Body-part templates for recovery of 2D human poses under occlusion

Ronald Poppe and Mannes Poel *

Human Media Interaction Group, Dept. of Computer Science, University of Twente
P.O. Box 217, 7500 AE, Enschede, The Netherlands
{poppe,mpoel}@ewi.utwente.nl

Abstract. Detection of humans and estimation of their 2D poses from a single image are challenging tasks. This is especially true when part of the observation is occluded. However, given a limited class of movements, poses can be recovered given the visible body-parts. To this end, we propose a novel template representation where the body is divided into five body-parts. Given a match, we not only estimate the joints in the body-part, but all joints in the body. Quantitative evaluation on a HumanEva walking sequence shows mean 2D errors of approximately 27.5 pixels. For simulated occlusion of the head and arms, similar results are obtained while occlusion of the legs increases this error by 6 pixels.

1 Introduction

Detection and analysis of humans in images and video has received much research attention. Much of this work has focussed on improving pose estimation accuracy, while partly ignoring the difficult localization task. Despite increased awareness, the two processes are still researched in relative isolation, inhibiting use in realistic scenarios. Another issue with the current state of the art is the sensitivity to cluttered environments and, in particular, partial occlusions.

In this paper, we aim at simultaneous human detection and 2D pose recovery from monocular images in the presence of occlusions. We do not model the background, thus allowing our algorithm to work in cluttered and dynamical environments. Moreover, we do not rely on motion, which makes this work suitable for estimation from a single image. The output of our approach can be used as input for a more accurate pose estimation algorithm.

Our contribution is a novel template representation that is a compromise between half-limb locators and full-body templates. We observe that, for a limited class of movements, there is a strong dependency of the location of body-parts. For example, given a walking motion, we can accurately predict the location of the left foot while observing only the right leg. To this end, we divide the human body into five body-parts (arms, legs and torso), each of which has associated

* This work was supported by the European IST Programme Project FP6-033812 (publication AMIDA-105), and is part of the ICIS program. ICIS is sponsored by the Dutch government under contract BSIK03024.

edge and appearance templates. Given a match of body-part template and image, we not only vote for the locations of joints within the body-part but for all joint locations. This approach allows us to recover the location of joints that are occluded (see Figure 1). We first apply the templates over different scales and translations, which results in a number of estimations for each 2D joint location. In a second step, we approximate the final joint locations from these estimations. In this paper, we focus on the matching, and keep the estimation part trivial.

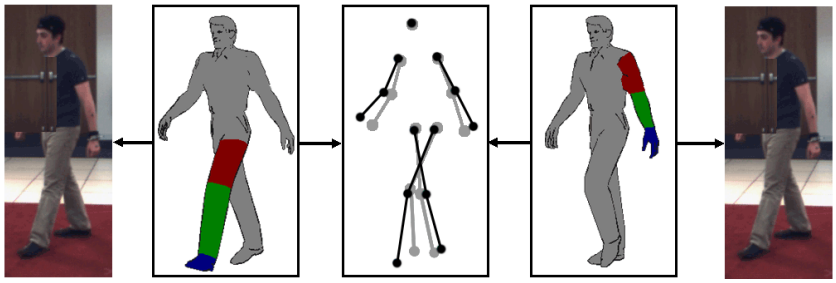

Fig. 1. Conceptual overview of our method. Templates from different exemplars and body-parts match with part of the image. Joint estimates are combined into a pose estimate. Anchor points are omitted for clarity. Occlusion of the right arm is simulated.

We first discuss related work on human pose recovery. The two steps of our approach, template matching and pose estimation, are discussed in Section 3 and 4, respectively. We present quantitative results on the HumanEva data set in Section 5, both on original image sequences and with simulated occlusion.

2 Related work on monocular pose recovery

Human motion analysis has received much attention [1]. Here, we focus on monocular approaches that can deal with cluttered, dynamic environments and occlusion. In general, we can distinguish two main classes of approach.

Discriminative approaches learn a mapping from image to human pose, where the image's region of interest is conveniently encoded in an image descriptor. Such approaches focus on poses that are probable, which is a subset of all physically feasible ones. Shakhnarovich et al. use histograms of directed edges and an efficient form of hashing to find similar upper-body examples from a database [2]. Agarwal and Triggs learn regression functions from extracted silhouettes to the pose space [3]. These approaches are efficient, but require accurate localization of the human from the image. Also, they are sensitive to noise in the region of interest, due to incorrect localization or segmentation, and occlusions. Some of these drawbacks have been partially overcome. Agarwal and Triggs suppress background edges by learning human-like edges [4], thus alleviating the need for

good segmentation. Howe uses boundary fragment matching to match partial shapes [5]. His approach requires that background and foreground are labelled, which limits its applicability to domains where such a segmentation is available.

The second class is that of *generative* approaches. These use a human body model that describes both the visual and kinematic properties of the human body. Pose estimation essentially becomes the process of finding the parameters that minimize the matching error of the visual model with the image observation. The direction of estimation is either top-down or bottom-up.

In top-down estimation, a projection of the human body is matched with the image observation, and usually improved iteratively. The process is hindered when occlusion occurs, since no image observation is present for the occluded part of the body. This can lead to unrealistic poses. A practical problem is the high dimensionality of the parameter space, which makes initialization difficult. A recent trend to overcome this problem is to use dimensionality reduction in the kinematic space, which can be regarded as a strong prior on the poses that can be observed. This reduction is motivated by the observation that there is a strong correlation in the movement of different body-parts, especially within a single movement class such as walking.

In bottom-up estimation, individual body-parts are found first and then assembled into a human body. In general, weak half-limb detectors are used, which results in many false positives. Many of the bottom-up works resemble the pictorial structures idea, which was applied to human pose recovery by Felzenszwalb and Huttenlocher [6]. The key idea is to model the appearance of each body-part individually, and represent the deformable assembly of parts by spring-like connections between pairs of parts. Most of this work relies on inference in a tree-like structure [6–8]. Again, there are two major drawbacks with the bottom-up approach. First, the templates are usually at the level of half-limbs (e.g. upper leg) which results in many false positives. Second, the 2D location of a template does not give any information about the rotation in 3D. This makes it difficult to enforce 3D constraints, such as joint limits, on the relative position between two adjacent parts. Such constraints are needed to be able to recover realistic poses when part of the body in the image is occluded.

In this paper, we propose an approach that combines several of the ideas above, while it aims at circumventing the major drawbacks. First, we use body-part templates that encode exactly one body-part (arm, leg or torso). Such a representation is more meaningful than that of half-limbs, and reduces false positives since the templates implicitly encode the view. Second, by voting over all joints, we can cope with occlusions and recover the pose even when only part of the human body is visible. See also Figure 1. Our work resembles that of Demirdjian and Urtasun [9], who vote over the pose space using patches that are similar to those in the image. Patches and their joint location densities are learned from a large annotated database. Our work differs since our few templates can be generated automatically using 3D modelling software. We focus on walking movements only. This effectively puts a strong prior on the poses that we can recover, which is a limitation of our work.

3 Template matching

We introduce full-body templates that consist of five, possibly overlapping, body-part templates, see Figure 2. We will call a full-body template *exemplar*. For the estimation of articulated poses, we use a collection E of n exemplars $\mathbf{E}_i \in E$ $(1 \leq i \leq n)$. Each exemplar consists of a tuple that describes the 2D pose $\mathbf{u}$ and the body parts $\mathbf{p}$, $\mathbf{E}_i = (\mathbf{u}, \mathbf{p})$. Note that we do not include information about the 3D orientation and scaling since this is implicitly encoded in the templates and 2D pose. For clarification of notation, subscripts are omitted where possible.

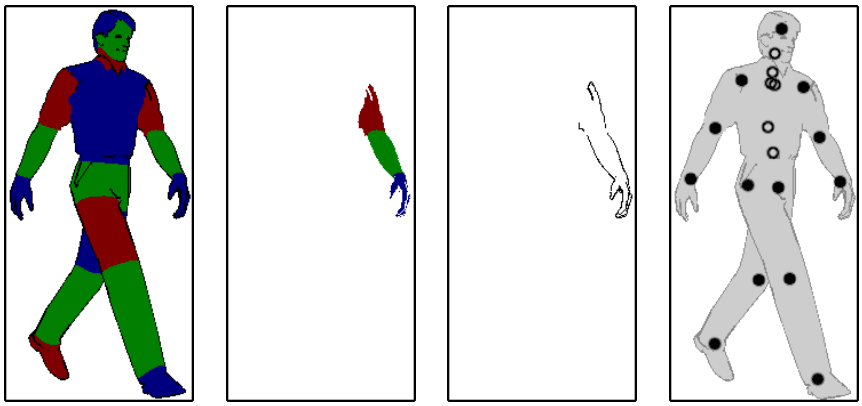

Fig. 2. (a) Exemplar with all color regions, (b) color regions and (c) edge template of left arm, (d) 2D joint locations (body shown for reference). Closed dots are used in the evaluation, see Section 5. The anchor point is the left-upper corner of the box.

Each element in $\mathbf{u}$ is a 2D joint location written as a tuple $\mathbf{u}_i = (x_i, y_i) \in \mathbf{u}$ $(1 \leq i \leq m)$. These locations are relative to the anchor point, which is by default the left upper corner of the minimum enclosing box of all templates. In principle, we use $m = 20$ joints, as shown in Figure 2(d).

Each $\mathbf{p}_i \in \mathbf{p}$ $(1 \leq i \leq 5)$ represents exactly one body-part (leg, arm or torso). We can write it as a tuple $\mathbf{p}_i = (\mathbf{t}, \mathbf{r})$. Here, $\mathbf{t}$ is an edge template, where each element $\mathbf{t}_i = (x_i, y_i) \in \mathbf{t}$ $(1 \leq i \leq |\mathbf{t}|)$ represents an edge pixel at a given location, relative to the anchor point. $|\mathbf{t}|$ is the number of elements in $\mathbf{t}$. Each body-part has several color regions, each of which is assumed to have a constant color. The number of regions per body-part, $|\mathbf{r}|$, is three for each leg and arm, and five for the torso (see also Figure 2(a-b)). Similar to our edge representation, each region $\mathbf{r}_i$ $(1 \leq i \leq |\mathbf{r}|)$ consists of a number of relative pixel locations, which can be considered the foreground mask. The total number of foreground pixels in a region is denoted with $|\mathbf{r}_i|$ for the i^{th} region.

In summary, each exemplar consists of five body-parts, each of which represents a limb or the torso. Each body-part has an associated edge template, and a number of color regions. Templates and 2D joint locations are positioned relative to the anchor point.

3.1 Template distance

For the matching, the notion of an exemplar is not needed, rather that of the body-parts individually. The match of a body-part and an image region is determined by calculating the distance of the edge and color templates individually.

For the edge matching, distance transforms such as the Chamfer distance are common. For such a transform, both image and template need to be converted to a binary edge map. The distance transform gives the (approximate) distance to the nearest edge pixel. The matching score is calculated by summing all distance transform values in the image "under" the edge pixels. One problem that we found while using this technique is that is favors areas that are densely covered with edges. Also, the performance proved to be very sensitive to the value of the edge magnitude threshold. Therefore, we use the edge magnitudes of the image directly, by calculating the derivative I_{edge} of the image in gray scale. Distance score Δ_{edge} for template $\mathbf{t}_j$ with the anchor at location (sx, sy) is given by:

$$\Delta_{edge}(I_{edge}, \mathbf{t}_j, sx, sy) = \frac{\sum_{(x,y)\in\mathbf{t}_j} I_{edge}(sx+x, sy+y)}{|\mathbf{t}_j|} \tag{1}$$

To evaluate the distance of the color template, we determine the color deviation score Δ_{color} for each region $\mathbf{r}_k$ at anchor point (sx, sy):

$$\Delta_{color}(I_{color}, \mathbf{r}_k, \mathbf{c}, sx, sy) = \frac{\sum_{c_i\in\mathbf{c}} \sum_{(x,y)\in\mathbf{r}_k} |I_{color}(sx+x, sy+y, c_i) - \mu|}{|\mathbf{c}||\mathbf{r}_k|} \tag{2}$$

Here, I_{color} is the color image, $\mathbf{c}$ is a vector with $|\mathbf{c}|$ color channels. We make no assumptions about the color space of the image. μ is the shorthand notation for $\mu(I_{color}, \mathbf{r_k}, c_i, sx, sy)$, the mean value of all "region" pixels in the color channel when no appearance assumptions are made. Alternatively, if the region colors are set beforehand, μ corresponds to $a(j, k, i)$ the specified value for body-part j, region k and channel i. Alternatively, we could have used color histograms, as in [8] but it seems unrealistic that these can be determined beforehand.

The distance for all color regions together is the sum of the means of each region, weighted on the size of the regions $|\mathbf{r}_k|$. We have distances rather than probabilities, so we need to determine when a match occurs. Therefore, we introduce thresholds η and θ for the minimum edge distance and maximum color deviation distance, respectively. We determine the values of these thresholds empirically.

4 Pose estimation

To estimate the 2D joint locations of a person in an image, we evaluate the distances of the body-parts of all exemplars in collection E. Each template is matched over multiple scales with the anchor point at different locations. For each match (with the distance scores satisfying the thresholds), a vote is made for the location of all joints in $\mathbf{u}$. The body-parts are unrelated to the exemplar, except for the common joint locations.

After processing the image, we have a number of estimates for each joint location. This "density" usually has multiple modes, depending on the number of persons, and the modes of uncertainty. For simplicity, we assume only a single person in the image. Therefore, we simply take the average location of all estimates for a given joint. This presents the risk of averaging over multiple modes. To be able to handle multiple persons, a more advanced density estimation scheme could be used such as the one described in [9].

5 Experimental results and discussion

To evaluate the performance of our technique, we evaluate the algorithm on the publicly available HumanEva benchmark set [10]. To our best knowledge, there is no data set that contains partially occluded human figures and 2D annotated joint positions. Therefore, we simulate occlusion on the HumanEva set.

5.1 Training set

One of the advantages of our approach is that templates can be generated using 3D modelling software. This makes it possible to generate templates that do not require manual labelling of joint positions, edge locations and color regions.

Curious Labs' Poser 5 was used, with the "Poser 2 default guy" as human model. We selected the "P4 Walk" as motion, and sampled 6 key frames within the cycle. The camera was placed at eye height, and was pointed slightly downwards. Each key frame was further viewed from 8 different angles at every 45° around the vertical axis. This yields 48 exemplars. For each exemplar, we used the "Cartoon with line" renderer to generate edge templates and color regions. See Figure 2(b-c) for examples of templates. In a post-processing step, the joint locations and templates are normalized with respect to the left-upper corner of the minimum enclosing bounding box of all templates.

5.2 Test set

We evaluated our approach on the HumanEva data set [10]. This set contains sequences with synchronized video and pose data. For the test sequences, ground truth is held back and validation is performed online. We present results for Walking and Jog sequence 2, performed by subject 2 and viewed from color camera 1. This sequence shows a man walking or jogging in circles.

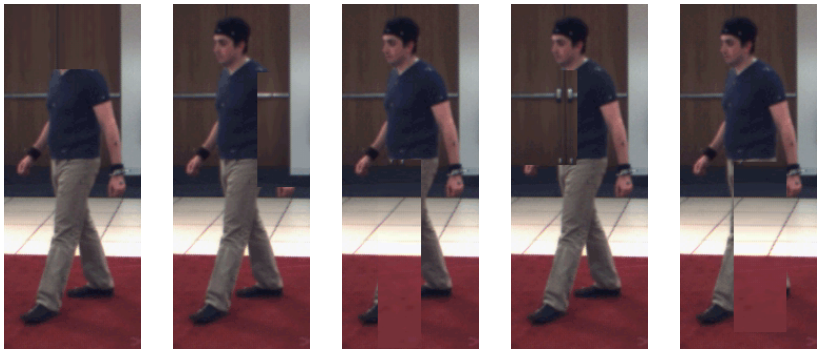

Fig. 3. Simulated occlusion. Left to right: head, left arm, left leg, right arm, right leg.

In order to test the accuracy of our approach against occlusions, we simulate occlusions for different body-parts. Instead of placing black boxes, as in [9], we remove the body-parts by replacing the foreground with background patches. Note that we do not model the background, so we have no knowledge where occlusion occurs. The location and size of the patch is determined by the 2D location of the shoulder and wrist, hip and ankle, and head for an arm, leg and head respectively. These locations were annotated manually. Figure 3 shows example frames with occlusion for different body-parts. The patches often occlude other body-parts, which is especially the case for the legs. Also, due to the location of the selected joints, there are still some parts visible, notably hands and feet. However, it is unlikely that this aids in the matching phase.

5.3 Results

We defined values for the color of the skin, shirt, pants, shoes and hair in HSV color space. Hue values were transformed to be the distance to the center value (180°), to avoid wrap-around errors. To reduce computation time, we assume that the person is entirely within the image. We move the anchor point through the image with steps of 10 pixels. This causes the edge term to function as a rejector of false positives. The human model in our exemplars is approximately 350 pixels high. We evaluate the templates at three scales: 90%, 100% and 110%. The human figure in our test sequence is between 275 and 410 pixels high, so in the range 79-117% of our exemplars. We further reduce computation time by ignoring body-parts with an area smaller than 1500 pixels. This excludes occluded or almost occluded limbs from being evaluated. These templates have a high probability of matching, while providing little information about the pose.

We used Walking sequence 1 of subject 2 to determine the thresholds for the templates. The exemplars and the HumanEva set use different joint sets. We selected the joints corresponding to the wrist, elbow, shoulder, ankle, knee, hip and head. In Figure 2(d), these are the closed dots. The mean 2D error over the

Walking test sequence is 27.48 pixels, with a SD of 3.26. This corresponds to an error of approximately 14 cm, if we average over the scales. We evaluated the Jog sequence, frames 100339, which corresponds to a full cycle. The average error is 30.35 pixels, with a SD of 3.90. For the evaluation of the occluded images, we used frames 1400 (one walking cycle) with an increment of 5 frames. We used the same settings as for the unmodified sequence. Specifically, we did not alter the thresholds. Results are presented in Table 1. In addition, we applied the method to some frames of the movie *Lola Rennt* (sample frames in Figure 4(a-b)).

Head	Left arm	Left leg	Right arm	Right leg
27.32 (3.64)	27.31 (3.49)	32.77 (8.57)	27.65 (3.40)	32.95 (7.52)

Table 1. Mean 2D error (and SD) in pixels on HumanEva Walking 2 sequence, subject 2, viewed with color camera 1. Results are given for different occlusion conditions.

5.4 Discussion and comparison with related work

Other works have reported 2D errors on the HumanEva data set [8, 11]. While these works are substantially different than ours, comparison may reveal the strong and weak aspects of the respective approaches. Siddiqui and Medioni present results on the Gesture sequence in the training set of subject 2. They report mean errors of approximately 13 pixels, for the upper-body. Templates are used, but for half-limbs, and colors specified as a histogram. In addition, motion information obtained from frame differencing is used. The background is modelled and temporal consistence is enforced through tracking. Their method can deal with a broader range of poses and is considerably faster.

Howe [11] uses a discriminative approach with a database of reference poses with corresponding image representation to retrieve similar observations. Temporal consistency is enforced using Markov chaining. On the Walking sequence in the training set of subject 1, mean errors of approximately 15 pixels are obtained. Howe's approach works in real-time, but requires good foreground segmentation. Also, it remains an open issue whether similar results can be obtained for subjects that are not in the training set.

Unlike both approaches above, our method is able to deal with occlusions, at a cost of higher computational cost and lower flexibility with respect to the range of poses that can be detected. Our larger errors are partly due to our evaluation method, and are partly inherent to our approach. There is a discrepancy between the joint locations in our exemplars and those defined for HumanEva. We selected those joints that have similar locations but differences are still present. The hips are placed more outwards in our exemplars, and the elbow and knee locations are more at the physical location of the joint. Also, the human model used in our exemplars differs in body dimensions, compared to subject in our test data (see Figure 1).

To reduce the computational cost, we used only 6 different key poses, viewed at 45° intervals. Also, the walking style differs substantially from the one observed in the test sequence. A similar observation can be made for the Jog sequence. Our results could be improved by adding exemplars and viewpoints, at the cost of increased computational complexity. By moving the anchor point with steps of 10 pixels, our edge template functions as a false positive rejector. Changing the matching function could improve results.

Closer analysis of our results shows that part of the error is caused by matches that are the 180° rotation of the real match. This happens especially when the person is facing the camera, or facing 180° away from it. This happens in frames 50, 250, 450, etc., see Figure 4(c). Consequently, we see lower error values around frames 150, 350, 550, etc. Here, the subject is either walking to the right or walking to the left. The number of left-right ambiguities are lower, resulting in a lower error. The relatively higher errors around frames 350 and 750 are caused by the subject being close to the camera. Here, matches with a larger scale are selected, which causes higher errors for ambiguities. Joints closer to the symmetry axis are much less affected by these errors, but these are not used in the evaluation. Overall, the estimated joint locations are closer to the symmetry axis than the real locations. A final observation can be made that the majority of the matches is from the leg templates, which explains the higher errors when one of the legs is occluded.

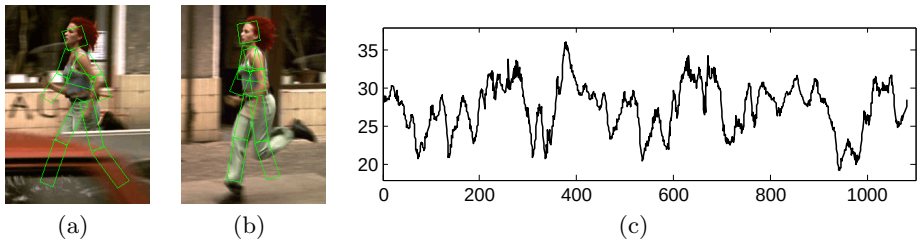

Fig. 4. Sample frames from *Lola Rennt*. (a) Legs occluded, (b) "out-of-vocabulary" pose. (c) Errors in pixels for Walking sequence 2, subject 1. See discussion in Section 5.4.

6 Conclusion and future work

We presented a novel template representation, where the body is divided into five body-parts. Each body-part implicitly encodes the viewpoint and is used to predict the location of other joints in the human body. By matching body-part templates individually, our approach is able to detect persons, and estimate their 2D poses under occlusions. We match edge and color templates associated with a body-part at different locations and scales. For a good match, an estimate for all joints is made. Subsequently, we determine the final pose estimate.

The HumanEva data set was used for evaluation. We simulated occlusion by replacing limbs with background patches. For the original walking sequence, and for occlusion of head and arms, we obtained mean 2D errors of approximately 27.5 pixels. Occlusion of the legs resulted in a 6 pixel increase.

These results can be improved by adding exemplars and viewpoints. Also, the edge matching could be improved to better fit the observations. A better pose estimation process would allow for multiple persons, and could favor matches close to the actual joint location. To reduce computational cost, we propose a coarse-to-fine matching approach. Other future work is aimed at combining our work with a discriminative or generative approach.

References

1. Poppe, R.: Vision-based human motion analysis: An overview. Computer Vision and Image Understanding (CVIU) **108**(1-2) (October 2007) 4–18
2. Shakhnarovich, G., Viola, P.A., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: Proceedings of the International Conference on Computer Vision (ICCV'03) - volume 2, Nice, France (October 2003) 750–759
3. Agarwal, A., Triggs, B.: Recovering 3D human pose from monocular images. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **28**(1) (January 2006) 44–58
4. Agarwal, A., Triggs, B.: A local basis representation for estimating human pose from cluttered images. In: Proceedings of the Asian Conference on Computer Vision (ACCV'06) - part 1. Number 3851 in Lecture Notes in Computer Science, Hyderabad, India (January 2006) 50–59
5. Howe, N.R.: Boundary fragment matching and articulated pose under occlusion. In: Proceedings of the International Conference on Articulated Motion and Deformable Objects (AMDO'06). Number 4069 in Lecture Notes in Computer Science, Port d'Andratx, Spain (July 2006) 271–280
6. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision **61**(1) (January 2005) 55–79
7. Ramanan, D., Forsyth, D.A., Zisserman, A.: Tracking people by learning their appearance. IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) **29**(1) (January 2007) 65–81
8. Siddiqui, M., Medioni, G.: Efficient upper body pose estimation from a single image or a sequence. In: Human Motion: Understanding, Modeling, Capture and Animation. Number 4814 in Lecture Notes in Computer Science, Rio de Janeiro, Brazil (October 2007) 74–87
9. Demirdjian, D., Urtasun, R.: Patch-based pose inference with a mixture of density estimators. In: Proceedings of the International Workshop on Analysis and Modeling of Faces and Gestures (AMFG'07). Number 4778 in Lecture Notes in Computer Science, Rio de Janeiro, Brazil (October 2007) 96–108
10. Sigal, L., Black, M.J.: HumanEva: Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report CS-06-08, Brown University, Department of Computer Science, Providence, RI (September 2006)
11. Howe, N.: Recognition-based motion capture and the HumanEva II test data. In: Proceedings of the Workshop on Evaluation of Articulated Human Motion and Pose Estimation (EHuM) at the Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN (June 2007)