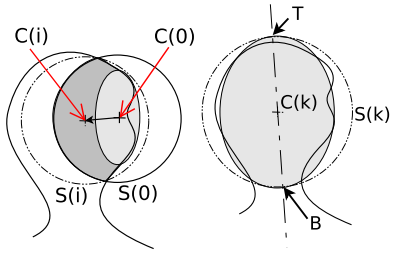

Figure 4. Left: sphere fitting (light gray denotes $\mathcal{V}^0_{\text{face}}$, dark gray denotes $\mathcal{V}^0_{\text{head}}(i)$), right: joints estimation.

from C^0 towards the centrod of $\mathcal{V}^0_{\text{face}}$ (note that voxels from the back of the head are not in $\mathcal{V}_{\text{skin}}$). At this point, we remove from $\mathcal{V}^0_{\text{act}}$ the set of elements that belongs to $\mathcal{V}^0_{\text{head}}$.

3.3. Torso Initialization

Let $\mathcal{V}^0_{\text{torso}}$ be the set of voxels that describes the torso, they are initialized using unused voxels $\mathcal{V}^0_{\text{act}}$. At step i, the algorithm estimates D^0_{torso} by fitting a generic cylinder $\mathcal{CYL}(i-1)$ in $\mathcal{V}^0_{\text{torso}}$(i) (see left figure 5). $\mathcal{CYL}(i)$ is defined by the center of one of its cap B^0, radius R_{torso}, length L_{torso} and orientation D^0_{torso}(i).

Torso fitting algorithm : $\mathcal{V}^0_{\text{torso}}(0)$ is initialized with $\mathcal{V}^0_{\text{act}}$, R_{torso} with L_{shld} and $D^0_{\text{torso}}(0)$ as the vector from N^0 toward the centrod of $\mathcal{E}_{\mathcal{V}^0_{\text{act}}(0)}$.

At step i, $\mathcal{V}^0_{\text{torso}}$(i) is computed as the set of elements from $\mathcal{V}^0_{\text{torso}}$(i-1) that lies in $\mathcal{CYL}(i-1)$. D^0_{torso}(i) will then

be the principal axis of $\mathcal{E}_{\mathcal{V}^0_{\mathrm{torso}}(i)}$

The algorithm iterates until step k when the distance between the center of $\mathcal{CYL}(k)$ and the centroid of $\mathcal{V}^0_{\mathrm{torso}}$(k) falls below a threshold $\epsilon_{\mathrm{torso}}$.

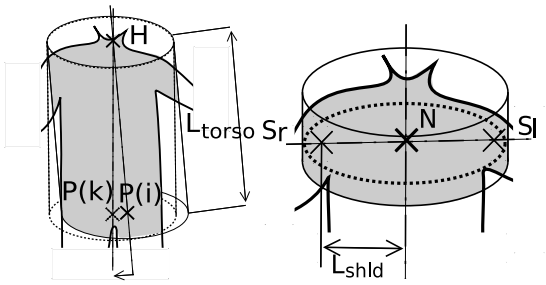

Figure 5. left: torso fitting, right: shoulders and thigh.

Thigh joints estimation : Let $\mathcal{V}^0_{\mathrm{hips}}$ be the set of voxels from $\mathcal{V}^0_{\mathrm{torso}}$(k) intersected by the lower cap of cylinder $\mathcal{CYL}(k)$ and v the principal axis of its inertia ellipsoid. The initial pelvis position P^0 is defined as the centroid of $\mathcal{E}_{\mathcal{V}^0_{\mathrm{hips}}}$. The torso's radius R_{torso} is updated with $|v|/2$ and left/right thigh joints are defined as :

$$\mathrm{T}^0{}_x = \mathrm{P}^0 \pm vR_{\mathrm{torso}}/2 \tag{1}$$

Shoulder joints estimation : Knowing the torso orientation, the neck position is defined as:

$$\mathrm{N}^0 = \mathrm{B}^0 + L_{\mathrm{neck}} D^0_{\mathrm{torso}}(k) \tag{2}$$

Let $\mathcal{V}^0_{\mathrm{shlds}}$ be the set of voxels from $\mathcal{V}^0_{\mathrm{torso}}$(k) that lies in cylinder $\mathcal{CYL}$ defined by its center N^0, axis D^0_{torso}(k), radius L_{shld} and length $2L_{\mathrm{neck}}$ (see right figure 5). v is the principal axis of its $\mathcal{E}_{\mathcal{V}^0_{\mathrm{shlds}}}$, left and right shoulder joints are defined as:

$$\mathrm{S}^0{}_x = \mathrm{N}^0 \pm vL_{\mathrm{shld}} \tag{3}$$

Global body orientation : The top-down orientation D^0_{t2d} of the subject acquired is given by $\mathrm{P}^0 - \mathrm{B}^0$. D_{b2f} was computed in 3.2. The left-to-right orientation D^0_{l2r} of the subject acquired is given by $D^0_{\mathrm{l2r}} = D^0_{\mathrm{t2d}} \times D^0_{\mathrm{b2f}}$.

Right and left side of shoulders and hips in equations 1 and 3 are determined by the orientation of v in respect to D^0_{l2r}. These vectors will help to differentiate the left from the right and the front from the back.

$\mathcal{V}^0_{\mathrm{act}}$ is updated by removing its elements that belong to $\mathcal{V}^0_{\mathrm{torso}}$.

3.4. Arms Initialisation

Let $\mathcal{V}^0_{\mathrm{hand}} = \mathcal{V}^0_{\mathrm{skin}} \bigcap \mathcal{V}^0_{\mathrm{act}}$ be the set of candidate voxels for hands. A maximum value for the length of an arm is given by $L_{\mathrm{stat}}/2$ (see 3.1). Hence left and right hands will be described by the two major connex components (noted $\mathcal{V}^0_{\mathrm{hand}0}$ and $\mathcal{V}^0_{\mathrm{hand}1}$) among the elements of $\mathcal{V}^0_{\mathrm{hand}}$ that lie within a sphere defined by its center N^0 and its radius $L_{\mathrm{stat}}/2$. If only one connex component is found, we assume that hands are joined and that this connex component describes both hands: $\mathcal{V}^0_{\mathrm{hand}0} = \mathcal{V}^0_{\mathrm{hand}1}$.

Hand joints estimation : From our hypothesis, the voxel set $\mathcal{V}^0_{\mathrm{hand}j}$ contains wrist joint and fingers extremity. From voxel set $\mathcal{V}^0_{\mathrm{hand}j}$ ($j \in [0,1]$), P_{j0} and P_{j1} are the two extremities of the principal axis of $\mathcal{E}_{\mathcal{V}^0_{\mathrm{hand}j}}$. We compute v_{j0} (resp. v_{j1}) as the volume described by the intersection between unused voxels ($\mathcal{V}^0_{\mathrm{act}}$) and a sphere defined by its center P_{j0} and its radius $L_{\mathrm{hand}}/2$.

The wrist being the connection point between forearm and hand, then if $v_{j0} \gg v_{j1}$, P_{j0} describes the wrist. Hence $\mathrm{W}^0{}_j = P_{j0}$ and $\mathrm{H}^0{}_j = P_{j1}$. Otherwise, $\mathrm{W}^0{}_j = P_{j1}$ and $\mathrm{H}^0{}_j = P_{j0}$.

Let us suppose that we know the side of $\mathcal{V}^0_{\mathrm{hand}0}$ and $\mathcal{V}^0_{\mathrm{hand}1}$, we can now estimate arm and hand joints for the left or right side x.

Elbow estimation : Let initialize the radius of the arm at the elbow position R_{farm} with the value of L_{shld}. Human morphology imposes constant lengths for arm L_{arm} and forearm L_{farm}. Then the potential voxels $\mathcal{V}^0_{\mathrm{elb}x}$ for elbow position are estimated as the elements of $\mathcal{V}^0_{\mathrm{act}}$ that lie in a torus defined by its center C_t, rotation axis V_t, tube radius R_{farm} and the distance from center to tube R_t (see Figure 6):

$$V_t = \mathrm{S}^0{}_x - \mathrm{W}^0{}_x \tag{4}$$

$$C_t = V_t \frac{L_{\mathrm{farm}}}{L_{\mathrm{farm}} + L_{\mathrm{arm}}} + \mathrm{W}^0{}_x \tag{5}$$

R_t is the altitude in $\mathrm{E}^0{}_x$ of the triangle defined by $\mathrm{S}^0{}_x$, $\mathrm{W}^0{}_x$ and $\mathrm{E}^0{}_x$. The last point is unknown but side's lengths are known so we can compute R_t. Hence, elbow position

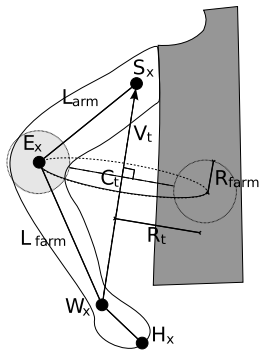

Figure 6. Elbow estimation, dark gray elements are already removed from $\mathcal{V}_{\mathrm{act}}$.

$\mathrm{E}^0{}_x$ is the centroid C_x of the connex component V_x of $\mathcal{V}^0_{\mathrm{elb}x}$ that maximizes the number of elements of $\mathcal{V}^0_{\mathrm{act}}$ intersected by segments $[\mathrm{W}^0{}_x, C_x]$ and $[C_x, \mathrm{S}^0{}_x]$. R_{farm} is then updated with the norm of the smallest axis of the $\mathcal{E}_{V_x}$.

Hand side : When the link between a set of voxel $\mathcal{V}^0_{\text{hand}\,j}$ (j stands for 0 or 1) and the shoulders is unknown, the elbow estimation is applied for both shoulders. **This is the key to resolve challenging poses of the arms.** Hence using $\mathcal{V}^0_{\text{hand}0}$, we have two potential positions for associated elbow: ${\text{E}^0}_{0l}$ computed with ${\text{S}^0}_l$ and ${\text{E}^0}_{0r}$ computed with ${\text{S}^0}_r$. $\mathcal{V}^0_{\text{hand}0}$ is then associated with the side x (x stands for l or r) that maximizes the number of elements of $\mathcal{V}^0_{\text{act}}$ intersected by segments $[{\text{W}^0}_0, {\text{E}^0}_{0x}]$ and $[{\text{E}^0}_{0x}, {\text{S}^0}_x]$. At this point, $\mathcal{V}^0_{\text{hand}1}$ is associated to the other side and elbow estimation is performed using the corresponding shoulder.

For each arm, $\mathcal{V}^0_{\text{act}}$ is updated, removing elements that lie inside the cylinders having $R_{\text{hand}}/2$ as radius and $[\text{H}^0, \text{W}^0]$, $[\text{W}^0, \text{E}^0]$ and $[\text{E}^0, \text{S}^0]$ as axis.

3.5. Legs Initialization

Note that at this point of the method, $\mathcal{V}^0_{\text{act}}$ contains the voxels that haven't been used for any other parts of the body.

Foot initialization : First we compute the set of connex components from elements of $\mathcal{V}^0_{\text{act}}$ having their height below $L_{\text{stat}}/8$. If there is less than 2 connex components, we assume that feet are joined and can't be distinguished. Otherwise, as for hands initialization, we use the two major connex components $\mathcal{V}^0_{\text{foot}\,l}$ and $\mathcal{V}^0_{\text{foot}\,r}$. Left and right assignation of voxel's set is done using D_{l2r} vector.

For the left/right side x, let v_x be the vector from ${\text{T}^0}_x$ to the centroid of $\mathcal{V}^0_{\text{foot}\,x}$. Knee and Foot joints are guesses using the following equations:

$$k_x = {\text{T}^0}_x + v_x \frac{L_{\text{thigh}}}{|v_x|} \tag{6}$$

$$f_x = {\text{T}^0}_x + v_x \frac{L_{\text{thigh}} + L_{\text{calf}}}{|v_x|} \tag{7}$$

Leg binding : At this point, all body parts but the legs have been estimated, hence $\mathcal{V}^0_{\text{act}}$ contains only the legs voxels. Our leg joints extraction is inspired from "point to line mapping" process used to bind an animation skeleton on a 3D mesh [16]. The elements of $\mathcal{V}^0_{\text{act}}$ are splitted into four sets $\mathcal{V}^0_{\text{thigh}_l}$, $\mathcal{V}^0_{\text{calf}\,l}$, $\mathcal{V}^0_{\text{thigh}_r}$ and $\mathcal{V}^0_{\text{calf}\,r}$ depending of their euclidean distance to segments $[\text{T}^0_l, k_l]$, $[k_l, f_l]$, $[\text{T}^0_r, k_r]$ and $[k_r, f_r]$ (see figure 7 left).

Leg joints : For the left/right side x, we compute the inertia ellipsoid $\mathcal{E}_{\mathcal{V}^0_{\text{calf}\,x}}$ and P_0 and P_1 its extrema points. The knee is at the intersection of thigh and calf, hence foot position ${\text{F}^0}_x$ is given by the extrema point the most distant from the inertia ellipsoid of $\mathcal{V}^0_{\text{thigh}_x}$ (let say it's P_1). The knee position will then be given by the following equation:

$${\text{K}^0}_x = L_{\text{calf}} \frac{P_0 - {\text{F}^0}_x}{|P_0 - {\text{F}^0}_x|} + {\text{F}^0}_x \tag{8}$$

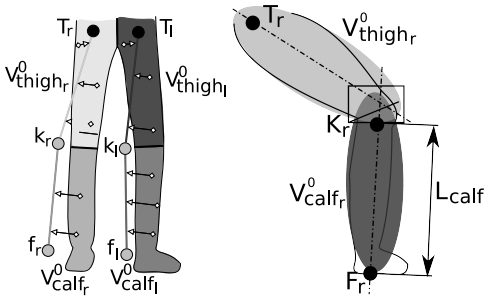

Figure 7. Legs initialization; left: voxel binding, right: joints estimation.

4. Body Parts Tracking

Using anthropometric measures initialization, the previous body pose and the labeled 3D shape estimation, we track the human body parts in real-time. The computation is made with the initialization steps ordering (Figure 3). First it estimates head joints. Next, we track torso joints. Finally we compute the limbs joints.

Head Tracking : Let $\mathcal{V}^n_{\text{face}}$ be the nearest connex component of $\mathcal{V}^n_{\text{skin}}$ from $\text{C}^{n-1} + R_{\text{head}} D^{n-1}_{\text{b2f}}$. The head fitting algorithm (section 3.2) is then applied using $\mathcal{V}^n_{\text{face}}$. Head joints estimation is then performed computing B^n, T^n and D^n_{b2f}.

Torso Tracking : The torso fitting algorithm (section 3.3) is applied using $\mathcal{V}^n_{\text{act}}$ as initial value for $\mathcal{V}^n_{\text{torso}}(0)$ and the vector from B^n to P^{n-1} as initial value for $D^n_{\text{torso}}(0)$. Thigh and shoulder joints estimation is then performed computing P^n, T^n_l, T^n_r, N^n, S^n_l, S^n_r, D^n_{l2r} and D^n_{t2d}.

Arms Tracking : Using spatial coherence property, each current hand position is described by the nearest skin connex component from last frame position H^{n-1}. If no connex component is found we use previous computed motion as an estimate for current frame motion, hence:

$$\text{H}^n = \text{H}^{n-1} + Motion^{n-1} \tag{9}$$

Once the new hand's positions are known, the elbow position is computed as described in 3.4.

Legs Tracking : For leg tracking, the binding (section 3.5) is performed using the knee and foot positions at previous frame using segments $[\text{T}^n_l, \text{K}^{n-1}_l]$, $[\text{K}^{n-1}_l, \text{F}^{n-1}_l]$, $[\text{T}^n_r, \text{K}^{n-1}_r]$ and $[\text{K}^{n-1}_r, \text{F}^{n-1}_r]$. Leg joints are then computed as for leg initialization.

5. Results

The acquisition infrastructure is composed of four calibrated cameras, each connected to a dedicated computer. To avoid network overload, background subtraction and skin segmentation produce down-sampled images at 30 frames per second at a resolution of 320×240 pixels. Results are

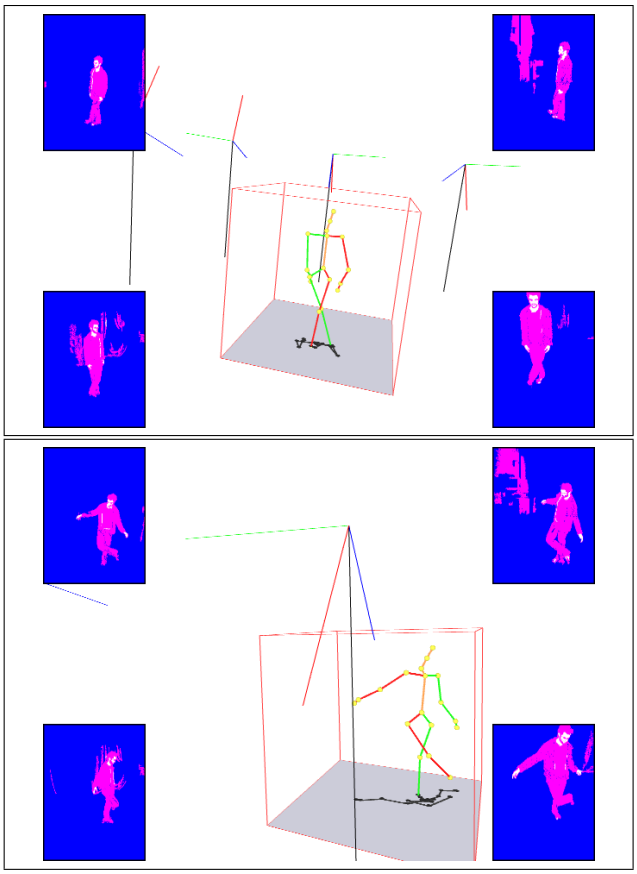

Figure 8. Results for challenging poses.

then transmitted to another computer which computes voxels and estimates joint positions.

Our method has been applied to different persons doing fast and challenging motions. In all pictures, right side skeleton's bones a represented in red, the left bones are in green and the others in orange. In the corresponding input images, blue parts represent background, red parts correspond to silhouettes, and white zones to skin parts.

Thanks to shape analysis and skin parts knowledge too, our system is able to acquire the joints position for a challenging pose outlined on the top image of the Figure 8. This pose is difficult because the 3D shape topology is not a human corresponding one.

The temporal coherence is the success key for the pose presented int the bottom picture of the Figure 8.The 3D Shape topology is a human one, but ambiguous as for the feet.

The images presented Figure 9 prove that our system works for the acquisition of a large range of motions. They also demonstrate our method robustness on noised input images.

In few cases, knees articulations positions are not coherent with the human kinematics chain constraints (see Figure 10), because they depend on the 3D shape estimation quality. We are currently working on adding kinematics constraint to our model. But we have to make it without increasing computation time.

As our algorithm is based on 3D shape analysis, it is independent of the number of cameras used, but it depends on the voxel grid resolution. Our current experimental implementation computes motion capture at 30 fps from a voxel grid composed by 64^3 voxels in a $6m^3$ box. This resolution is sufficient for human machine interfaces. For more precise acquisition we have to use better resolution. To conserve real-time computation, our method can be used only from sets of surface voxels.

Our motion capture system is based on a Shape-From-Silhouette algorithm. This algorithm computes an object 3D shape estimation from its silhouettes. The result directly depends on the silhouette segmentation quality, which is always an opened problem of the computer vision science. If the silhouette mask contains some noises like camera noise or object shadows, the volume reconstruction will be very noised. Thus the results of the motion capture will be worse. But our method is also based on a skin segmentation which is a more robust faced to camera noise. Then the hand and head articulations are more noise-resistant, than others articulations.

The segmentation we have selected is based on a skin-colored stochastic learning from colored image set. It is important to make the leaning process on a big data set, with different kind of skin sample. If the skin sampling is biased then the system will provide worse results, especially when the skin color of the person filmed is not learned.

6. Conclusion

In this paper, we describe a new marker-free human motion capture system from a camera set. Fully automated and working under real-time constraint, the system is based on a 3D shape analysis, human morphology constraints, and a 3D shape skin segmentation. Combining different 3D information, the approach is robust to self-occlusion and poor 3D shape approximation provided by voxel estimation subsystem.

We are able to achieve this by classifying voxels in skin and non skin voxels, and carefully and orderly assigning the voxels to body parts to determine and disambiguate joint positions.

The current system provides real-time motion capture for only one person. Current work aims at providing motion capture of multiple persons filmed in the same acquisition area, even when they are in contact.

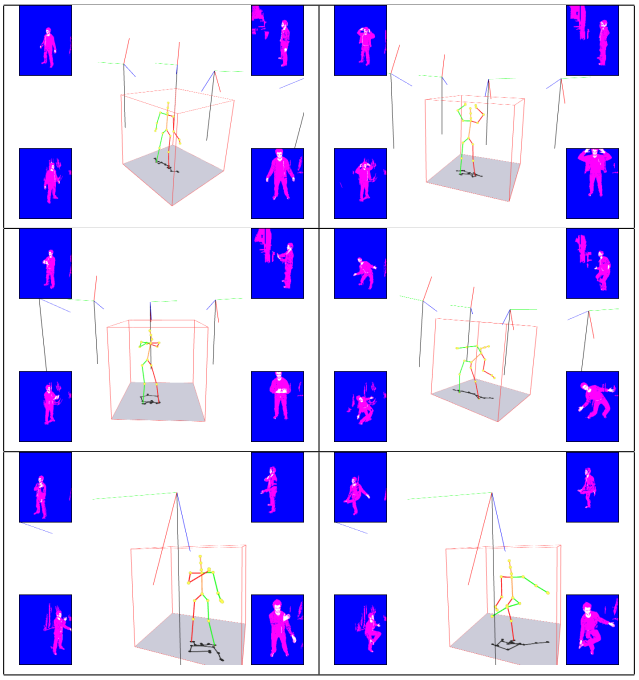

Figure 9. Results for a wide range of movements.

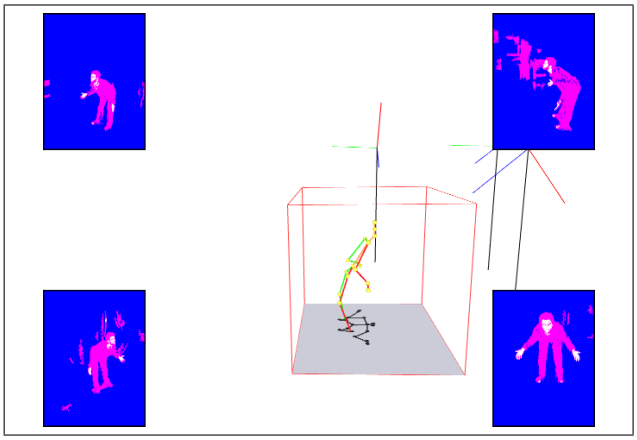

Figure 10. The knees estimations are not coherent with the human kinematics constraints.

References

[1] A. Agarwal and B. Triggs. Monocular human motion capture with a mixture of regressors. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, page 72, Washington, DC, USA, 2005. IEEE Computer Society.

[2] F. Caillette, A. Galata, and T. Howard. Real-Time 3-D Human Body Tracking using Variable Length Markov Models. In *Proceedings of British Machine Vision Conference (BMVC)*, volume 1, pages 469–478, September 2005.

[3] K. M. Cheung, T. Kanade, J.-Y. Bouguet, and M. Holler. A real time system for robust 3d voxel reconstruction of human motions. In *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '00)*, volume 2, pages 714 – 720, June 2000.

[4] D. Conte, P. Foggia, M. Petretta, F. Tufano, and M. Vento. Evaluation and improvements of a real-time background subtraction method. In M. S. Kamel and A. C. Campilho, editors, *ICIAR*, volume 3656 of *Lecture Notes in Computer Science*, pages 1234–1241. Springer, 2005.

[5] E. de Aguiar, C. Theobalt, M. Magnor, H. Theisel, and H.-P. Seidel. M^3: Marker-free model reconstruction and motion tracking from 3d voxel data. *pg*, 00:101–110, 2004.

[6] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *Int. J. Comput. Vision*, 61(2):185–205, 2005.

[7] H. Dreyfuss and A. R. Tilley. *The Measure of Man and Woman: Human Factors in Design*. John Wiley & Sons, 2001.

[8] P. Fua, A. Gruen, N. D'Apuzzo, and R. Plankers. Markerless Full Body Shape and Motion Capture from Video Sequences. In *Symposium on Close Range Imaging, International Society for Photogrammetry and Remote Sensing, Corfu, Greece*, 2002.

[9] J.-M. Hasenfratz, M. Lapierre, and F. Sillion. A real-time system for full body interaction with virtual worlds. *Eurographics Symposium on Virtual Environments*, pages 147–156, 2004.

[10] C. Ménier, E. Boyer, and B. Raffin. 3d skeleton-based body pose recovery. In *Proceedings of the 3rd International Symposium on 3D Data Processing, Visualization and Transmission, Chapel Hill (USA)*, june 2006.

[11] B. Michoud, E. Guillou, and S. Bouakaz. Shape from silhouette: Towards a solution for partial visibility problem. In *Eurographics 2006*, Eurographics 2006 Short Papers Preceedings, pages 13–16, Sept. 2006.

[12] I. Mikic, M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *Int. J. Comput. Vision*, 53(3):199–223, 2003.

[13] R. Motmans and E. Ceriez. Anthropometry table. Ergonomie RC, Leuven, Belgium, 2005.

[14] none. Anthropometry for designers. In *Anthropometric source book. Volume 1*. NASA Report Number: NASA-RP-1024, S-479-VOL-1, 1978.

[15] M. Soriano, B. Martinkauppi, S. Huovinen, and M. Laaksonen. Skin detection in video under changing illumination conditions. In *ICPR*, pages 1839–1842, 2000.

[16] W. Sun, A. Hilton, R. Smith, and J. Illingworth. Layered animation of captured data. *The Visual Computer*, 17(8):457–474, 2001.

[17] T. Tangkuampien and D. Suter. Human motion de-noising via greedy kernel principal component analysis filtering. In *ICPR (3)*, pages 457–460. IEEE Computer Society, 2006.

[18] Z. Zhang. Flexible camera calibration by viewing a plane from unknown orientations. In *ICCV*, pages 666–673, 1999.