A Robust Integrated Framework for Segmentation and Tracking

Prabhu Kaliamoorthi and Ramakrishna Kakarala

Nanyang Technological University,
Singapore
{prab0009,ramakrishna}@ntu.edu.sg

Abstract. Recent studies on human motion capture (HMC) indicate the need for a likelihood model that does not rely on a static background. In this paper, we present an approach to human motion capture using a robust version of the oriented chamfer matching scheme. Our method relies on an MRF based segmentation to isolate the subject from the background, and therefore does not require a static background. Furthermore, we use robust statistics and make the likelihood robust to outliers. We compare the proposed approach to the alternative methods used in recent studies in HMC using the Human Eva I dataset. We show that our method performs significantly better than the alternatives despite of not assuming a static background.

1 Introduction

Human motion capture (HMC) is a challenging and an important area of research in the vision community. Though there is a large body of literature on HMC [18,7,22,1,21,19,8,15,12], most recent methods on HMC [20,7,8] rely on either chroma keying or a static background for background subtraction. Consequently, they are not suitable for outdoor motion capture where lighting change and non static background makes the assumptions invalid. This is observed to be a critical problem for HMC in recent studies [20]. In order to overcome this limitation, some studies augment the visual data with other sensors such as IMU's [16]. Recent studies [20] indicate that these problems can be mitigated by designing a suitable cost function that does not make strong assumptions about a static background and yet would enable tracking with an acceptable computational overhead. We address this aspect of HMC systems in this paper.

Our approach draws inspiration from oriented chamfer matching (OCM) in object recognition [17] and exemplar based pose estimation [14]. Shotton et al. [17] note that OCM has high discriminative properties for articulated objects and it is robust to noise. Though these properties have been exploited in areas such as exemplar based pose estimation [14] and articulated hand tracking [23], these techniques have not been applied to model based HMC.

In this paper, we formulate a cost function for HMC using robust oriented chamfer matching. The analysis by synthesis framework that is widely used for HMC minimizes an energy functional that measures the disparity between the observed image and the synthesized output. The energy functional is generally [8,20] composed of a model to

R. Klette, M. Rivera, and S. Satoh (Eds.): PSIVT 2013, LNCS 8333, pp. 443–453, 2014.

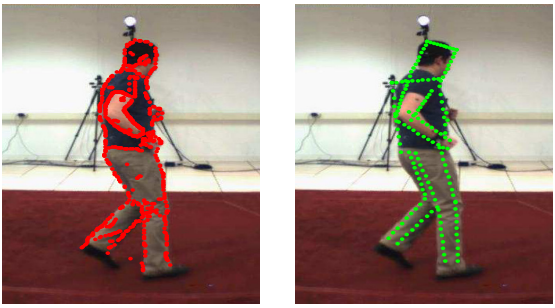


Fig. 1. Edge fragments from the observed image (red) and synthesized output (green) superimposed on the observed image.

observation and an observation to model matching (bidirectional) terms. Our method extracts a set of edge fragments from the observation and the synthesized output, and minimizes a bidirectional oriented chamfer distance between the two. Since our observation is a general set of oriented edge fragments, our method does not make any assumption about a static background. Furthermore, since the observation could include significant noise, we extend the standard oriented chamfer distance with a robust function. This is combined with an MRF based shape prior which segments out the subject from the background to ensure an acceptable computational overhead.

Our approach is compared with the alternative approaches used in recent publications [20,8] using the data from the Human Eva I dataset. We demonstrate with quantitative results, that our method on an average reduces the tracking error by up to 15%, despite of not making limiting assumptions such as a static background.

2 Previous Work

Local optimization and optical flow is used for HMC by studies such as [6,3]. These methods are known to get stuck in incorrect hypothesis [9]. Hence they are not suitable for situations where the model is not exact, or the image observation is uncertain. Some studies [1,4] directly infer the human pose from a set of image features. However, these methods are known to be sensitive to the appearance of the subject and to require extensive training. Exemplar based methods [15,14] have been in use for pose estimation. However, they are not accurate enough for motion capture. Model free approaches to HMC [2,21], loosely assemble human parts to a plausible pose. However, anatomically correct reconstruction of the human motion cannot be ensured by these methods.

Widely used methods such as [7] rely on a kinematic skeleton fleshed with parametric surfaces and achieve motion capture using stochastic search. Gall et al. [8] describe a multi-pass solution that performs initial tracking with a stochastic search method, which is refined with a smoothing filter and local optimization. They use a deformable surface model of the human for tracking. Sidenbladh et al. [18] describe a complete generative framework to model based human motion capture using sequential Monte Carlo tracker. Our method can be incorporated within their framework.

Kohli et al. [13] present an integrated framework for segmentation and pose estimation. The MRF based shape prior we use in this work is similar to [13]. However, their approach uses background and foreground statistics. Hence their approach is not suitable for outdoor situations where the statistics are expected to change due to lighting changes. Since we don't explicitly use background and foreground statistics for segmentation, our approach is insensitive to lighting changes that are expected in outdoor situations such as those considered in [16].

3 Oriented Chamfer Matching

3.1 Overview

Oriented chamfer matching is a technique used in object recognition for matching edge fragments [17,14,11]. Let $\mathcal{P}$ be the space $\mathbb{R}^2 \times [0, \pi)$ that denote the coordinates of an edge fragment along with the edge orientation. Let the respective sets of observation and synthesized edge fragments be $O = \{o_i \in \mathcal{P}\}$ and $S = \{s_i \in \mathcal{P}\}$. The oriented chamfer distance between the two is defined as

$$d_{OCM}(O, S) = \frac{1}{|O|} \sum_{o_i \in O} \underset{s_i \in S}{min}\, d_{\mathcal{P}}(o_i, s_i) \tag{1}$$

where $|O|$ is the cardinality of the set O, and $d_{\mathcal{P}} : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$, is a distance metric on $\mathcal{P}$ described below

$$d_{\mathcal{P}}(A, B) = \sqrt{\frac{\parallel A_x - B_x \parallel_2^2}{\lambda} + \parallel A_a - B_a \parallel_a^2} \tag{2}$$

where $\parallel A_a - B_a \parallel_a = min(|A_a - B_a|, \pi - |A_a - B_a|)$ is a distance metric for the angular component. Let W be a warping function parametrized by x, then the optimal parameter $\hat{x}$ that aligns the synthesized edge fragments S with the observation O can be obtained by minimizing the distance d_{OCM} as below

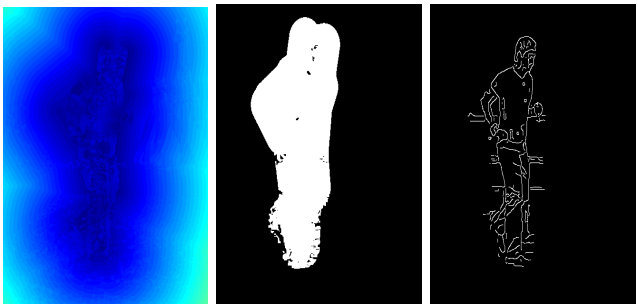
$$\hat{x} = \arg\min_{x} d_{OCM}(O, W(S; x)) \tag{3}$$

When matching edge fragments, the warping function W is generally a planar affine warp, for example, $x \in SE(2)$. In this paper, we consider the warp W to be a nonlinear map that assembles the articulated model in its pose in $\mathbb{R}^3$ and projects it to calibrated cameras to obtain the synthesized output S. We treat x to be the parameters of the articulated model.

Furthermore, as observed in earlier studies [10], human pose tracking requires a bidirectional matching. Hence we extend the oriented chamfer matching scheme for HMC by defining a bidirectional likelihood term for the camera C as below

$$\begin{aligned} -lnP(O; X, C) = {} & d_{OCM}(O, W(X, C)) \\ & + d_{OCM}(W(X, C), O) \end{aligned} \tag{4}$$

The overall likelihood is expressed as the mean from all the cameras used by the multi-view HMC system. The above formulation of the likelihood performs well for HMC. In this paper, we extend this with a robust formulation which is described next.



(a) Cost (b) Mask (c) Canny Edges

Fig. 2. Segmentation using oriented chamfer distance. The sub-figure a) shows the likelihood term used by the MRF based segmentation technique, b) shows the segmented foreground region, and c) shows the Canny edges extracted from the foreground region that are used for tracking.

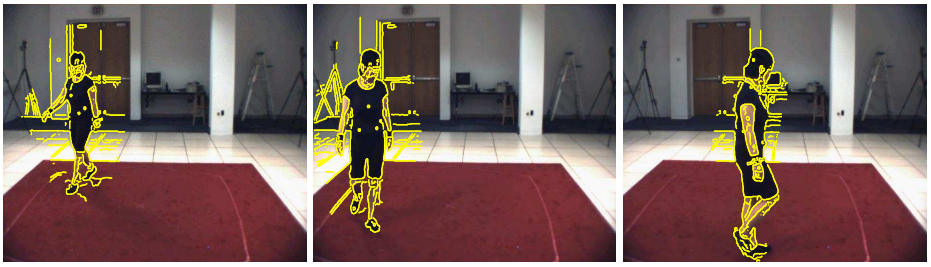


Fig. 3. Canny edge fragments obtained by the MRF based segmentation for the Subject 1 walking sequence superimposed on the input image. It can be observed that the subject and a small region around the subject is carved out of the input.

3.2 Robust OCM

The likelihood described in last section assumes that an error free observation O representing the subject is available. This could be achieved in practice by using a silhouette obtained by background subtraction. However, this would result in a method that does not work well in outdoor situations. Hence in this work we use an MRF based segmentation technique that isolates the subject from the background, which we describe next.

Let Z_i ($i \in \{1, \ldots, n\}$) represent the random variables corresponding to the individual pixels. The segmentation task is formulated as the problem of classifying the pixels into two coherent regions represented by the labels $\mathcal{L} \in \{f, b\}$. Spatial coherency between the segmented regions is ensured by a neighborhood influence function. Let $\mathcal{N}_i$ represent a set of neighbors of Z_i, which we consider to be the four adjacent pixels. The posterior probability of the set of pixels in the image Z taking a configuration $z \in \mathcal{L}^n$, is expressed as

$$Pr(Z = z) \propto \prod_{i \in \{1,\ldots,n\}} \Big(\phi(z_i) \prod_{j \in \mathcal{N}_i} \psi(z_i, z_j) \Big) \tag{5}$$

where $\phi(z_i)$ is the data likelihood term that indicates how likely is a given pixel to belong to a specific label, and $\psi(z_i, z_j)$ is the prior term that enforces spatial coherency.

It can be shown [5] that finding the MAP configuration is equivalent to finding the minimum energy configuration

$$E(z) = \sum_{i \in \{1,\ldots,n\}} -log\ \phi(z_i) + \sum_{\substack{i \in \{1,\ldots,n\} \\ j \in \mathcal{N}_i}} -log\ \psi(z_i, z_j) \tag{6}$$

The minimum energy configuration of Z can be efficiently obtained using graph cuts [5]. In this paper, we consider the prior term $\psi(z_i, z_j)$ to be the generalized Potts model [13]. The likelihood using the oriented chamfer distance can be expressed as

$$-log\ \phi(z_i = f) = \min_{y \in \Pi(X_{t-1})} d_{\mathcal{P}}(\mathcal{P}(z_i), y) \tag{7}$$

where $\Pi(X_{t-1})$ is the synthesized model from the previous time frame, and the operator $\mathcal{P}(z_i)$ returns the 3-tuple $\in \mathcal{P}$ for the pixel z_i. We fix $\phi(z_i = b)$ to be a constant in this work. It should be noted that our framework does not make use of the appearance for the segmentation and it only makes use of the oriented chamfer distance. As a result, our segmentation is robust to lighting changes that are expected to happen in outdoor scenario.

Figure 2 shows the segmentation using the MRF. The cost $-log\ \phi(z_i = f)$ is shown in Figure 2a. The mask obtained using graph cut is shown in Figure 2b. Figure 2c shows the Canny edge segments in the foreground region. Figure 3 shows the segmented Canny edge fragments for the Subject 1 walking sequence.

3.3 Robust Likelihood

It can be observed from Figure 2c, that the edge fragments obtained from segmentation include considerable noise. Consequently, the likelihood described in Section 3.1 can be biased by the noise. We extend the likelihood model in order to overcome this. The likelihood described in Section 3.1 is composed of two parts. The first part measures the fit of the observation to the synthesized model, and the second term measures the fit of the synthesized model to the observation. The observation edge fragments O include substantial noise as a result of the segmentation. However, synthesized edge fragments remain the same as before. Hence we modify the first term to use a robust function that is less sensitive to noise than the squared error. Formally,

$$d_{ROCM}(O, S) = \frac{1}{|O|} \sum_{o_i \in O} \rho\Big(\inf_{s_i \in S} d_{\mathcal{P}}(o_i, s_i) \Big) \tag{8}$$

where $\rho(x)$ is the Geman-McClure error function described below.

$$\rho(x) = \frac{x^2}{x^2 + \alpha^2} \tag{9}$$

The robust likelihood is formally expressed as

$$\begin{aligned} -lnP(O;X,C) = d_{ROCM}(O, W(X,C)) \\ + d_{OCM}(W(X,C), O) \end{aligned} \tag{10}$$

We refer to the oriented chamfer matching scheme that uses the above likelihood and the MRF based segmentation described in Section 3.2 as robust oriented chamfer matching (**ROCM**).

4 Experiments and Results

In this section, we compare the proposed method with the alternative methods used for HMC in recent publications [16,8,20]. The Human Eva I dataset that provides video and motion capture data was used for validating the proposed approach. We used input from 3 RGB cameras and 2 grayscale cameras in all the sequences except the Subject 3 jogging and walking sequence. For the Subject 3 jogging and walking sequence, we used an additional camera input, since we found the segmented video had a blind spot near a corner, where nearly 60% of the foreground was invisible, resulting in a tracking failure.

Initialization was performed with the ground truth provided with the Human Eva I dataset as commonly done [20] in recent work. We registered a set of markers provided by the ground truth to the model in the first frame in order to measure the tracking error. Mean Euclidean distance (in mm) between the registered markers and the ground truth in subsequent frames are reported as the tracking error. A constant position model [7] was used as a motion prediction strategy for tracking. We used a kinematic model with 25 degrees of freedom and 10 links. The rigid links were fleshed with truncated cones with elliptical cross section [20].

We obtained silhouettes by background subtraction using the background statistics provided with the Human Eva I dataset. The likelihoods used in current methods [20,8] require the silhouettes. However, our method **ROCM** does not need the silhouettes. Our method was compared with two common likelihoods used with model based motion capture systems. The study in [20] finds bidirectional silhouette (**BiS**) based likelihood to produce the best results. We briefly describe the method here, **BiS** is composed of two terms,

$$\begin{aligned} -lnP_{\mathbf{BiS}}(Sil;X,C) = \frac{1}{|\sum_p Sil(p)|} \sum_p Sil(p)(1 - \Pi(p)) \\ + \frac{1}{|p \in \Pi(X)|} \sum_{p \in \Pi(X)} (1 - Sil(p)) \end{aligned} \tag{11}$$

where Sil is the smoothed silhouette and $\Pi(X)$ is the synthesized projection. The first terms in Eq. (11) penalizes the set of pixels in the silhouette that are not explained by the synthesized projection and the second term penalizes the set of pixels in the synthetic projection that are not explained by the silhouette. This is similar to the bidirectional matching we perform.

(a) Subject1 (b) Subject2 (c) Subject3

Fig. 4. Image features used for tracking. The silhouette of the subjects are shown as the foreground region, appearance is shown by the intensity values and the oriented edge fragments are shown as red dots.

Appearance (**App**) based likelihood [16,8] is used in recent studies, since it is more informative than likelihoods such as **BiS**. It uses appearance cues in addition to the silhouette (**BiS**). The likelihood requires the segmentation of the projection $\Pi(X)$ into those of the individual body parts $\Pi_i(X)$, which takes occlusion into consideration. The subscript i in $\Pi_i(X)$ indicates that it is the projection for the part i. In the initial frame, the statistics for the part i, Ψ_i is extracted and is assumed to remain unchanged. The appearance based likelihood is formally expressed as below using the statistics for the parts Ψ_i

$$-lnP_{\textbf{App}}(I;X,C) = \frac{1}{|parts|}\sum_{i\in parts} D_B(\Psi_i||\Psi(\Pi_i(X),I)) \tag{12}$$

where $D_B(X||Y)$ is the Bhattacharya distance between the two distributions, and $\Psi(\Pi_i(X),I)$ is the statistics extracted from the image I in the region corresponding to $\Pi_i(X)$. The appearance statistics were represented as histograms. In order to make the likelihood invariant to lighting change, similar to [8], we use the statistics from the a and b channels alone of the CIELab representation of the image.

Figure 4 shows the image features used for tracking for the three subjects in Human Eva I. The silhouette of the subjects are shown by the foreground region. It can be observed that the shadow of the subject is part of the silhouette region. The appearance is shown by the intensity values. The set O of observation edge fragments are shown as red dots.

Figure 5 shows the time and ensemble averaged tracking error and deviation of 5 different runs for each sequence. It can be noticed that for Subject 1, in general **BiS** has a very high error. We found the reason for this to be the poor model fit for the subject, which results in a highly multimodal likelihood. Using appearance cues (**App**), improves the tracking performance. Our method (**ROCM**) has the least tracking error. The tracking performance of Subject 2 is equally good with **BiS** and **App**, since Subject 2 is well segmented from the background. Our method performs better than the other

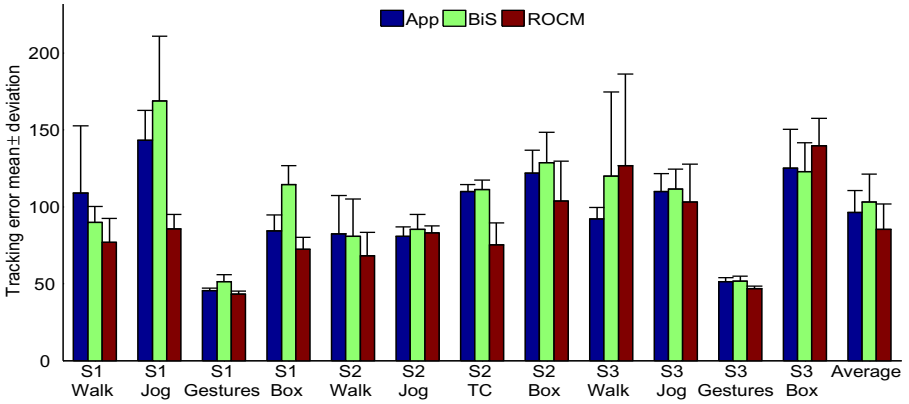


Fig. 5. Tracking error for the twelve sequences compared. Overall mean error for all the sequences compared is shown in the last group. It can be observed that ROCM has the least overall error and it perform better than the other configurations in 9 out of 12 sequences.

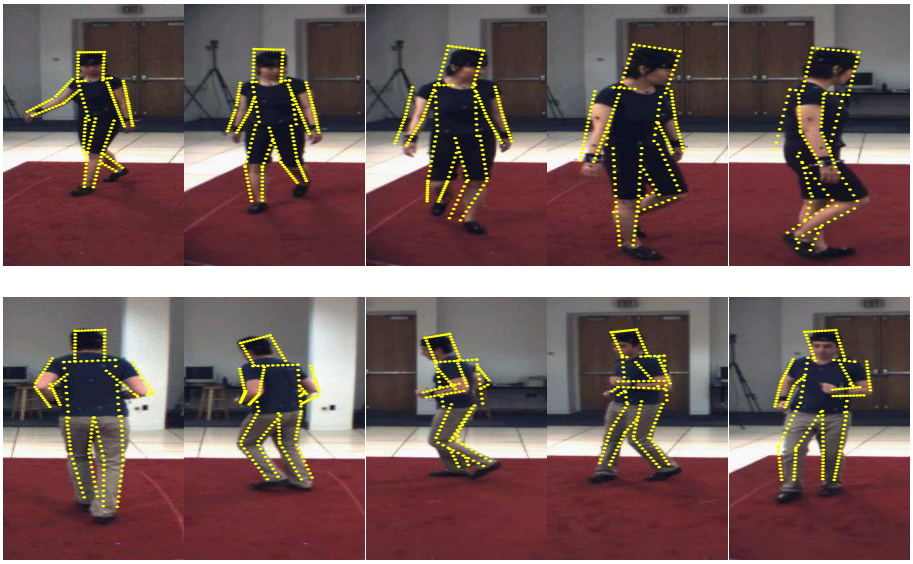


Fig. 6. Tracking results for the Subjects 1 (top row) and 2 (bottom row) using the configuration **ROCM** (without using silhouettes). The best result out of five runs for each sequence is shown.

two methods for this subject too. Silhouettes obtained from Subject 3 are highly reliable, but the subject is completely dressed in black. As a result there are hardly any edges visible due to depth discontinuity (as observed in 4c). Hence, for this subject our method performs worse for two sequences.

Table 1. Runtime and overall error for the three configurations compared. It can be observed that the proposed method has a lower error and runtime.

	App	**BiS**	**ROCM**
Runtime per frame (sec)	29.3	26	13 (+3.3)
Overall error (mm) mean $\pm$ std	96 ± 14	103 ± 18	85 ± 16

The configuration **ROCM** has a lower error than **BiS** and **App** in 9 out of 12 sequences. Overall mean error for **ROCM** (shown in Table 1) is more than 15% and 10% lower than that of **BiS** and **App** respectively. The best tracking results for the Subject 1 walking and the Subject 2 jogging sequences, obtained using the configuration **ROCM** is shown in Figure 6. Figure 7 shows two situations where the proposed method switches to an incorrect hypothesis. It can be observed that ROCM at times gets misled by edges that appear to be from the subject. However, it did not result in a complete tracking failure. We believe that such problems can be reduced by using a more accurate model of the subject.

Our HMC framework is currently implemented in Matlab. However, critical computational blocks such as the evaluation of the cost function are implemented in C. We further use data structures such as KD trees [23] to reduce the complexity of the robust oriented chamfer matching scheme. We ran the code on a standard PC and the processing time. It can be observed that our method is significantly faster than the other two methods. The preprocessing time necessary for the MRF segmentation of the proposed method is shown separately (3.3 seconds).

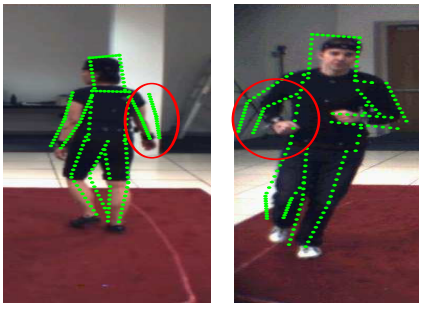


Fig. 7. Incorrect tracking results obtained by using ROCM. It can be observed that ROCM gets misled by the presence of edges that are similar to those from the human subject.

5 Conclusion

In this paper, we describe a robust oriented chamfer matching scheme for human motion capture. The method relies on an MRF based segmentation technique to segment out the subject from the background. Furthermore, we extend the well-known OCM scheme

with robust statistics. We compare our approach to the alternative methods used in HMC in recent studies, and show that our method reduces the tracking error by up to 15%, despite of not using background statistics. In the future, we hope to extend our method to scenarios where multiple subjects interact.

References

1. Agarwal, A., Triggs, B.: 3d human pose from silhouettes by relevance vector regression. In: CVPR, pp. 882–888 (2004)
2. Andriluka, M., Roth, S., Schiele, B.: Monocular 3d pose estimation and tracking by detection. In: CVPR, pp. 623–630 (2010)
3. Ballan, L., Cortelazzo, G.M.: Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In: 3DPVT (June 2008)
4. Bo, L., Sminchisescu, C.: Twin Gaussian processes for structured prediction. IJCV 87, 28–52 (2010)
5. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient n-d image segmentation. IJCV 70, 109–131 (2006)
6. Bregler, C., Malik, J.: Tracking people with twists and exponential maps. In: CVPR, pp. 8–15 (1998)
7. Deutscher, J., Reid, I.: Articulated body motion capture by stochastic search. IJCV 61, 185–205 (2005)
8. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.-P.: Optimization and filtering for human motion capture. IJCV 87, 75–92 (2010)
9. Gall, J., Stoll, C., de Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.-P.: Motion capture using joint skeleton tracking and surface estimation. In: CVPR, pp. 1746–1753 (2009)
10. Gavrila, D.M., Davis, L.S.: 3-d model-based tracking of humans in action: a multi-view approach. In: CVPR, pp. 73–80 (1996)
11. Kaliamoorthi, P., Kakarala, R.: Directional chamfer matching in 2.5 dimensions. IEEE Signal Processing Letters, 1–4 (in press, 2013)
12. Kaliamoorthi, P., Kakarala, R.: Parametric annealing: A stochastic search method for human pose tracking. Pattern Recognition 46, 1501–1510 (2013)
13. Kohli, P., Rihan, J., Bray, M., Torr, P.H.: Simultaneous segmentation and pose estimation of humans using dynamic graph cuts. IJCV 79, 285–298 (2008)
14. Liu, M.-Y., Tuzel, O., Veeraraghavan, A., Chellappa, R.: Fast directional chamfer matching. In: CVPR, pp. 1696–1703 (2010)
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, Part III. LNCS, vol. 2352, pp. 666–680. Springer, Heidelberg (2002)
16. Pons-Moll, G., Baak, A., Gall, J., Leal-Taixé, L., Müller, M., Seidel, H.-P., Rosenhahn, B.: Outdoor human motion capture using inverse kinematics and von Mises-Fisher sampling. In: ICCV, pp. 1243–1250 (2011)
17. Shotton, J., Blake, A., Cipolla, R.: Multiscale categorical object recognition using contour fragments. TPAMI 30, 1270–1281 (2008)
18. Sidenbladh, H., Black, M.J., Fleet, D.J.: Stochastic tracking of 3D human figures using 2D image motion. In: Vernon, D. (ed.) ECCV 2000. LNCS, vol. 1843, pp. 702–718. Springer, Heidelberg (2000)
19. Sigal, L., Balan, A.O., Black, M.J.: Combined discriminative and generative articulated pose and non-rigid shape estimation. In: NIPS, pp. 1337–1344 (2007)

20. Sigal, L., Balan, A.O., Black, M.J.: HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV 87, 4–27 (2010)
21. Sigal, L., Isard, M., Haussecker, H.W., Black, M.J.: Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. IJCV 98, 15–48 (2012)
22. Sminchisescu, C., Triggs, B.: Estimating articulated human motion with covariance scaled sampling. IJRR 22, 371–392 (2003)
23. Sudderth, E.B., Mandel, M.I., Freeman, W.T., Willsky, A.S.: Visual hand tracking using non-parametric belief propagation. In: IEEE Workshop on Generative Model Based Vision, p. 189 (2004)