3D Human Motion Reconstruction Using Video Processing

Nadiya Roodsarabi and Alireza Behrad

Faculty of Engineering, Shahed University
Tehran, Iran
n_rsrb20@yahoo.com, behrad@shahed.ac.ir

Abstract. One of the important problems in human motion analysis is the 3D reconstruction of human motion, which utilizes the anatomic point's positions. These points can uniquely define the position and orientation of all anatomical segments. In this paper, a new method for reconstruction of human motion from the image sequence of a single static camera is described. In this method 2D tracking is used for 3D reconstruction, which a database of selected frames are used for the correction of tracking process. We use Discrete Cosine Transform (DCT) block as a "Matrix des criptor" used in the matching process for finding appropriate frame in the database and tracking process. Finally, 3D reconstruction is performed using Taylor method. By using DCT, we can select best frequency region for various tasks such as tracking, matching, correcting joints and so on. Experimental results showed the promise of the algorithm.

Keywords: Discrete Cosine Transform (DCT), Human motion reconstruction, Video processing, Occluded limb, Pose corresponding, Tracking, Matching.

1 Introduction

The realistic modeling and animation of human characters is one of the most difficult tasks in the vision and graphic community. In particular, body modeling from video sequences is a challenging problem that has been investigated a lot in the last decade. Recently the demand of 3D human models is drastically increased for applications like movies, video games, ergonomic, e-commerce, virtual environments and medicine.

A classical approach to build human shape models uses 3D scanners [1]. They are some disadvantages such as expensive instrument, limited flexibility (heavy wires) and freedom constraints, but they are simple to use and various softwarse are available to model the resultant measurements.

Due to the high number of degrees of freedom of the human body, motion tracking is a difficult problem. The problem is particularly challenging for visionbased (marker-less) approaches because of self occlusion, unknown kinematic information, perspective distortion and cluttered environments.

The existing approaches to human motion analysis can be roughly divided into two categories: model-based methods and model-free methods. In the model-based methods, a priori human model is used to represent the observed subjects, while in model-free methods; the motion information is derived directly from a sequence of

A. Elmoataz et al. (Eds.): ICISP 2008, LNCS 5099, pp. 386–395, 2008.

images. The main drawback of model-free methods is that they are usually designed to work with images taken from a known viewpoint. Model-based approaches, on the other hand, support viewpoint independent processing and have the potential to generalize across multiple viewpoints.

One decisive factor determining the used methodologies is the input data characteristics provided to the system by the acquisition stage. Some authors use monocular views [2-5], while others focus on multi-camera video streams [6-10]. Some limit their work to the use of calibrated views [8,9],[11-13], while others choose noncalibrated images [2]. Extracting monocular 3D human motion poses a number of difficulties such as Depth 3D-2D Projection Ambiguities, High-Dimensional Representation, Physical Constraints, Self-Occlusion and Observation Ambiguities.

Nowadays, monocular uncalibrated video sequences are the most common source of human motions. These methods which enable the extraction of detailed information about a specific uncalibrated sequence would greatly benefit applications related to video compression, video content-based classification and annotation industries. At a different level, the idea of being able to recover the motion of actors or historical celebrities from old movies and bringing them to life in new movies, animations, games or virtual environments is also very attractive. To compensate the lack of calibration, manual specification of key features such as joints or adjustment of a reference skeleton to specific frames are crucial [14]. To facilitate the process, motion databases can be used, becoming indispensable the containment of similar motion clips to the motion being recovered [15, 16].

Monocular methods for motion reconstruction are divided in two categories: 1-discriminative methods [4-5] 2- estimating and tracking methods [3]. In deiscriminative methods, 3D joint coordinates are found by using database, motion libraries and so on. In estimating and tracking methods, 3D information at a step is estimated using a sequence of images, prior and posterior to it.

Table 1. Tasks and required frequency regions

Tasks	Low frequency	Middle frequency	High frequency
Used frequency region for matching process		×	
used frequency region for database matching process	×	×	
Check tracking errors		×	
Updating RDMs if true tracking	×	×	×
Updating RDMs if false tracking			×
Updating RDMs if database matching occure		×	

In this article, we introduce a new descriptor with Discerete Cosine Transform used to various tasks in the algorithm. Tracking and matching is based on Reference

Matrix Descriptors (RMDs). These RDMs must be updated after each stage based on the frequency regions. Advantage of using descriptor is the capability of selecting required frequency in various tasks which results in better tracking and pose matching. For example، we use low and middle frequency in tracking for intensity and edge tracking. Also, we pass up color of clothes in database matching with avoiding low frequency.

Table 1 shows all tasks and their required frequency regions.

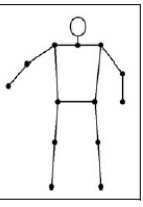
Organization: The paper is organized as follows. In section 2 we review the model considered to model 3D human motion. In Section 3 we give an overview of the algorithm of reconstructing 3D human motion using sequences of images acquired with a single video camera. Finally, we report the practical results in section 4.

2 Human Body Model

Human skeleton system is treated as a series of jointed links (segments), which can be modeled as rigid bodies. In the motion reconstruction applications, it is common to use a simple skeleton system modeling the important segments. We describe the body as a stick model consisting of a set of thirteen joint (plus the head) connected by thirteen segments as shown in Table 2.

Table 2. Relative lengths of the segments

Segment	Relative Length (MC) [cm]	Relative Length (L) [unit]
Height	175	8 i
Lower arm	35	2 i
Upper Arm	25	1 ½ i
Neck-Head	25	1 ¼ i
Shoulder Girdle	44	2 i
Torso	53	2 ½ i
Pelvic Girdle	30	1 ½ i
Upper leg	46	2 i
Lower leg	52	2 i
Foot	*22*	*1 i*



The algorithm needs the knowledge of relative lengths of the segments for the 3D reconstruction purpose, which can be obtained from anthropometric data. Table 2 shows this relative length. With known 2D position and using the knowledge of length of the segments and enforceing some constraints such as dynamic smoothing, we can obtain 3D reconstruction.

3 Overview of Algorithm

In this method, we locate 2D joints position using a fixed and uncalibrated monocular video and use them to estimate 3D skeletal configuration. As regards no enough information is available from monocular video; we save several 2D exemplar of various body poses in the database that used to correct tracked points.

In the preposed algorithm, joint tracking is performed based on a n*n block of DCTs coefficient (descriptor matrix). Algorithm starts with background subtraction and is initialized by user through specifying 2D joint positions in first frame. Then for each joint, descriptor matrix is computed and saved as “Reference Descriptor Matrix”

for the same joint. In the next stage, all joints are tracked with its own RDM. After finding joint positions in next frame, RDMs are updated based on DCT block frequency regions considering occlusion problem and tracking errors. Then, human pose in current frame is compared with poses in database based on middle frequency. If corresponding occurred, joint positions are corrected and RDMs are updated. The reason for using middle frequency is that we want pass up clothing color (low frequency) and body deformation details (high frequency).

A major problem may be encountered is the occlution of joints. We solve this problem with "occluded" label for each tracked joints. After tracking process for all frames, "occluded" labeled joints are corrected by interpolation. Given the 2D joint locations, the 3D body configuration and pose are estimated using the algorithm of Taylor [18]. Figure 1 shows the overview of algorithm.

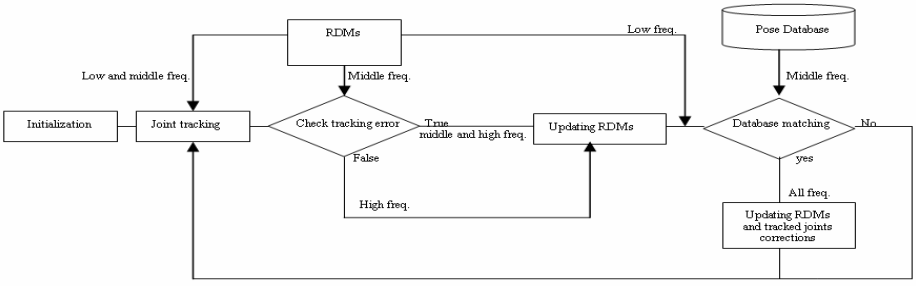


Fig. 1. Overview of algorithm

3.1 Descriptor Matrix

In this article, we use DCT block for tracking and matching purposes. "Discriptor Matrix" for the point p_i is n*n DCT coeffitient matrix. By putting the image window of fixed size (n*n) centerd at point p_i into matrix A, a descriptor matrix (DM) for p_i is then compute by equation 1.

$$F(u,v)=C(u)C(v)\left[\sum_{x=px-\frac{n}{2}}^{x=px+\frac{n}{2}}\sum_{y=py-\frac{n}{2}}^{y=py+\frac{n}{2}} f_i(x,y)*\cos\frac{(2x+1)u\pi}{2n}\cos\frac{(2x+1)v\pi}{2n}\right] \quad (1)$$

where px and py are p_i coordinate. Also, if x=0; $C(x)=1/\sqrt{n}$ otherwise $C(x)=\sqrt{2/n}$.

There are n^2 coefficients in each DM matrix divided into three frequency regions accorging to figure 2. White region is low frequency region, gray region is middle frequency region and black region is high frequency region. We use this three frequency region for tracking and matching algorithm.

Matrix distance in special frequency region is defined according to:

$$Mdis_{frequencyregion}(M,N)=\sqrt{\sum_f\left(M_f^2-N_f^2\right)} \qquad f\in frequency \quad region \quad (2)$$

DC	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Fig. 2. Discerete Cosine Transform coeffitients and frequency regions for a 8*8 block

3.2 Reference Descriptor Matrix (RDM)

These matrices save tracked joints information using previous frames and database information and are used to find corresponding joints in tracking process. We generate a reference descriptor matrix for each joint ($RDM_1 \ldots RDM_{13}$). Reference descriptor matrix for joint j (RDM_j) are loaded from descriptor matrix for joint j after initialization and updated after finding new tracked joint j in next frame. Updating routine is different in each frequency region:

Low frequency region: This region consists of general shape and intensity information of tracked joint. So its changes are small in successive frames. Tracking process may lose tracked object for several reasons such as occluding problem or large distortion and tracked joint information may be incorrect. For safekeeping of general object information, we leave the low frequency coeffitients unchanged after tracking. These matrices will be updated only when corresponding to database occure.

Middle frequency region: This region consists of general edge information. Because the individual limbs are deformable due to moving muscle and clothing, we apply changes in middle frequency coeffitients after tracking if tracking is accurate. In order to determine the accuracy of the tracked joint j at frame t+1, we calculate matrix distance in middle frequency between descriptor matrix of tracked point ($DM_j(t+1)$) and RDM_j .

$$Mdis_{middle}(RDM_j, DM_j(t+1)) \begin{cases} < \Delta & j(t+1) \quad is \quad true \\ > \Delta & j(t+1) \quad is \quad false \end{cases} \tag{3}$$

We update this frequency region if j(t+1) is true according to equation 3 after tracking. Also this region is updated when corresponding to database occure.

High frequency region: this region is consists of noise and object details and must be updated after each tracking phase.

3.3 Tracking

The tracking process is based on frequency domain matching techniques.

Tracking process aims to find body joints in successive frames. Because of temporal correspondences between subsequent frames, search for corresponding joint is local. In two successive frames, limbs and joints have the same intensity and general shape, but they are different in details. So we use low and middle frequency in tracking process.

The tracking process is based on DCT matching techniques. Its basic idea is to track joints through the sequence. Descriptor matrices are computed for every pixel in the search window (SWDMs). Finding best match is performed by selecting minimum matrix distance between low and middle frequency of RDM_j and search window descriptor matrices (SWDMs).

Assuming a first estimate of the pose has been given, the tracking algorithm can be summarized in 2 steps as follows:

1- Generate descriptor matrices for all pixels in search window at frame t+1 (SWDMs)
2- Determine best matching point in search window by computing matrix distance between RDM_j and SWDMs.

3.4 Database Matching Process

The database consists of different poses required information of video sequences of number of subjects. This information is body joint positions and middle frequency of their descriptor matrices and necessary labels for 3D reconstruction. Head position has been selected as reference joint to adjust two poses and other joint positions have determined towards it.

Measuring similarity between human pose in current frame (p_f) and human pose in database (p_d) requires two kinds of matrix: DDMs and FDMs, which will be defined later. If pose distance is smaller than Δ, corresponding occurs; therefore points and middle frequency of RDM must be corrected. Pose distance is defined by:

$$Pdis_p(p_f, p_d) = \sum_{j=1}^{13} Mdis_{low,mid}(DDM_j, FDM_j) \tag{4}$$

Database descriptor matrix (DDM) is generated using low frequency of RDM (for intensity joint similarity) and midlle frequency of database (for edge similarity).

$$DDM_f = \begin{cases} RDM_f & f: low \quad frequency \\ Database & f: middle \quad frequency \\ 0 & f: high \quad frequency \end{cases} \tag{5}$$

Frame descriptor matrix (FDM) is generated by following process:

1- Search locally around the previous head position to find corresponding to RDM_{head} point in the current frame.
2- Determine other joints in current frame with adjusting head position.
3- Generate descriptor matrices for each joint and save them as FDMs.

Measuring similarity algorithm between human pose in current frame (p_f) and human pose in database (p_d) can be summarized as follows:

1- Generate DDMs.
2- Search locally to find head position in current frame.
3- Determine other point positions in the current frame.
4- Compute matrix distance for DDM and FDM in low and middle frequency.
5- Correct points if corresponding occure.
6- Updating RDMs.

3.5 Estimating 3D Reconstruction

We use Taylor's method [18] to estimate the 3D configuration of a body given the keypoint position estimates. Taylor's method works on a single 2D image, taken with an uncalibrated camera.

It assumes that we know:

1- the image coordinates of keypoints (u, v).
2- the relative lengths l of body segments connecting these keypoints.
3- a labelling of "closer endpoint" for each of these body segments.
4- that we are using a scaled orthographic projection model for the camera.

In our work, the image coordinates of keypoints are obtained via the deformable matching process. The "closer endpoint" labels are supplied on the exemplars, and automatically transferred to an input image after the matching process. The relative lengths of body segments are fixed in advance, but could also be transferred from exemplars.We use the same 3D kinematic model defined over keypoints as that in Taylor's work.

We can solve for the 3D configuration of the body {(Xi, Yi, Zi): i ∈ keypoints} up to some ambiguity in scale s. The method considers the foreshortening of each body segment to construct the estimate of body configuration. For each pair of body segment endpoints, we have the following equations:

$$l^2 = (X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2 \tag{6}$$

$$(u_1 - u_2) = s(X_1 - X_2) \tag{7}$$

$$(v_1 - v_2) = s(Y_1 - Y_2) \tag{8}$$

$$dZ = (Z_1 - Z_2) \tag{9}$$

$$dZ = \sqrt{l^2 - \left((u_1 - u_2)^2 + (v_1 - v_2)^2\right)/s^2} \tag{10}$$

To estimate the configuration of a body, we first fix one keypoint as the reference point and then compute the positions of the others with respect to the reference point. Since we are using a scaled orthographic projection model the X and Y coordinates are known up to the scale s. All that remains is to compute relative depths of endpoints dZ. We compute the amount of foreshortening, and use the user-supplied "closer endpoint" labels from the closest matching exemplar to solve for the relative depths.

Moreover, Taylor notes that the minimum scale smin can be estimated from the fact that dZ cannot be complex.

$$s \geq \frac{\sqrt{(u_1 - u_2)^2 + (v_1 - v_2)^2}}{l^2} \tag{11}$$

This minimum value is a good estimate for the scale since one of the body segments is often perpendicular to the viewing direction.

4 Experimental Results

The propsed algorithm is applied for reconstructiong of human subjects from single-camera video. The database consists of some poses of number of subjects, performing different types of motions from the CMU MoBo Database. On this collection of poses, we manually determined joint locations of each pose and "closer endpoint" labels for each body segment used in Taylor's method. Also, we save middle frequency of descriptor matrix for each labeled joint.

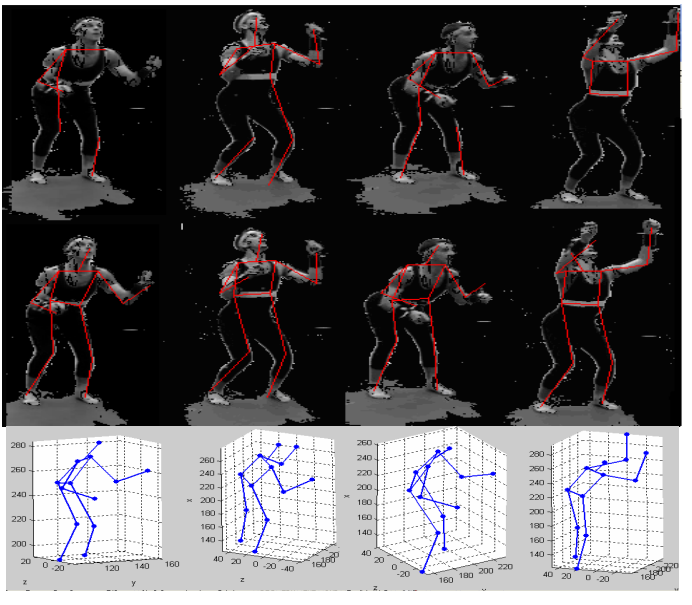


Fig. 3. Reconstruction results for 4 frames. Row 1 shows 2D tracking before interpolation for some frames of the video sequence (7, 14, 54 and 79). Row 2 shows 2D tracking after interpolation. Row 3 shows 3D reconstruction results.

Background subtraction was used to facilitate tracking. We use a skeleton with 13 joints in our experiment and apply a 16*16 DCT block as a descriptor for each point. Then 2D and 3D reconstruction is performed. Figure 3 shows sample results of 2D body joint localization before and after interpolation and finally 3D reconstruction on the CMU dataset. Note that some joints are occluded or failed in 2D tracking. These joints are reconstructed by interpolation. The same body parameters (lengths of body segments) are used in all 3D reconstructions. Figure 4 shows a comparison between 2D optical flow tracking using iterative Lucas-Kanade method in pyramids [20] and our 2D tracking method. As it is shown our method has better results.

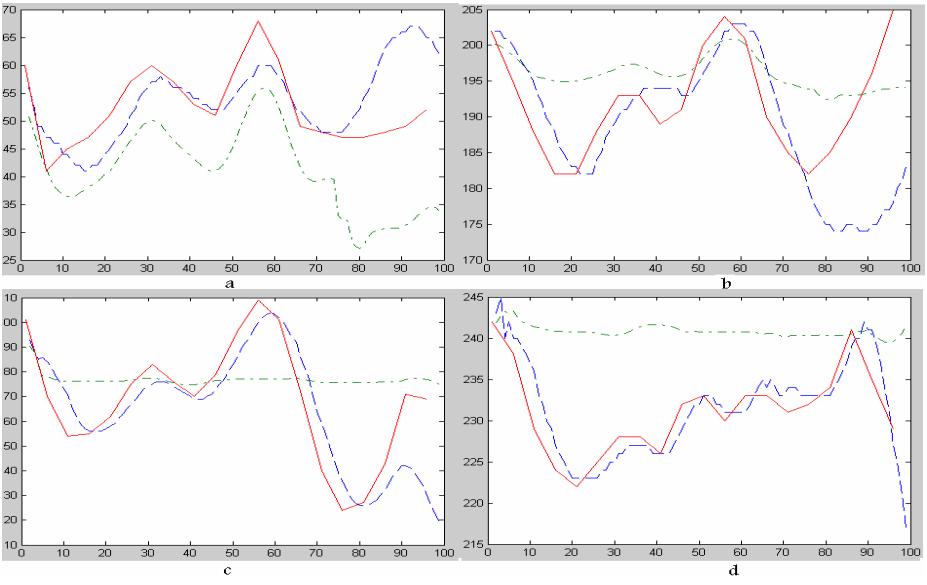


Fig. 4. Comparison between our tracking method and optical flow method using iterative Lucas-Kanade method in pyramids. a) True head position along the x axis (___), our method (- - -) and optical flow method (-.-.-). B) True head position along the y axis (___), our method (- - -) and optical flow method (-.-.-). c) True right hand position along the x axis (___), our method (- - -) and optical flow method (-.-.-). d) True right hand position along the y axis (___), our method (- - -) and optical flow method (-.-.-).

5 Conclusion

In this paper, a new method for reconstruction of human motion from the image sequence of a single static camera is described. In this method, 2D tracking is used for 3D reconstruction, which a database of selected frames are used for the correction of tracking process. We used Discrete Cosine Transform (DCT) block as a "Matrix descriptor" used in the matching process for finding appropriate frame in the database and tracking process.

The reconstruction algorithm was tested with several sequences and experimental results showed the reliability of our algorithm. This method is robust in 2D tracking and holding properties of each joint along tracking process.

References

1. Horiguchi: Body Line Scanner.The development of a new 3-D measurement and Reconstruction system. In: International Archives of Photogrammetry and Remote Sensing, vol. 32, pp. 421-429 (1998)
2. Barrón, C., Kakadiaris, I.A.: Estimating anthropometry and pose from single uncalibrated image. Computer Vision and Image Understanding 81(3), 269–284 (2001)

3. Sminchisescu, C., Triggs, B.: Kinematic Jump Processes for Monocular 3D Human Tracking. In: IEEE International Conference on Computer Vision and Pattern Recognition (2003)
4. Chen, C., Zhuang, Y., Xiao, J.: Towards Robust 3D Reconstruction of Human Motion from Monocular Video. In: Pan, Z., Cheok, A.D., Haller, M., Lau, R.W.H., Saito, H., Liang, R. (eds.) ICAT 2006. LNCS, vol. 4282, pp. 594–603. Springer, Heidelberg (2006)
5. Loy, G., Eriksson, M., Sullivan, J., Carlsson, S.: Monocular 3D Reconstruction of Human Motion in Long Action Sequences. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 442–455. Springer, Heidelberg (2004)
6. Hilton, A., Beresford, D., Gentils, T., Smith, R., Sun, W., Illingworth, J.: Whole-body modelling of people from multiview images to populate virtual worlds. The Visual Computer, International Journal of Computer Graphics 16, 411–436 (2000)
7. D´Apuzzo, N., Plänkers, R., Fua, P., Gruen, A., Thalmann, D.: Modeling human bodies from video sequences. In: El-Hakim, S.F., Gruen, A. (eds.) SPIE. Videometrics VI, vol. 3461, pp. 36–47 (1999)
8. Plänkers, R., Fua, P.: Tracking and modeling people in video sequences. Computer Vision and Image Understanding 81(3), 285–302 (2001)
9. Mikić, I., Triverdi, M., Hunter, E., Cosman, P.: Articulated body posture estimation from multi-camera voxel data. In: IEEE Conference on Computer Vision and Pattern Recognition, Kauai, HI, USA, vol. I, pp. 455–460 (2001)
10. Cheung, K.M., Kanade, T., Bouguet, J.Y., Holler, M.: A real time system for robust 3D voxel reconstruction of human motions. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 714–720 (2000)
11. Iwasawa, S., Ohya, J., Takahashi, K., Sakaguchi, T., Ebihara, K., Morishima, S.: Human body postures from trinocular camera images. In: IEEE International Conference on Automatic Face and Gesture Recognition, pp. 326–331 (2000)
12. Plänkers, R., Fua, P., D'Apuzzo, N.: Automated body modeling from video sequences. In: International Conf. of Computer Vision (Workshop on Modeling People), pp. 45–52 (1999)
13. Rosales, R.E., Sclaroff, S.: Learning and synthesizing human body motion and posture. In: IEEE International Conf. on Automatic Face and Gesture Recognition, pp. 506–511 (2000)
14. Park, M.J., Choi, M.G., Shin, S.Y.: Human motion reconstruction from inter-frame feature correspondences of a single video stream using a motion library. In: ACM SIGGRAPH Symposium on Computer Animation, pp. 113–120 (2002)
15. Mori, G., Malik, J.: Estimating human body configurations using shape context matching. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 666–680. Springer, Heidelberg (2002)
16. Park, M.J., Choi, M.G., Gawa, Y.S., Shin, S.Y.: Video-Guided Motion Synthesis Using Example Motions. ACM Transactions on Graphics 25(4) (2006)
17. Mori, G., Belongie, S., Malik, J.: Shape contexts enable efficient retrieval of similar shapes. In: Proc. IEEE Comput.Soc. Conf. Computer Vision and Pattern Recognition, vol. 1, pp. 723–730 (2001)
18. Taylor, C.J.: Reconstruction of articulated objects from point correspondences in a single uncalibrated image. In: CVIU, vol. 80, pp. 349–363 (2000)
19. Remondino, F., Roditakis, A.: 3D Reconstruction of Human Skeleton from Single Images or Monocular Video Sequences. In: Michaelis, B., Krell, G. (eds.) DAGM 2003. LNCS, vol. 2781, pp. 100–107. Springer, Heidelberg (2003)
20. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI, pp. 674–679 (1981)