HAL
open science

3D Markerless Human Limb Localization through Robust Energy Minimization

Marco Marcon, Massimiliano Pierobon, Augusto Sarti, Stefano Tubaro

To cite this version:

Marco Marcon, Massimiliano Pierobon, Augusto Sarti, Stefano Tubaro. 3D Markerless Human Limb Localization through Robust Energy Minimization. Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2 2008, Andrea Cavallaro and Hamid Aghajan, Oct 2008, Marseille, France. inria-00326757

HAL Id: inria-00326757

https://inria.hal.science/inria-00326757v1

Submitted on 5 Oct 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

3D Markerless Human Limb Localization through Robust Energy Minimization

Marco Marcon, Massimiliano Pierobon, Augusto Sarti, Stefano Tubaro

Politecnico di Milano - Dipartimento di Elettronica e Informazione

Abstract. Markerless human tracking addresses the problem of estimating human body motion in non-cooperative environments. Computer Vision techniques combined with Pattern Recognition theory serve the purpose of extracting information on human body postures from video-sequences, without the need of wearable markers. Multi-camera systems further enhance this kind of application providing frames from multiple viewpoints. This work tackles the application of multi-camera posture estimation through the use of a multi-camera environment, also known as "smart space". A 3D skeleton structure and geometrical descriptors of human muscles are fitted to the volumetric data to directly recover 3D information. 3D skeleton deformations and bio-mechanical constraints on joint models are used to provide posture information at each frame. The proposed system does not require any pre-initialization phase and automatically adapt the skeleton and the volumetric occupation of each limb to the actor physiognomy independently from the pose. Exhaustive tests were performed to validate our approach.

Gesture tracking and posture estimation require the localization of the human body in the scene and the estimation of each limb position during motion. This process can be considered as a combination of two major components, namely, Model Definition and Pose Estimation. Model definition aims at the moulding of the human body model that best fits to the specific shape and size of the tracked human subject, whereas pose estimation is devoted to the proper model adaptation to the postures and the body joint configurations during the motion evolution [1]. Only a few human motion tracking approaches use images from multiple cameras in order to obtain cues on 3D information. On the other hand, a broad range of applications is based on a single monocular acquisition device,e.g.[2]. Motion tracking solutions can be parsed into three main classes: marker-based, model-based and model-free. Our aim is the realization of a markerless system able to estimate the human posture evolution exploiting visual features such as colors, edges, silhouettes and textures. The output of such a system is a set of angles relating each limb to each other, submitted to appropriate constraints for each joint (articulations differ in possible rotation angles and angular extensions). Various approaches that explicitly model the human body as an assembly of rigid parts can be found in the literature[3].Most of them adopt a two-stage strategy: a bottom-up detector is first applied on the image in order to extract the candidate parts, then, a top-down procedure is used to

properly infer about the joint configuration and to find the assembly that best fits; Dynamic Programming is also frequently used to solve the same problem with greater efficiency; e.g. Lee and Cohen [4] accurately estimate the human 3D pose though the adoption of a data-driven Markov Chain Monte Carlo (MCMC) framework. Image observations of different cues are used to infer possible positions of body joints. They translate these inferences into pose hypotheses though the use of the 'proposal maps' as an efficient way of consolidating the evidence and generating 3D pose candidates during the MCMC search, anyway the need of a proper detection of non-frontal postures requires a good face detection algorithm. Moreover, the computational cost is quite high. In the work of Mori et al. [5], Normalized Cuts algorithms are used to detect salient human body parts. Brute-force approaches are also applied to solve the body parts assembly problem but they need to store templates for all possible configurations, as in an exemplar-based approach. Unfortunately this approach will dramatically increase the computational cost. On the other hand, Ren et al. [6] detect the candidate body parts in a bottom-up fashion through the use of parallel cues, assembled together into body configurations using various pairwise constraints between human body parts. Actually most of the works presented in literature require an initialization phase to define the body geometry of the tracked person and the problem of recovering human body configurations without assuming an *a priori* knowledge of the scale, pose or appearance becomes extremely challenging since it demands the use of all possible sources of information: The self-occlusion is one of the most hard-to-solve problems in a monocular environment. Poppe et al. [7] use a computer vision-based approach to estimate the body pose of a presenter in a meeting environment. They extract a 2D silhouette and match it against a planar projection of a 3D human body model with 16 Degree of Freedom (DoF). They use skin color to locate hands and head. The algorithm is able to cope with low resolution image sequences from a monocular device. The constraints imposed by Poppe are:

1. the person has to face the camera all the time
2. the presenter is supposed to wear short sleeve attire

. In [8] Siddiqui and Medioni describe an efficient system for the detection and tracking of the human limbs. The human model consists of a collection of 2D part models related to the lower and upper limbs, the head and the torso. The individual parts are organized into a tree structure, having the head at its root. The detection of proper visual features is made easier through the assumption that the users wear short sleeve shirts, thus leaving their forearms exposed. The main focus is on the tracking of image regions related to the hands and the forearms rather than on a complete articulated object fitting. Capo et al. [9] present a computer vision algorithm able to automatically model the human body skeleton. The body is decomposed into its main parts through the computation of the curvature of a B-Spline coming out of a parametrization of the human body contour. They rely on the assumption that the user stands in a predefined pose during a calibration/initialization phase. Our aim is the release of most of the

aforementioned constraints estimating body posture from a 3D model obtained from a multi-camera set-up frame-by-frame.

1 The acquisition set-up

The acquisitions of 3D body models with their temporal evolution were made at our laboratories where a controlled environment is monitored using eight synchronized and calibrated cameras. We adopted the *Running Gaussian Average* algorithm, directly derived from the work by Wren et al. [10], in order to provide a segmentation of the actor silhouette from the image background. This solution stems from a statistical description of the foreground and the background obtained from a collection of the previous frames; This approach provided the best trade-off for accuracy and fast frame processing; obviously, for real environments, especially in outdoor scenes, more complex approaches may be required for accurate model definition. Once several silhouettes of the same subject are simultaneously extracted from different viewpoints, it is possible to provide a reconstruction of the actor volume, or at least an approximation of it. Every pixel within the silhouette can be traced back from the image plane to the 3D environment, exploiting the information on the optical center and the viewpoint direction of the corresponding camera. Thus, the set of all these lines that connect the silhouette to the volume is a cone with the silhouette as its basis. The intersection of all cones coming from different viewpoints is a volumetric object named Visual Hull (VH, in the following), and it is supposed to enclose the actor body (examples are given in Fig. 1. The closeness of the VH to the actor volume

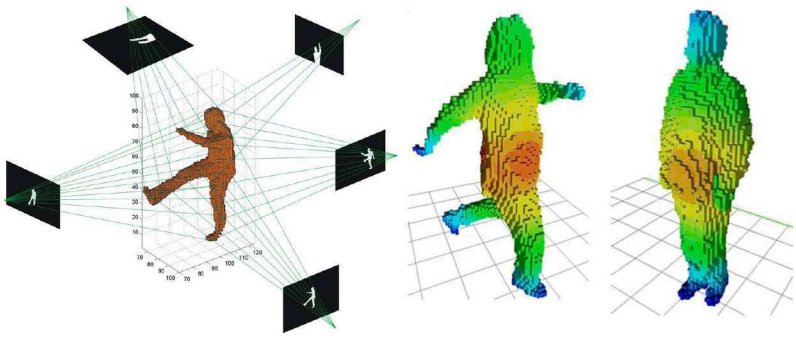

Fig. 1. (*left*)Visual Hull reconstruction from intersection of reprojection cones from silhouettes. (*right*)Two examples of Visual hull reconstruction; different colors relates to different geodesic distances from the model centroid

is strictly related to the position of the cameras and to the shape of the volume itself, i.e. loose clothing may result in an inaccurate reconstruction.

2 Human Model definition

The posture tracking problem requires the definition of a human body model and its proper adaptation within the actor volume. The human model gives to the algorithm the opportunity of exploiting the a priori information about the human body structure and, therefore, the search space related to possible body part configurations can be reduced through the definition of a set of constraints, such as human body proportions, limb connections and their possible relative motion. The skeleton model defined for our experiments comes as simplified representation of the human skeleton, named stick figure, where the joints are modeled as spheres and the bones as cylinders. Moreover, we used a flesh figure that extends the stick figure with truncated cones and ellipsoids, as shown in Fig.2, making the model closer to the real human shape. The aim of our work

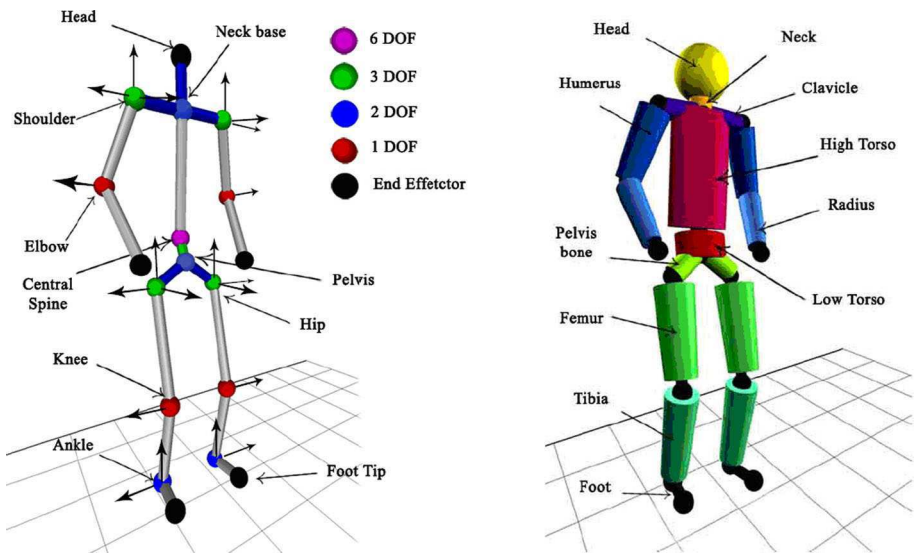

Fig. 2. The stick figure represents a human skeleton, defined by connections and bounds on angles. The image shows the possible axis of rotations and the DOF of rotation (left). The flesh figure extends the stick figure with volumes around bones (right).

stems in the realization of a robust system able to track human body without making any assumption on the initial actor position. In this situation the skeleton size and the definition of volumes for each limb in the flesh model becomes a challenge. We tackled these problem through the application of the following approach. Starting from the head, whose shape can be clearly distinguished from the rest of the body, the other parts detection proceeds with the torso, the shoulders and the hips. Subsequently, the procedure estimates the positions of legs and then arms, which are often partially or completely adjacent to the

torso. The knowledge about the position and the orientation of the other body parts offers some clues about the configuration of arms and legs, thus allowing to cope with adjacency problems.

2.1 Anchor points localization

Most of the body parts are localized starting from a geodesic description of the VH surface: the first step is the determination of center of mass C of the VH,it is usually localized in the middle of the torso and is simply computed with the classical center of mass formula $C_i = \frac{\sum_{x,y,z} iV(x,y,z)}{\sum_{x,y,z} V(x,y,z)}$ where i is x, y or z the sum is over the whole Voxel-Set and the function V is 1 if the voxel is inside the VH and 0 otherwise. The next step is the analysis of the surface curvature. In order to obtain an accurate estimation of it the vectors normal to the surface must be accurately computed: due to the discretization of the voxel-set the surface requires a filtering step before normal computation. The approach we followed is based on 3D Sobel filtering applied on the whole voxel-set: The filter is a 3D matrix (a $3 \times 3 \times 3$ cube) where the planes are defined as:

$$S_1 = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, S_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, S_3 = \begin{bmatrix} -1 & -2 & -1 \\ -2 & -4 & -2 \\ -1 & -2 & -1 \end{bmatrix} \quad (1)$$

The application of this filter gives the amount of variation (first derivative) in a direction orthogonal to the matrix planes. Defining two other 3D matrices obtained by a 90 degrees rotation along two orthogonal axes of the previous matrix (1) is then possible to obtain for each surface voxel the 3 components of a vector normal to the surface itself. The resulting vector is normalized to become a "unit vector": a result is shown in Fig. 3 .

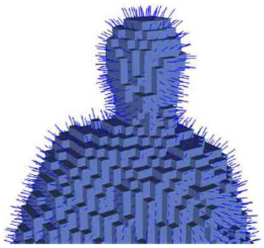

Fig. 3. Determination of normal vectors using Sobel filters on voxel-set slices

The curvature is then approximated for each surface voxel as the average angular displacement of the normals in a neighborhood of the voxel from its normal. If each voxel in the nearest neighborhood would lay on the same plane

of the considered voxel then all voxels would have the same normal and the curvature would be 0.

Degrees of Freedom					
Feature	Rot.	Trans.	Length	Radii	DOF
Joints					
Central Spine	3	3			6
Neck bone	2				2
Shoulder(2)	3				2 x 3
Elbow (2)	1				2 x 1
Pelvis	2				2
Hip (2)	3				3
knee (2)	1				2 x 1
Ankle (2)	2				2 x 2
Flesh descriptors					
Head				1	1
High Torso			1	2	3
Humerus (2)			1	2	3
Radius (2)			1	2	3
Femur (2)			1	2	3
Tibia (2)			1	2	3
Total					43

Table 1. Degrees of Freedom related to each model feature of the union of both Stick Figure and Flesh Figure; the Central Spine gives the relative reference of the backbone with the viewing system.

The following step is the determination of face; it is easily located since it shows a high variability of the curvature in almost all of its shape and, in particular, in the frontal part. These considerations allow an easy identification of the head and of its frontal part as shown in Fig.4. We further underline the fact that our algorithm do not use any information about face color and, in general, about textures making this approach robust to illumination changes, actor race and clothing color. Once the face and the center of mass is defined (the two main anchor points in our approach) it is possible to define a map-subdivision of the whole surface based on the geodesic distance from the center of mass. The definition of the '*geodesic distance*' we used is derived from the *graph theory* where the graph $G(V, E)$ is defined with the vertex set V and edge set E, with $n = |V|$ and $m = |E|$. Each edge $e \in E$ has an associated cost (or weight), $cost_G(e)$, which is a non-negative number. An $x - y$ shortest path $P_{x,y}(G) = e_1, e_2, ..., e_k$ is a collection of k edges of G forming a shortest path from x to y. The path $P_{x,y}(G)$ can also be defined as a sequence of nodes lying on the path. The length (total weight) of $P_{x,y}$ is denoted by the geodesic distance $d_G(x, y)$;

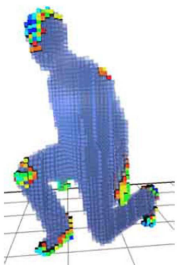

Fig. 4. curvature radius of the VH: points with lower radius (high curvature) are represented in blue and are mainly localized in the facial region while all the other colors represent bigger radii and points with neutral color represents very high radii (surface almost flat)

further details can be found in [11]. Examples of the application of this distance definition are given in Fig.1(right). Good results are obtained also in complex configurations, as shown below, even if it needs to be pointed out that this method suffers from the closed loop geodesic distance problem: in fact, when an arm gets in contact with the torso, the distance obtained from one point belonging to the torso and the distance of the tip of the hand from the center of mass will most likely be equivalent to the euclidean distance, since those two points would be directly connected (see, e.g., Fig.1(right)). This problem, even if unsolved in the geodesic distance definition is correctly tackled in the model fitting part.

2.2 The stick and flesh figure adopted

In Fig.2 we gave the representation of the stick and flesh figures we adopted that are mainly derived from the pioneer works of prof. Johansonn [12] and further developments; the stick figure (in the following SF) is not sufficient to correctly tackle complex poses without any initialization with a reference pose, but since we do not want to add constraints, a second layer based on a flesh figure (in the following FF) is used (similar approaches were followed by [13]and [14]). We described each articulation using one of these three shapes: a frustum of cone with circular bases (3 DOF, 1 for each radius and 1 for its height) a cylinder with elliptical bases (3 DOF, the major and minor axes and its height) and a sphere (1 DOF, its radius, used just for the head). In table 1 we report the DOF for each articulation. The proposed approach requires a set of assumptions (heuristics) about limbs articulation, flesh and cloth deformation and movement constraints; the main problem can be stated as: the *pose detection* is the search in the space of possible configurations of the FF where the best fitting with the voxels is achieved. The best fitting will correspond to the minimization of an

energy defined as:

$$E(\theta) = \frac{\sum\limits_{voxel-set} (V_{VH} \oplus V_{FF}(\theta))}{\sum\limits_{voxel-set} V_{VH}} \tag{2}$$

where the sum is made over each voxel of the voxel-set and $V_(VH)$ represents the VH: it is '1' if the voxel belongs to the VH and 1 otherwise. $V_(FF)$ is a sampling of the Flesh Figure in the Voxel-set and is '1' if the considered voxel belongs to the FF and '0' otherwise; the parameter θ indicates the values of the 43 DOFs. The operand $\oplus$ indicates the XOR logical operation and the denominator normalizes the whole sum to the VH volume. The energy definition we used is then a normalized Hamming distance between the VH and the FF and its minimization corresponds to the search for the optimal θ: $\bar{\theta} = \min\limits_{\theta}(E(\theta))$.

Many constrictions must be enforced to reduce the minimum search in the 43-dimensions space. It follows a list of the different types of constrictions we applied to the search and their major effects: distance of joints:

1. The lengths of each bone must be inside a predefined range which is scaled to the head length detected for the subject. The standard human body proportions are used. This is the base starting point for the search of each joint.
2. Every pair of symmetric bones or muscles is considered with the same size.
3. Inclusion in the voxel set: each bone must fully included in the voxelset. Assuming the characteristic of the VH to contain the entire volume of the captured subject, a bone which even partly laying outside has an extra high-penalty in the Energy value.
4. Angle checks: shoulders, hips and knee joints implement articulations limits to account for real human motion range.
5. Penetration check: bones defined by the flesh figure should not penetrate each other. This constriction add a further energy penalty to these configurations.

2.3 Detection sequence

To reduce computational time and to avoid local minima in the optimal parameters search we followed a branched fitting sequence, starting from the Torso and Head and then moving towards the peripheral 'leafs'. The Head/Face localization is performed using the surface curvature as described above, the next step is the fitting of a sphere within those voxels, this is done following locally an energy minimization approach similar to the one previously defined in (2) but applied only to this part. The radius of estimated sphere gives a first raw approximation of the skeleton size that must be used in the global fitting.

The next step consists in the accurate determination of the Torso: we used 6 DOFs for the Central Spine, 3 for the Pelvis and 2 for the Neck bone, this allowed us to deal with a Torso capable of 'twist and bend' action. This approach is quite different from the previous model of Trivedi [13] reported in Fig. 5(left) and allows us to tackle complex configurations as shown in Fig. 5(right). The first rough estimation of the Torso position is coaxial with the segment joining the

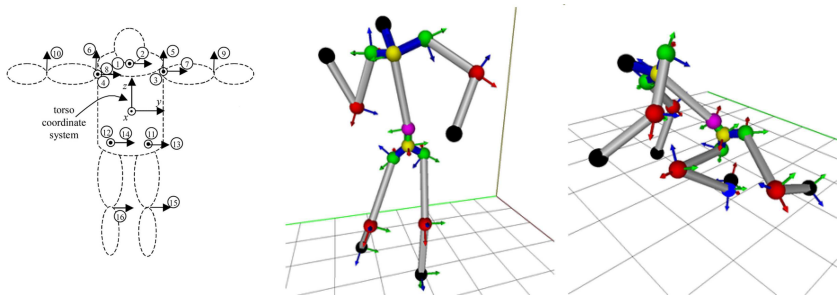

Fig. 5. (*left*) The skeleton model adopted bi Trivedi (*right*)some twisted and blended configurations that our model can honor

center of the head with the VH barycenter and further refinements are based on the previously described approach for local energy minimization. The next step concerns the lower body limb detection, in particular, starting from the lower part of Torso we will look for knees: we found them analyzing the curvature variations in a sphere centered at the lower part of the Torso and with radius equal to the femur; the center of mass of these high curvature zones is recognized as the Knees. Ankles are estimated using the same approach based on a sphere centered at the knee and with radius equal to the Tibia. These approaches based on local curvature could also be strengthened identifying local maxima in the geodesic distance from a reference point: knees and ankles, besides their curvature, present smooth local maxima of geodesic distance from a generic surface points, furthermore their slow decreasing behavior in the neighborhood of the maximum allows us to identify a wide maximum region discarding local curvature and/or distance maxima due to clothing or ouliers. The foot bone is then identified looking for the principal direction of the surface voxels farthest than the ankle. The localization of knee, ankle and feet is shown in Fig.6

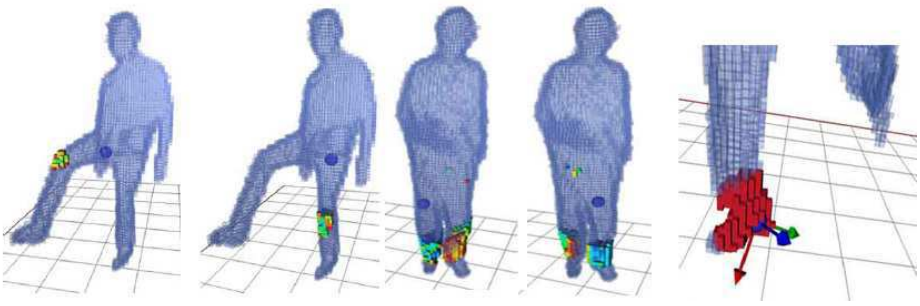

Fig. 6. From left to right: the localization of the knees, ankles and the identification of the feet bone based of the principal direction of farthest voxels.

2.4 Higher body Limbs detection

Elbows and hands detection follows a process similar to the lower body part: starting from shoulders geodesic distance is evaluated and elbows usually represent wide local maxima and hands represent absolute maxima. Both of them can be obtained with an exhaustive search in spheres respectively centered on shoulders and elbows with radii of humerus length and radius length. Anyway this kind of articulations present a lot of ambiguous situations when elbows or hands are in contact with torso or legs originating seamless VHs. In these cases arms or hands in the FF are assumed in contact with the torso. An example of higher body limbs localization is shown in Fig. (7).

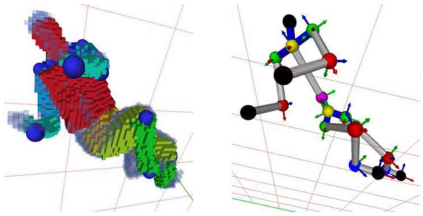

Fig. 7. Arms and hand configuration in complex configuration where Elbows are in contact with the Torso

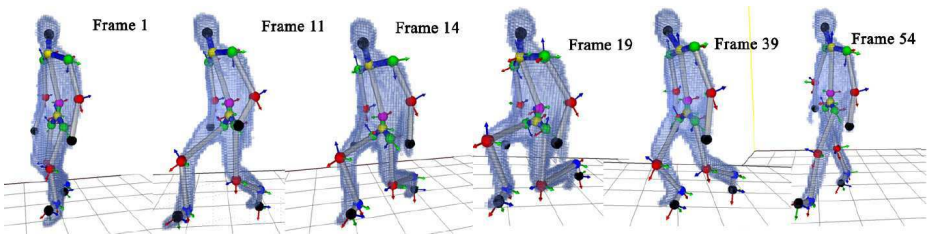

Fig. 8. 6 different poses identified during a movement

2.5 global refinement

After the rough estimation of the limbs position described below we can provide the energy functional described in (2) with a good guess of the SF and of FF; hopefully close to the global minimum. Further minimization is then acquired using an improved version of the simplex algorithm described by Nelder and Mead [15].

3 Experimental results

Our algorithm as been widely tested on numerous real world data acquired in our lab of different people also with complex configuration; The good results validated our heuristic approach for robust limb localization. They are examined through visual feedback and through the analysis of three dimensional paths of the articulations during a performed actions. Visual analysis allows an immediate gross evaluation of the results, as well as the individuation of critical poses. The plot of the three dimensional paths followed by the articulations allow detecting discontinuities in the movements. In Fig.(9) we show the SF fitting for 6 frames while in Fig.(10) the head position is shown for 66 consecutive frames. In Fig. we give some examples of correct fitting for complex configurations.

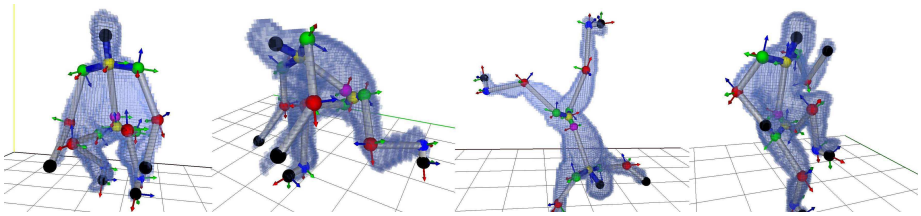

Fig. 9. 4 complex positions correctly recognized

4 Conclusions

We presented a novel approach to robust markerless human pose estimation. The presented system uses a VH obtained by 8 digital video cameras and does not require any assumption on the starting pose of the subject or on its physiognomy. Furthermore, there is no demand upon post processing of all frames to produce a reliable result since each frame is processed independently and the algorithm can be parallelized on different frames. The accuracy achieved in the estimate of the joints positions is not yet suited for a biomedic research and movement analysis, being lower than what such systems usually demand. It is on the contrary perfectly suited for the use in all those applications, which require an absence of restrictions imposed by the system and that ask for precision within 5cm in order to grasp the action. Actually there is no analysis between consecutive limbs estimations in different frames but further developments will implement these aspect to strengthen the localization with noisy VH.

References

1. K.Choo, D.J.Fleet: People tracking with hybrid monte carlo. In: IEEE International Conference on Computer Vision. Volume II. (2001) 321–328

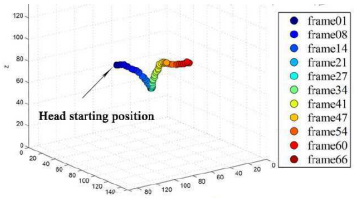

Fig. 10. Head 3D position during action performing

2. C.Bregler, J.Malik: Tracking people with twists and exponential maps. In: Computer Vision and Pattern Recognition. (1998) 8
3. S.Ioffe, Forsyth, D.: Probabilistic methods for finding people. International Journal of Computer Vision **43** (2001) 45–68
4. M.W.Lee, I.Cohen: Proposal maps driven mcmc for estimating human body pose in static images. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2004) 334–341
5. G.Mori, X.Ren, A.A.Efros, J.Malik: Recovering human body configurations: combining segmentation and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Volume 2. (2004) 326–333
6. X.Ren, A.C.Berg, J.Malik: Recovering human body configurations using pairwise constraints between parts. In: Proceedings of the 10th IEEE International Conference on Computer Vision. Volume 1. (2005) 824–831
7. R.Poppe, D.Heylen, A.Nijholt, M.Poel: Towards real-time body pose estimation for presenters in meeting environments. In: Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision. (2005)
8. M.Siddiqui, G.Medioni: Robust real-time upper body limb detection and tracking. In: ACM International Workshop on Visual Surveillance and Sensory Networks. (2006) 53–60
9. A.J.Capo, M.G.Hidalgo, R.Mas, F.J.Perales: Automatic human body modelling for vision-based motion capture. In: Proceedings of the International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision. (2006)
10. C.R.Wren, A.Azarhayejani, T.Darrell, A.P.Pentland: Pfinder: real-time tracking of the human body. IEEE Transaction on Pattern Analysis and Machine intelligence **19** (1997)
11. J. Bouttier, P. Di Francesco, E.G.: Geodesic distance in planar graphs. Nuclear Physics B **663** (2003) 535–567
12. G.Johanson: Visual perception of biological motion and a model for its analysis. Perception and Psychophysics **14** (1973) 201–211
13. I.Mikic, M.Trivedi, E.Hunter, P.Cosman: Human body model acquisition and tracking using voxel data. International Journal on Computer Vision **5** (2003) 199–223
14. M.Magnor, H.Theisel, Aguiar, E.D., C.Theobalt, H.P.Seidel: M3: Marker-free model reconstruction and motion tracking from 3d voxel data. In: In Proceedings of the 12th Pacific Conference on Computer Graphics and Applications. (2004) 101–110
15. McKinnon, K.: Convergence of the nelder-mead simplex method to a non-stationary point. SIAM J Optimization **9** (1999) 184–158