Multimed Tools Appl (2016) 75:6809–6827
DOI 10.1007/s11042-015-2611-8

Pose estimation of soccer players using multiple uncalibrated cameras

Reza Afrouzian[1] · **Hadi Seyedarabi**[1] · **Shohreh Kasaei**[2]

Received: 20 June 2014 / Revised: 10 February 2015 / Accepted: 6 April 2015 /
Published online: 23 April 2015

Abstract Fully automatic algorithm for estimating the 3D human pose from multiple uncalibrated cameras is presented. Unlike the state-of-the-art methods which use the estimated pose of previous frames to restrict the candidates of current frame, the proposed method uses the viewpoint of previous frame in order to obtain an accurate pose. This paper also introduces a method to incorporate pose estimation results of several cameras without using the calibration information. The algorithm employs a rich descriptor for matching purposes. The performance of the proposed method is evaluated on a soccer database which is captured by multiple cameras. The dataset of silhouettes, in which the related 3D skeleton poses are known, is also constructed. Experimental results show that the proposed algorithm has a high accuracy rate in estimation of 3D pose of soccer players.

Keywords Shape context · 3D human pose estimation · Soccer match · Uncalibrated cameras · Silhouette

1 Introduction

Human pose estimation is one of the active research fields in computer vision. It has shown to be effective in many real applications such as human-computer interaction, surveillance, entertainment industry, video-based rendering, sport motion analysis, etc.

Human pose estimation is a challenging and complicated task. Hence, usually human motion capture systems (along with using image information) employ supplementary tools

✉ Reza Afrouzian
afrouzian@tabrizu.ac.ir

Hadi Seyedarabi
seyedarabi@tabrizu.ac.ir

Shohreh Kasaei
skasaei@sharif.edu

[1] Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran

[2] Department of Computer Engineering, Sharif University of Technology, Tehran, Iran

such as markers (e.g., Vicon) or complement sensors (e.g., Kinect) to find the 3D position of human body joints. Most of significant problems that make it to be sophisticated are: few number of cameras in the scene, self-occlusion of human body, high dimensionality of pose configuration space, calibration error in cameras, Loose-fitting clothes with varying appearance, variability in human body physique, and complicated background with a variations in lighting conditions [19]. Hence, there is no algorithm that can deal with all of these problems and the state-of-the-art methods only focus on some few cases.

In order to categorize the current works in pose estimation and motion capture field, it is appropriate to consider some basic factors which may be used or not be used in each algorithm. These factors can be summarized as follow:

1. calibrated cameras vs. uncalibrated cameras,
2. 2D image features vs. 3D reconstruction of human body,
3. video sequences from monocular camera vs. footage from several cameras,
4. spatial and temporal information vs. only spatial information,
5. model-based vs. model-free pose estimation methods,
6. using of learned motion models vs. not using them.

This paper proposes a method to estimate 3D human body poses of soccer players. In order to achieve this purpose, a dataset of silhouettes is used which has various pose configurations and viewpoints with respect to the camera. In comparison with similar works, instead of monocular camera, several cameras (two or three) are used to find the nearest pose in the dataset. However, since the cameras have to track the fast motions of players and ball in the scene, their calibration information should be updated in each frame; due to variations in their internal and external parameters. As a result, implementation of 3D pose estimation systems, which employ calibration information, would be costly. In order to overcome this problem, in this paper, a novel algorithm for 3D pose estimation of human bodies is introduced that shares the information of several cameras without using any calibration data (unlike other state-of-the-art algorithms).

Shape context descriptor extracts features to compare query images against the dataset samples. Instead of finding an optimal assignment between edge points on the query image and those on the dataset [2, 20], the proposed method localizes several points on the image and transfers shape context information of edge points to them. Then, the comparison is performed among the histograms of shape contexts in the known points and there is no need to use an assignment problem solver to find the optimal matching among the edge points of bodies.

Exploiting temporal information has an important role in the tracking stage. In fact, it helps to restrict the search area in the human pose configuration space. On the other hand, it also increases the dependency to previous frames and thus yields to some difficulties when the tracking algorithm fails. Fortunately, the proposed method does not require the estimated pose of previous frames to restrict the candidates of current frame. Instead, it employs the viewpoints of previous frame to enhance the estimation accuracy and to eliminate irrelative candidates. Also, this paper incorporates the offered nearest 3D pose candidates of cameras by using the similarity matrix without any need to employ the calibration information of cameras.

The main contributions of this paper were highlighted in the above three paragraphs. The rest of this paper is organized as follows. Section 2 discusses about related work (literature review). General framework of the proposed method is outlined in Section 3. Section 3.1

describes how to employ the histogram of shape context descriptor in the proposed algorithm. In order to reduce the effect of irrelative poses, the viewpoint constraint is introduced in Section 3.2. Section 3.3 provides a method to aggregate the information of several cameras and find the nearest pose from dataset. Experimental results are presented in Section 4. Section 4.1 explains how to construct a dataset of silhouette images. Evaluation of the proposed method is performed in Section 4.2. Section 4.3 gives the performance comparison among the proposed method and other existing algorithms. Finally, Section 5 concludes the paper with some discussions and directions for future work.

2 Related work

Due to the importance of human body pose estimation topic, several survey papers have been published in recent years [12, 14, 17, 18, 21, 23]. Sigal, et al. have categorized pose estimation and tracking methods into three major classes including: generative (model-based), discriminative (model-free), and part-based methods [19, 23]. Also, Moeslund, et al. have divided pose estimation process into some stages including: initialization, tracking, pose estimation and recognition. They have also separated pose estimation algorithms into model-free-based, direct model-based, and indirect model-based methods according to a prior human model [17, 18].

In direct model-based methods, the human body model is explicitly used which describes both shape and kinematic features to reconstruct the pose. Majority of current works in 3D human pose estimation belong to this category [8, 11]. Some of them use 3D features for pose estimation (e.g., voxel data) and the others employ 2D features which are extracted from images. The direct model-based algorithms need the calibration information for either 3D reconstruction of human body or projection of 3D human volume on images. Also, such approaches utilize an analysis-by-synthesis methodology to optimize the similarity between the model projection and observed images

Model-free methods employ a bottom-up approach to track and label body parts in 3D reconstructed volumes or images. Also, the part-based methods are bottom-up approaches in which the body can be represented as an assembly of parts which are connected together. This connection is determined by considering the constraints which are imposed through the joints within the skeleton structure [27]. The application of such methods arises from their ability to allow inference in very complex natural scenes from just a single image without involving any temporal information [19]. But, these methods are limited to 2D cases and are unable to recover 3D poses from monocular images.

Indirect model-based methods use a priori model in pose estimation as a reference or look-up table to guide the interpretation of measured data. This class is itself divided into two main categories: learning-based and example-based. In the learning-based approaches, a function from image space to pose space is learned by using some training data whereas example-based approaches avoid learning this kind of mapping. Instead, a collection of exemplars in different poses and orientations is stored in a dataset, for which the related 3D skeleton poses are known. In this category, majority of works have employed a single camera. Shakhnarovich, et al. [22] employed exemple-based methods for pose estimation. Since the computational complexity of the method increases with the number of examples in the dataset, it employs a set of hashing functions that efficiently index examples relevant to a particular estimation task. Agarwal and Triggs [1] proposed a learning-based method to recover 3D human body pose

from a monocular camera. They employed shape context descriptor for feature extraction and *relevance vector machine* (RVM) regression to recover the pose using shape descriptor vectors. In another work, Mori and Malik [20] focused on the images captured by an uncalibrated monocular camera. The query image is matched to the dataset by using the technique of shape context matching in conjunction with a kinematic chain-based deformation model. Howe [13] introduced an effective method for tracking human body pose based on looking up observations within a collection of known poses. Chamfer distance was used as the shape descriptor and Markov chaining exploited the temporal dependency of human motion to eliminate unlikely pose sequences.

As mentioned in Section 1, particular focus of this paper is on estimating the 3D pose of soccer players during an uncontrolled match environment. In recent years, several papers have been published in the literature in which their proposed methods have considered this purpose. Germann, et al. [10] performed a rough 2D pose estimation for each camera using a spatial and temporal silhouette-based search in a dataset containing different poses. Also, they employed calibration information and some optimization techniques which combines spatial and temporal constraints in order to achieve the final 3D pose. In a similar work, 3D pose estimation of soccer players was used for video-based rendering [9]. Burenius, et al. [5] focused on pictorial structures (a famous method in 2D human body parts detection [7]) and proposed a new framework to employ it in three dimensions. They evaluated their method in terms of speed, accuracy, and required memory. Kazemi, et al. [16] utilized 3D pictorial structure. They focused on learning the 2D part likelihood for each part and aggregated the likelihoods from the different views to obtain the 3D part likelihoods. However, they used a random forest classifier which has the advantages of capturing the variation in the appearance of body parts in 2D images. In the other work by Kazemi [15], the *flexible mixture of parts* (FMP) model [26] was employed to detect body parts in the images. In the mentioned works by Kazemi and Burenius [5, 15, 16], calibration information is known and no temporal information is exploited.

The proposed algorithm of this paper belongs to example-based methods (a subset of indirect model-based category) which use a dataset of silhouettes in different poses and orientations. This dataset are employed independently for each camera without incorporating information of other cameras. Unlike multi-camera pose estimation methods, this paper incorporates the selected candidates of all cameras without employing any calibration information after obtaining nearest 3D poses in each camera.

3 Proposed method

The main framework of the proposed method is described in this section. Figure 1 illustrates the three main steps of the method. In the first step, query images of multiple cameras are segmented and the player body area is subtracted from the background. In this paper, it is assumed that a coarse segmentation of the body is available by employing foreground-background subtraction methods. In the second step, the shape contexts of edge points are calculated and used to obtain the features of marked points (which are localized in the normalized image after clustering the edge points of dataset samples) on the image. By using shape context features of the marked points, the nearest 3D body configuration of dataset samples is assigned to the query image. This step is performed for each camera independently without employing the information of other cameras. In some cases, estimation of the 3D

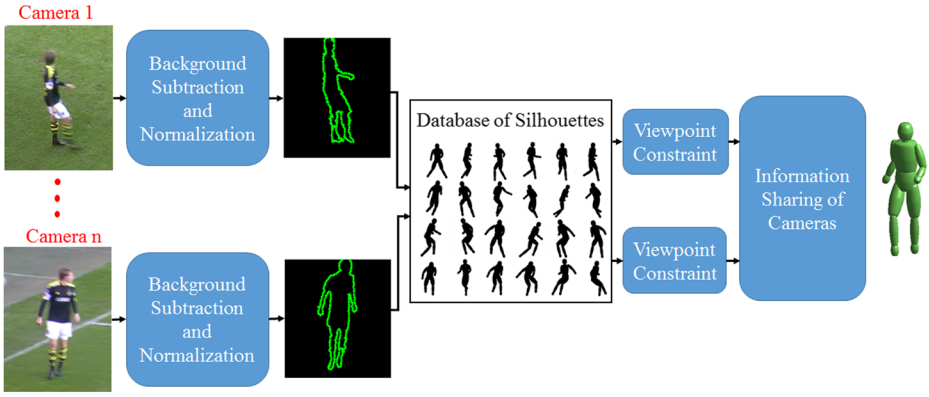

Fig. 1 Overview of the proposed algorithm

human body pose from a monocular camera yields incorrect outputs. These errors are caused by: inaccuracy in foreground-background segmentation, complicated topology of human body, similarity of various 3D poses in 2D views, and finally employment of silhouette images which only makes use of shape information (and not the texture information). Hence, in the next stage, the viewpoint of previous frame is used to remove incorrect poses of each camera.

Up to the third step, information of each camera is processed independently. In this step, the final 3D pose of query images is estimated by incorporating pose candidates of all cameras in the scene. This stage does not use the calibration information and instead it employs the similarity matrix to mutually measure the likeness among 3D pose candidates. It then chooses the nearest 3D human body pose from the dataset.

3.1 Shape context descriptor

Shape context is a rich descriptor used in object recognition and matching [2]. In this paper, after calculating the shape context histogram of edge points, instead of solving for the correspondence between sample points on the test body and those on the dataset, the marked points are employed for comparison purposes. Information within the histogram bins of edge points is transferred to the marked points. As a result, the assignment problem solver (the optimization process), which has relatively heavy computational burden, is converted to a simple and computationally less complex comparison among marked points.

In order to obtain the marked points, first each human body image in the dataset is normalized to an image with size of 200×200 pixels such that the mean of edge points is placed in the center of the normalized image. Next, the 2D position of body edge points belonging to all dataset samples (16,614 images) are considered as the input data to the K-means algorithm for categorizing them to 100 clusters. The center of each cluster is then considered as the marked point. Figure 2 shows the position of marked points in the normalized image.

Since the marked points have constant positions, we have to normalize the input images in order to avoid mismatching caused by two types of distortions. The first distortion is caused by

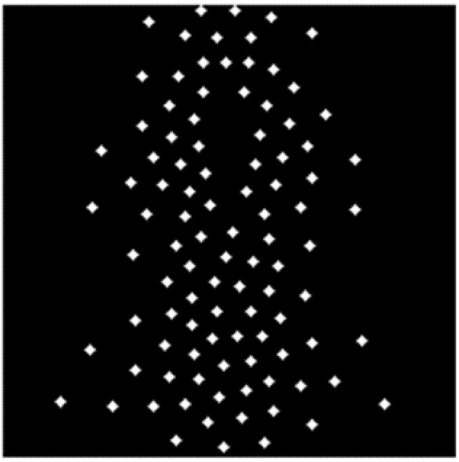

Fig. 2 Position of marked points in the normalized image

the varying distance of soccer players from the camera (which leads to size change of player in the image). The second one is caused by holding the player in various borders of the image (which shifts the position of player from the image center).

The information of shape context histogram of edge points should be transferred to the marked points such that each marked point includes the data existing around it. In order to achieve this goal, the edge points which are within a neighborhood of the marked point should have greater weight in forming the shape context histogram of that marked point. In this paper, Eq. (1) is used to calculate the normalized histogram of marked points in each image. In this equation, the distance criterion $diss(e_p,m_p)$ has to satisfy the above-mentioned constraint about the weight of neighboring edge points. This paper proposes to employ the exponential function of Eq. (2) as a distance criterion. It amplifies the impact of nearer edge points to the marked point and attenuates the effect of farther edge points. Furthermore, for edge points which are farther than a threshold (T), the output of the distance function $diss(e_p,m_p)$ should be set to zero, as it has been considered in Eq. (2). The reason for this constraint is that if the distances of the marked point from all edge points is higher than a threshold (T), the histogram of that point cannot contain the information of the shape context histogram of neighboring edge points. This is because the normalization factor in the denominator of Eq. (1) amplifies the information of edge points which are farther from the marked point. The human body is represented by a discrete set of k points $E_p=\{e_{p1},e_{p2},\ldots,e_{pk}\}$, $e_{pi}\in\mathbb{R}^2$ which are subsampled from edge points of the body. As mentioned above, the histogram in marked points, is defined by

$$h_{mn}=\begin{cases}\dfrac{\sum_{i=1}^{k} h_{ei}\times diss\left(e_{pi},m_{pn}\right)}{\sum_{i=1}^{k} diss\left(e_{pi},m_{pn}\right)} & \sum_{i=1}^{k} diss\left(e_{pi},m_{pn}\right)\neq 0\\ \mathbf{0} & otherwise\end{cases} \tag{1}$$

where h_{mn} is the histogram of the n'th marked point, $n=1,2,\ldots,100$; h_{ei} is the histogram of e_{pi} which includes 60 bins (here shape contexts contain 12 angular×5 radial log-polar bins, giving rise to 60

dimensional histogram), and **0** indicates a vector of 60 zeros. As mentioned above, the distance criterion is defined by

$$diss(e_{pi}, m_{pi}) = \begin{cases} e^{-\lambda\left\|e_{pi}-m_{pi}\right\|} & \left\|e_{pi}-m_{pi}\right\| \le T \\ 0 & otherwise \end{cases} \quad (2)$$

where $\|e_{pi}-m_{pi}\|$ represents the Euclidean distance between edge points e_{pi} and marked points m_{pi}, λ is a degree of freedom which should be set by the user, and T indicates the distance threshold. The proper values for parameters λ and T are chosen by cross-validation method (trial and error). Of course, if the mean of pairwise distances between mutual marked points is considered in the cross-validation stage, the parameter selection process will be faster and more logical (because this task biases the initial choices toward relatively rational values). After some tests, the best empirical values for these parameters are obtained (these are given in Section 4).

In summary, Eqs. (1) and (2) produce higher coefficient values for edge points near the marked point and lower coefficient values to farther edge points. Thus, the histogram of marked point is mostly influenced by the histogram of nearer edge points.

In order to quantify the similarity between the query image and dataset samples, the Chi squared distance

$$C_{ti} = \frac{1}{2}\sum_{n=1}^{100}\sum_{j=1}^{60}\frac{\left[h^t_{mn}(j) - h^i_{mn}(j)\right]^2}{h^t_{mn}(j) + h^i_{mn}(j)} \quad (3)$$

is used where h^t_{mn} corresponds to the histogram of the query image in the n'th marked point, and h^i_{mn} is the histogram of the n'th marked point in the i'th sample of dataset.

It is worth mentioning that in addition to simplicity and higher speed of the proposed shape context matching with the marked points, it can represent the difference between two shapes based on their locations. This property can be employed in the optimization stage. It increases the speed and accuracy of the system.

Despite the benefits of using shape context descriptor in body matching, it has some disadvantages because of: poor foreground-background subtraction, loose-fitting clothing of soccer players, different body styles of players, humanoid of dataset, and problems corresponding to silhouette images (which only use the body edge data and do not provide any information of the image texture and thus, they cannot resolve problems such as self-occlusion). Due to the above-mentioned disadvantages, shape context alone cannot provide accurate pose estimation results and dataset sample which has the most similarity to the input image, in terms of shape context feature, is not necessarily the most nearest 3D pose. Hence, instead of employing the best matching result, N samples of dataset which are the most similar to the query image are selected. The number of selected samples (N) is chosen based on the number of dataset images and existing similarities among them, such that the final result usually lies among N selected samples (this number is given in Section 4). Viewpoint constraint is then applied to the selected poses for each camera and after removing irrelevant poses, remaining samples of each camera are employed to find the final nearest pose.

3.2 Viewpoint constraint

As mentioned in the previous section, shape context descriptor alone cannot provide accurate results in human body pose estimation. Therefore, temporal information are

described in this section to enhance the accuracy of results. Majority of works have used the estimated poses of previous frames to obtain human body configuration in the current frame (such as tracking methods). In this paper, the viewpoint orientation is employed in the current frame to remove irrelevant dataset samples.

Viewpoint orientation is defined as the angle between two lines: camera's principal axis and the line that connects two hips of human body, as shown in Fig. 3. In the scene, there are two types of rotations: first is the rotation of human body around its vertical axis and second is the camera pan (rotation in a horizontal plane of it). Here, the tilt rotation is not considered due to distance of the player from the camera. Considering two rotation types and paying attention to camera frame rate, there is no significant difference between viewpoint orientation of the current frame and the previous frame. Hence, given the viewpoint orientation in the previous frame, it can be estimated in the current frame.

For dataset construction, 18 virtual cameras are employed to capture images of each pose in different views (the reason for this choice is explained in Section 4). Hence, the minimum orientation difference between virtual cameras is 20 degrees ($360^{\circ}/18=20^{\circ}$). Figure 4 shows an example of the accepted viewpoints by knowing the viewpoint orientation of previous frame. If the viewpoint orientation in previous frame is 100° ($5\times20^{\circ}=100^{\circ}$), dataset images which have been captured by virtual cameras in Positions 4, 5, and 6 according to the above-mentioned remark will be accepted. Then, the viewpoint constraint is applied on N selected samples which are obtained from the previous stage. The samples whose orientations lie in the acceptable area are selected. The acceptable area is determined by the information of previous frame. If the viewpoint orientation between camera and player in the previous frame is α degrees, the dataset images with the viewpoint orientation in the range of $[\alpha-20^{\circ}\ \alpha+20^{\circ}]$ will be chosen and the rest will be discarded.

3.3 Sharing information of several cameras

One way to enhance the accuracy of human pose estimation algorithm is incorporation of the information obtained from several cameras instead of a single camera. In order to

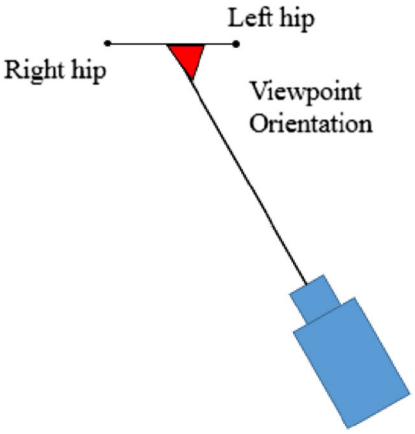

Fig. 3 Definition of viewpoint orientation

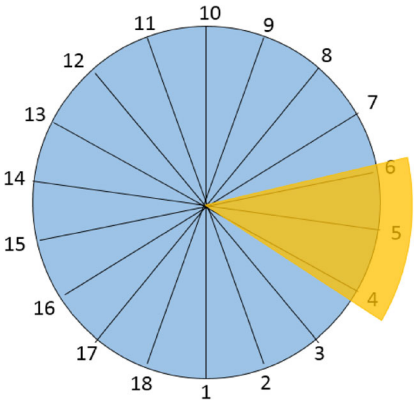

Fig. 4 An example for accepted viewpoints by knowing the viewpoint orientation of previous frame

share the information of several cameras, a majority of methods employs the calibration data which causes increment in both accuracy and speed of the system. Although various algorithms have been proposed for camera calibration, it is still a challenging problem in soccer matches [4, 24, 25].

In this paper, a method is proposed to employ the information of several cameras without using the calibration data. The method works by defining a similarity matrix which represents the likeness between pairs of 3D pose samples in the dataset. In order to calculate the similarity matrix between two 3D pose configurations, Eq. (4) is used. This equation computes the reverse sum of Euclidean distances between corresponding joints, as

$$S_{mn} = \frac{1}{\sum_{i=1}^{12} (P_{mi} - P_{ni})^2} \tag{4}$$

where P_{mi} and P_{ni} are 3D positions of joint i in pose samples of m and n, respectively. Figure 5 shows the joints which are used to calculate similarity matrix. With k pose samples in the dataset, the dimension of similarity matrix would be $k \times k$. Each 3D pose sample is determined with 12 joints and each joint in 3D space has three components. Hence, each 3D sample has 36 parameters. In other words, each 3D pose in space of 36 dimensions is specified with one vector (a 36-dimentional point).

According to previously mentioned steps, after applying the matching and viewpoint constraint stages, the nearest 3D poses available in the dataset are selected for each camera. These selections are performed independently without using information of the other cameras. Final 3D body pose is selected from 3D pose candidates which have been chosen by all cameras. As such, the candidate will be considered as the final pose if density of the candidates in the final pose is high. Therefore, to select the final 3D pose, all of the candidates introduced by cameras are considered in a 36-dimensional space. Next for each candidate, a hypersphere is considered such that its center is placed on the position of that candidate with the radius of r (r is defined by averaging distances among all of the candidates, divided by two) and then, all candidates within the hypersphere are counted according to Eqs. (5) and (6). The above

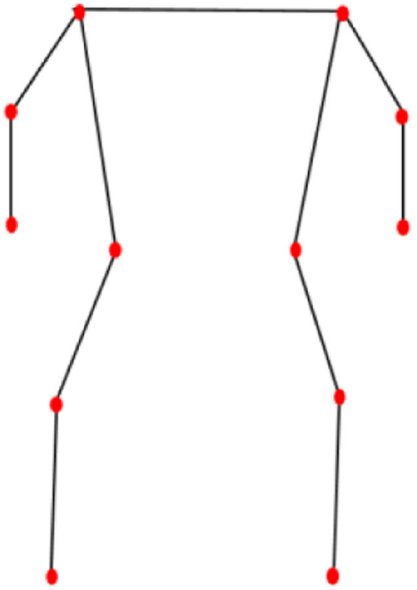

Fig. 5 Red points mark 12 joints which are used for calculation of the similarity matrix

procedure is repeated by decreasing the radius of hypersphere to $r/4$, $r/8$, and $r/16$. Next, the number of candidates are summed. The candidate which has the highest number of samples in its neighborhood is selected as the final 3D body pose. The number of candidates in the neighborhood of i'th pose is proposed to compute by

$$N_i = \sum_{k=1}^{4} \sum_{j \in C_a} T_k(S_{ij}) \tag{5}$$

$$T_k(S_{ij}) = \begin{cases} 1 & S_{ij} \geq \frac{1}{r_k} \\ 0 & S_{ij} < \frac{1}{r_k} \end{cases} \tag{6}$$

where N_i is the number of candidates in the neighborhood of i'th pose, C_a is the set of all candidates which is introduced by scenes cameras, r_k is the radius of k'th hypersphere, and S_{ij} is related to the similarity measure between samples i and j in the dataset.

4 Experimental results

In this section, at first, the process of dataset construction is explained. Next, the proposed method which was presented in Section 3 is evaluated using the KTH multi-view soccer database images [16]; and experimental results are shown. Then, the performance comparison among the proposed method and the state-of-the-art methods are given.

4.1 Dataset construction of silhouette images

In the example-based methods, construction of the dataset has an important role. Since the attention in this paper is on the sport scenes, selecting the dataset samples related to the poses of soccer players can have a significant impact on the performance of such systems. Therefore, to select the pose samples of dataset, the information of human motion capture system presented in [3] is used. In such a system, a soccer player performs prevalent actions and 3D positions of 12 markers on the body joints of the player are calculated by using the motion capture system. Having 3D positions of the joints, Matlab software is used to build the 3D reconstruction of human body. Also, as mentioned in Section 1, in an example-based method (like ours), a collection of exemplars in different poses and viewpoints is produced. This collection of samples is named as dataset of silhouettes which includes various pose configurations and viewpoints with respect to the camera. In order to create such a dataset, we have assumed that 18 virtual cameras are placed around the 3D human body. Note that these cameras are virtual and are employed by a software to produce the silhouettes.

In order to justify this choice for the number of virtual cameras (18), it is worth mentioning that the shape context feature extraction method is robust against minor out-of-plane rotations and can perform the matching despite of these rotations. On the other hand, increasing the number of virtual cameras in dataset construction grows the number of images and this increases the computational complexity and similarity among dataset images. Therefore, choosing the number of virtual cameras is a trade-off between the computational cost and the coverage of all viewpoints in the body pose. The selection of 18 virtual cameras can be a good empirical choice which is adopted in the proposed algorithm. These virtual cameras capture images from the different views. Thus, each 3D human pose produces 18 silhouettes. Then, 923 human poses with various styles are selected to form the dataset. Hence, our dataset includes 16,614 silhouettes which are used in the matching stage. Before silhouettes are generated for each 3D pose, the 3D human body is rotated such that it has 90° viewpoint orientation regard to the first camera. Figure 6 shows the construction process of the dataset.

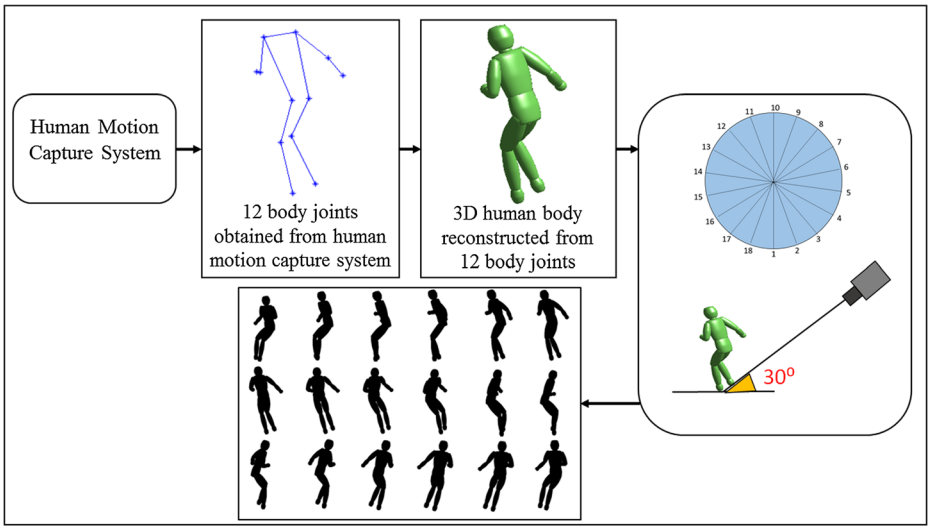

Fig. 6 Dataset construction steps

4.2 Proposed method evaluation

Before evaluating the performance of the proposed algorithm, the parameters values in Eq. (2) have to be selected. As mentioned before, this is done by a cross-validation method. After some tests, we set the parameter λ (the slope of exponential curve) to 0.3 and the T (threshold distance) to 40 pixels. These empirical numbers are used in the implementation of the proposed pose estimation algorithm. However, the number of selected samples (N) for each camera is chosen based on the number of dataset images as mentioned in the Section 3.2. According to the number of dataset images (16,614), the parameter N is empirically set to 100 in this paper.

The proposed method is evaluated on the KTH multi-view soccer database that contains images of professional soccer players which are captured from 3 views during a real match [16]. This database contains two different players and two sequences per player. In all of these video sequences, each frame is processed independently and no temporal information is used except the viewpoint information which is based on viewpoint orientation of the previous frame.

Our goal in this paper is to find the nearest pose in the dataset to the query images of several cameras. While designing a quantitative criterion to evaluate 3D pose estimation systems is necessary, it has some challenges. This is due to the infinite number of poses and the lack of suitable benchmarks for evaluation of the similarity between two different poses of the body. A criterion which is recently employed to evaluate human pose estimation systems is the *percentage of correctly estimated body parts* (PCP) [5, 6]. It appraises each 3D body pose based on the evaluation of different body parts which belong to that pose. An estimated body part is correct if

$$\frac{\left\|m_k^{'}-m_k\right\| + \left\|n_k^{'}-n_k\right\|}{2} \leq \alpha \left\|m_k^{'}-n_k^{'}\right\| \tag{7}$$

where $m_k^{'}$ and $n_k^{'}$ are ground truth 3D coordinates of end points in *k'th* part (note that ground truth is the information about the real poses of body parts which has to be provided by a standard database to objectively evaluate the human body pose estimation systems). Also, m_k and n_k are determined by the proposed method, and α is between 0.2 and 0.5.

Experimental results, which are summarized in Table I, display performance of the proposed method according to the PCP scores for α=0.5 and 0.2. The score of each correctly estimated body part is calculated and the results are shown in Table I.

Looking carefully at the results listed in Table I, a fine corollary can be observed. As shown in this table, the PCP score of lower arms is less than other parts. To justify this observation, it is notable that arms have more degrees of freedom even at daily activities in comparison with

Table I PCP scores of the proposed 3D human pose estimation method

	With viewpoint constraint		Without viewpoint constraint	
Parts	α=0.5	α=0.2	α=0.5	α=0.2
Torso	1.0000	0.9689	1.0000	0.9595
Upper arms	0.7212	0.1112	0.6736	0.0692
Lower arms	0.3454	0.0143	0.2637	0.0095
Upper legs	0.9983	0.7301	0.9829	0.6096
Lower legs	0.9627	0.1772	0.6928	0.1157
All parts	0.8055	0.4003	0.7226	0.3527

other body parts. Due to the limited number of dataset samples (923 poses are used in dataset construction), the whole positions of arms in input images are not found in the dataset. On the other hand, it is worth mentioning that the lower part of arms and legs has a less PCP score in comparison with the upper part. A reason for this phenomenon is that the generated error in the pose estimation of upper part changes the location of the common joint (elbow in arm and knee in leg) and this variation in the position of joint reduces the PCP score in the lower part.

Another discussion which is worth mentioning about the results is that some of the errors shown in Table I do not root from the lack of accuracy in the proposed method. In fact, the PCP score is not fully compatible with subjective criterion and has some differences. On the other hand, because of calibration error, the ground truth information of KTH database in end points of body parts does not have sufficient accuracy and there are always some errors in the calculations. Also, despite the correct estimation of the final pose, the PCP score reports some errors due to the lack of posture associated with the query image in dataset samples.

The experimental results with and without viewpoint constraint are also compared in Table I. It clarifies that the use of viewpoint constraint in our proposed algorithm has a significant impact on the accuracy of human body pose estimation. Figure 7 displays the estimated 3D poses for six different frames. It should be mentioned that the proposed method

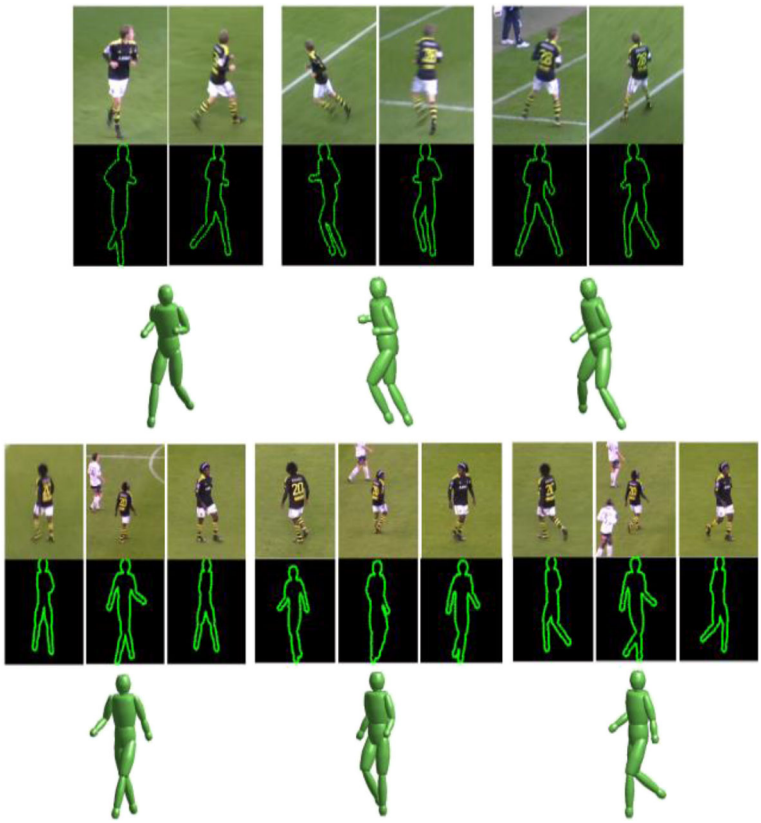

Fig. 7 Multi-camera 3D pose estimation for 6 different frames; first row: query images, second row: nearest matched image in dataset, third row: 3D estimated pose

is applicable in multi-camera environments. Figure 7 shows the results of the proposed method for cases with 2 and 3 cameras. It is observable that the proposed algorithm is able to find the correct nearest pose from the dataset, in the most cases.

As mentioned before, in order to find the best match between the query image and dataset samples, the histograms of shape context in marked points of the query image and dataset images are compared. The dataset samples whose histograms have the most similarity to histograms of the query image are selected as the best matches. Figure 8 shows the best match for query images. The red-marked points in Fig. 8c show the points whose histograms of shape context have greater differences between the query image and the matched one. As depicted in Fig. 8c, these red-marked points involve areas of the matched image where some differences between the query image and the matched one are obvious. This property can be utilized in the optimization stage. In other words, instead of applying the optimization algorithm to all pose configurations, only the parts of body pose are considered in the optimization stage that their positions have been displayed by red-marked points. This method can have a significant impact on the speed of optimization stage in the human body pose estimation systems.

As mentioned in Section 3.1, the number of clusters in the K-means algorithm, which is used to categorize body edge points, is set to 100. It should be noted that the selection of 100

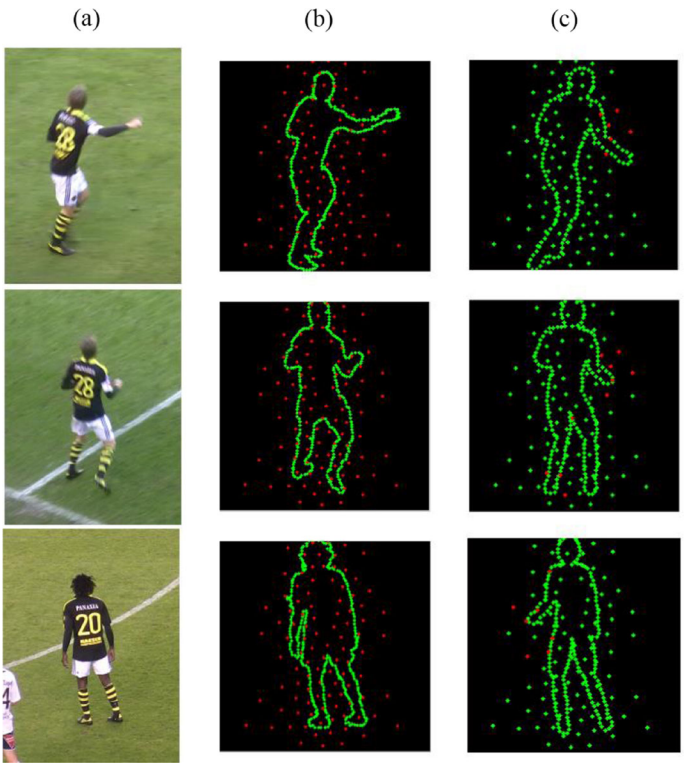

Fig. 8 **a**. Query images; **b**. Foreground-background segmented of query image; **c**. Nearest image in the dataset (areas of matched image which are different from query image are shown by red-marked points)

Table II Impact of variation in the number of cluster points on the algorithm accuracy and required run-times

Number of cluster points	20	30	40	50	60	70	80	90	100	110
Require run-time (in sec)	4.4	6.3	7.9	9.6	11.5	13.3	14.8	16.6	18.5	20.6
Average PCP score	80.32	80.02	79.39	80.32	80.4	80.41	79.97	80.26	80.55	80.49

clusters in the K-means algorithm is an empirical choice. In fact, we tested other choices for the number of clusters to clarify the effect of this selection on the accuracy of the method and also its impact on the computational complexity of the algorithm. The results of these experiments are shown in Table II, when the number of cluster points is varied from 20 to 110.

It is inferable from Table II that changing the number of clusters in the K-means algorithm does not considerably affect the accuracy of the method based on PCP scores. But, the computational complexity of the algorithm is significantly influenced by this selection. For instance, running the algorithm with 20 clusters is almost 14 seconds faster than running it with 100 clusters. Based on these results, at the first glance, we could conclude that it is logically better to select less cluster numbers in the K-means algorithm to avoid unnecessary complexity. On the other hand, choosing more number of clusters in the K-means algorithm leads to better localization of places which introduces error in the matching process (see Fig. 8c). Hence, 100 clusters are selected for K-means algorithm to provide the information that can be used in the optimization stage as explained above.

4.3 Performance comparison

It is worth mentioning that the performance of various pose estimation algorithms can be compared with each other if the conditions and constraints (imposed on the setup) are similar. For example, we cannot compare the results of a pose estimation method which is based on calibrated cameras with those which do not utilize the calibration information. As it is revealed in Section 1, one of the main contributions of this paper is the information sharing of several cameras without using any calibration data. Since in the most demanded applications (such as soccer matches), the calculation of calibration data for the cameras is a challenging task, our proposed algorithm is potentially more practical in real applications. As such, we cannot directly compare the results of our proposed algorithm, which is based upon uncalibrated cameras, with ones that employ calibration data because such a comparison would be logically unfair. In other words, the assumption taken by these two methods (with/without calibration data) is completely different and therefore the results are

Table III Comparative results between our proposed algorithm and methods presented in [5] and [16], based on PCP score

Methods	Proposed method	Algorithm [5]	Algorithm [16]
Parts of the body			
Torso	1.00	1.00	1.00
Upper arms	0.7212	0.60	0.89
Lower arms	0.3454	0.35	0.68
Upper legs	0.9983	1.00	1.00
Lower legs	0.9627	0.90	0.99
All Parts	0.8055	0.77	0.912

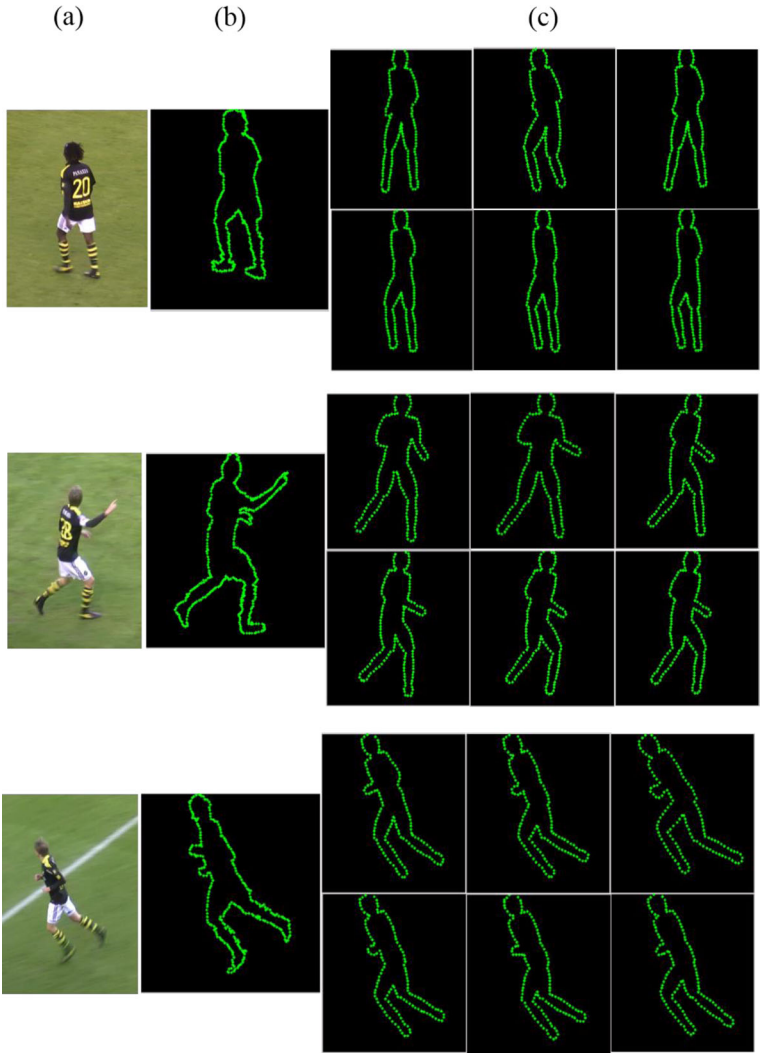

Fig. 9 **a**. Query image; **b**. Foreground-background segmented of query image; **c**. First row: top three nearest images in the dataset for the shape context matching proposed in [2]; second row: top three nearest images in the dataset for the shape context matching proposed in Section 3.1

incomparable. On the other hand, there is another condition to make the results of different algorithms to be comparable. This condition states that the algorithms have to use a similar database; otherwise, comparison of results will be ineligible. In summary, to choose a proper state-of-the-art pose estimation algorithm which is comparable with the proposed method, care should be taken about the above-mentioned limitations. But then, in current state-of-art pose estimation methods, one can seldom find the algorithm which has similar conditions with the proposed method in this paper (which estimates 3D poses of soccer players using several uncalibrated cameras). Hence, choosing proper methods for comparison is a challenging task.

Despite the above-mentioned discussions about the unfairness of comparisons between the proposed method with other state-of-the-art ones, but in order to have some comparisons, papers

[5, 16], which have been published in the same context (human 3D pose estimation in soccer game) by using the KTH football database for evaluation, are selected for this purpose. Table III shows these comparative results in terms of PCP scores. However, we reemphasize here that such comparisons are practically irrational because the methods proposed in the mentioned papers have been constructed based on calibration information while the proposed scheme does not use such information. Also, even if calibration information was employed by a method (like in [5, 16]), it would have to be updated in each frame due to variations in the internal and external parameters of cameras. These variations are created because the cameras have to track the fast motions of players and ball in the scene. Thus, the higher PCP scores in [16] do not necessarily indicate that such method outperforms the algorithm proposed in this paper.

We perform another comparison between the proposed shape context matching (introduced in Section 3.1) and the shape context matching presented in [2] which employs optimal assignment algorithms to solve the correspondence between edge points on the two shapes. The results of qualitative comparison are shown in Fig. 9 for three images of KTH multi-view soccer database. Experimental results reveal that two shape context matching strategies have approximately similar outputs.

5 Conclusion

A novel method for recognition of human body poses is proposed. In order to estimate 3D poses, a rich dataset which have been captured from soccer players are considered. The shape context descriptor based on 100 marked points are used for silhouette matching. This descriptor simplifies the matching process. It can also mark the areas where the query image and the nearest selected pose differ from each other. The temporal information is used without employing the pose configuration in the previous frame. Since no calibration information is employed, a method for incorporation of several cameras is also proposed. The proposed method finds the nearest pose in the dataset. In our future work, an optimization stage based on the proposed shape context descriptor is going to be used to increase the accuracy of pose estimation. Since the methods employing just silhouette images cannot distinguish poses with similar silhouettes, feature descriptors (which utilize texture of the body region) are going to be incorporated. We expect that combination of several distinct features yields more accurate results.

References

1. Agarwal A, Triggs B (2006) Recovering 3D human pose from monocular images. IEEE Trans Pattern Anal Mach Intell 28(1):44–58
2. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522
3. Burenius M, Sullivan J, Carlsson S, Halvorsen K (2011) Human 3D motion computation from a varying number of cameras. Scandinavian Conference on Image Analysis (SCIA):24–35
4. Burenius M, Sullivan J, Carlsson S (2011) Motion capture from dynamic orthographic cameras. IEEE International Conference on Computer Vision Workshops (ICCV Workshops): 1634-1641
5. Burenius M, Sullivan J, Carlsson S (2013) 3D pictorial structures for multiple view articulated pose estimation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR): 3618-3625
6. Eichner M et al (2012) 2D articulated human pose estimation and retrieval in (almost) unconstrained still images. Int J Comput Vis 99(2):190–214
7. Felzenszwalb PF, Huttenlocher DP (2005) Pictorial structures for object recognition. Int J Comput Vis 61(1):55–79

8. Gall J, Rosenhahn B, Brox T, Seidel HP (2010) Optimization and filtering for human motion capture. Int J Comput Vis 87(1–2):75–92
9. Germann M, Hornung A, Keiser R, Ziegler R, Würmlin S, Gross M (2010) Articulated billboards for video-based rendering. Comput Graph Forum 29(2):585–594
10. Germann M, Popa T, Ziegler R, Keiser R, Gross M (2011) Space-time body pose estimation in uncontrolled environments. International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT):244-251
11. Hofmann M, Gavrila DM (2012) Multi-view 3D human pose estimation in complex environment. Int J Comput Vis 96(1):103–124
12. Holte MB, Tran C, Trivedi MM, Moeslund TB (2012) Human pose estimation and activity recognition from multi-view videos: comparative explorations of recent developments. J Sel Topics Signal Process 6(5):538–552
13. Howe NR (2007) Silhouette lookup for monocular 3D pose tracking. Image Vision Comput 25(3):331–341
14. Ji X, Liu H (2010) Advances in view-invariant human motion analysis: a review. IEEE Trans Syst Man Cybern Part C 40(1):13–24
15. Kazemi V, Sullivan J (2012) Using richer models for articulated pose estimation of footballers. IEEE British Machine Vision Conference (BMVC):1-10
16. Kazemi V, Burenius M, Azizpour H, Sullivan J (2013) Multi-view body part recognition with random forests. IEEE British machine vision conference (BMVC)
17. Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vis Image Underst 81(3):231–268
18. Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. Comput Vis Image Underst 104(2–3):90–126
19. Moeslund TB, Hilton A, Krüger V, Sigal L (2011) Visual analysis of humans: looking at people. Springer
20. Mori G, Malik J (2006) Recovering 3D human body configurations using shape contexts. IEEE Trans Pattern Anal Mach Intell 28(7):1052–1062
21. Poppe R (2007) Vision-based human motion analysis: an overview. Comput Vis Image Underst 108(1–2):4–18
22. Shakhnarovich G, Viola P, Darrell T (2003) Fast pose estimation with parameter-sensitive hashing. Ninth IEEE International Conference on Computer Vision (ICCV):750-757
23. Sigal L, Black M (2010) Guest editorial: state of the art in image- and video-based human pose and motion estimation. Int J Comput Vis 87(1–2):1–3
24. Thomas GA (2006) Real-time camera pose estimation for augmenting sports scenes. IET Conf Publ 2006 (CP516): 10–19
25. Thomas G (2007) Real-time camera tracking using sports pitch markings. J Real-Time Image Proc 2(2):117–132
26. Yang Y and Ramanan D (2011) Articulated pose estimation with flexible mixtures-of-parts. IEEE conference on Computer Vision and Pattern Recognition (CVPR):1385–1392
27. Yi Y, Ramanan D (2013) Articulated human detection with flexible mixtures of parts. IEEE Trans Pattern Anal Mach Intell 35(12):2878–2890

Reza Afrouzian Received the B.Sc. and M.Sc. degrees in Electrical Engineering from Azad University of Tabriz, Iran, and University of Zanjan, Iran in 2006 and 2010, respectively. Currently, he is a student at University of Tabriz, Tabriz, Iran. His research interests include 3D computer vision, human pose estimation, human motion capture and sport scenes analysis.

Hadi Seyedarabi Received B.S. degree from University of Tabriz, Iran, in 1993, the M.S. degree from K.N.T. University of technology, Tehran, Iran in 1996 and Ph.D. degree from University of Tabriz, Iran, in 2006 all in Electrical Engineering. He is currently an associate professor of Faculty of Electrical and Computer Engineering in University of Tabriz, Tabriz, Iran. His research interests are Image Processing, Computer Vision, Video Coding, Human-Computer Interaction, Facial Expression Recognition and Facial Animation.

Shohreh Kasaei Received the B.Sc. degree from the Department of Electronics, Faculty of Computer and Electrical Engineering, Isfahan University of Technology, Iran, in 1986. She worked as research assistance in Amirkabir University of Technology, for 3 years. She then received the M.Sc. degree from the Graduate School of Engineering, Department of Electrical and Electronic Engineering, University of the Ryukyus, Japan, in 1994, and the Ph.D. degree from Signal Processing Research Centre, School of Electrical and Electronic Systems Engineering, Queensland University of Technology, Australia, in 1998. She was awarded as the best graduate student in engineering faculties of University of the Ryukyus, in 1994, the best Ph.D. student studied in overseas by the ministry of Science, Research, and Technology of Iran, in 1998, and as a distinguished researcher of Sharif University of Technology, in 2002 and 2010, where she is currently a full professor. She is the director of Image Processing Lab (IPL). Her research interests are in image/video processing with primary emphasis on 3D computer vision, 3D pose estimation, 3D object tracking, 3D model building, multiresolution texture analysis, scalable video coding, image retrieval, video indexing, face recognition, hyperspectral change detection, video restoration, fingerprint authentication, and watermarking.