Multi-view Body Part Recognition with Random Forests

Vahid Kazemi
vahidk@csc.kth.se
Magnus Burenius
burenius@csc.kth.se
Hossein Azizpour
azizpour@csc.kth.se
Josephine Sullivan
sullivan@csc.kth.se

CVAP / KTH
The Royal Institute of Technology
Stockholm, Sweden

Abstract

This paper addresses the problem of human pose estimation, given images taken from multiple dynamic but calibrated cameras. We consider solving this task using a part-based model and focus on the part appearance component of such a model. We use a random forest classifier to capture the variation in appearance of body parts in 2D images. The result of these 2D part detectors are then aggregated across views to produce consistent 3D hypotheses for parts. We solve correspondences across views for mirror symmetric parts by introducing a latent variable. We evaluate our part detectors qualitatively and quantitatively on a dataset gathered from a professional football game.

1 Introduction

In this paper we address the problem of automatically estimating the 3D pose of a person seen from multiple calibrated cameras outside a studio environment [14]. Our particular focus is the estimation of the 3D pose of football players during a professional game. Football footage have several key characteristics some of which are shared between different sports. Most notably the images are commonly disturbed by motion blur because of the fast moving players and cameras. There is also a large variation in the players' 3D pose. On the other hand the variation in the players' clothing is limited and background clutter is not as severe as in less structured environments. Yet, low quality images and fast motion make it hard to perform background subtraction reliably.

Currently, the most successful solutions to 2D pose estimation are discriminatively trained part-based models [1, 8, 12, 16, 21]. This class of methods are attractive as they enable efficient inference by reducing the conditional dependencies between parts, and demand less labeled training data as they can generate new poses at test time. Part-based models have also been used for 3D pose estimation [2, 4, 13, 17, 19], but to our knowledge good performance has only been reported in studio environments. In this paper we focus on computing efficient and accurate 3D part appearance likelihoods that can be plugged into any 3D part-based

model. We show that these 3D part appearance likelihoods allow for 3D pose estimation outside of the studio without imposing a strong pose prior.

To discriminatively learn the 3D part likelihoods directly for the individual parts would require labeled 3D data and the associated calibrated views. We want to avoid this potentially expensive and non-trivial labeling task. Therefore, in this paper we discriminatively learn 2D part likelihoods for each part and aggregate the likelihoods from the different views to obtain the 3D part likelihoods. This means that we only need labelled 2D images from uncalibrated cameras. However, to get good performance this requires solving a part correspondence problem across views during the aggregation phase. We return to this issue later in the introduction, but now we turn to the issue of how to learn and compute the 2D part likelihoods.

State of the art 2D part based models for human pose estimation rely on SVM classifiers applied to a HOG descriptor of an image patch [21]. However we opt to use a more efficient random forest approach for estimating the part likelihoods. We take our inspiration from the recent success of the *Kinect* system. Shotton *et al*. [18] use a random forest to estimate a person's 3D pose from a depth image. They divide the human body into a set of parts and a random forest is used to estimate the probability of each pixel belonging to each part. From these probabilities the 3D location of the skeletal joints are then independently estimated. Their work clearly demonstrates that given sufficiently diverse training data, one can learn a compact random forest classifier which at test time efficiently recognizes parts across a very varied set of 3D poses. In this paper we consider ordinary visual images, as opposed to depth images, but similarly use a random forest to assign to every pixel a probability of being a particular part or background. These probabilities form the basis for our part likelihood scores in 2D and 3D.

We create 3D part appearance likelihoods by aggregating the 2D likelihoods across all camera views. Care must then be taken regarding the correspondence of joints across the views. Because of the similar appearance of mirror symmetric parts, such as left and right arms and legs, and also the local nature of our part detectors, we can not directly distinguish the correct correspondences for each part. In this paper this issue is handled by introducing a latent variable into our model which represents the correspondence. At inference time we optimize for both the best pose and the best values of our latent variable. We show that this approach is both feasible and effective (fig. 2).

We now summarize the contributions of this paper.

1. We introduce a new dataset, KTH Multiview Football Dataset, of annotated football images consisting of 5900 images with 2D annotations and 1167 with 3D annotations and calibrated multi-view camera parameters.

2. We benchmark the performance of a 2D part-based model, which uses our random forest based 2D part appearance likelihoods. We show that given sufficient labelled training data our method outperforms the state-of-the-art methods for 2D pose estimation on football footage.

3. We show how multi-view 3D appearance likelihoods can be computed from 2D likelihoods. We solve correspondences across different views for mirror symmetric parts by introducing a latent variable. We demonstrate how our 3D likelihoods can be plugged into a 3D part-based model and used to estimate 3D poses outside a studio environment without imposing a strong pose prior.

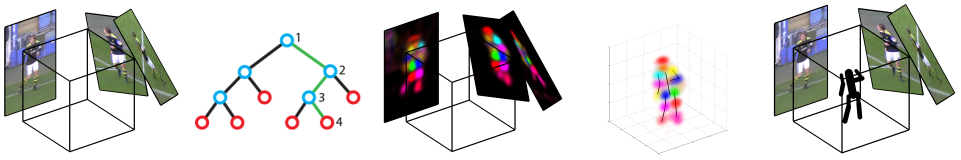


Figure 1: A general overview of our multi-view pose estimation framework. A random forest is first used to classify each pixel in each image as belonging to a part or the background. The results are then back-projected to a 3D volume. We find corresponding mirror symmetric parts across views by introducing a latent variable. Finally, a part-based model is used to estimate the 3D pose.

2 Method

Given a set of calibrated cameras viewing a person, our goal is to estimate the location of body joints in 3D. Figure 1 shows a general overview of our framework. First a random forest is used to classify each pixel in each image as a part or the background, as described in section 2.1. We then discuss how the resulting 2D part appearance likelihoods can be used for 2D pose estimation in section 2.2. This process is performed so that we can compare 2D part detectors to previous work for 2D pose estimation. The results from section 2.2 are not used for performing 3D inference. For 3D part appearance likelihoods we back-project the result of the random forest pixel classification to a 3D volume, as described in section 2.3. We then discuss how our 3D part appearance likelihoods can be plugged into any multi-view part-based model in section 2.4. The problem of mirror ambiguity for symmetric parts is addressed in section 2.5.

2.1 Appearance likelihoods in 2D using random forests

We use a random forest of classification trees to estimate the probability that a pixel v belongs to a skeletal joint or the background class. The split decisions made in each tree are based on thresholding a dimension of the UoC-TTI HOG descriptor [9] of the image window. This dimension is defined by three numbers as follows. First there is a 2D offset vector u. It is computed within which cell of the HOG descriptor the point $u+v$ falls. The final number defines the dimension of $(u+v)$'s cell descriptor to be accessed. It is this entry which is thresholded in the split decision. The offsets considered are constrained to be within a certain distance of v.

We have training images that have the position of the 2D skeleton joints labelled. From these labelled images we generate a new labelled dataset $\{(h^k, y^k)\}_{k=1}^K$ where h^k is the HOG descriptor of an image centered at a pixel having a class label $y^k \in \{0,1,\ldots,N\}$. The label 0 corresponds to the background class and the other numbers to the skeletal joints. This is the labelled data we use to train the random forest. We use the standard procedure for training random forests similar to [5, 13].

When we apply a learnt decision tree to a test image i and a pixel location v in an image with HOG descriptor h we will reach a leaf node m. The posterior probability of pixel v having label y is equal to the proportion of the training samples that reach node m and have label y. The output of our random forest is the average of the probabilities returned by the trees in the forest. After the random forest is run on all pixels in the image we separately smooth the posterior probability maps obtained for each part. The final response image for each part n is denoted by $f_n(i,v)$.

2.2 Inferring the 2D pose

We first formulate the pose estimation problem in 2D. This is done so we can introduce our notation for part-based models and can compare the random forest results to previous work for 2D pose estimation. However, the results from this sub-section are not used when performing the multi-view 3D inference.

Let V_n be a random variable representing the 2D position of part n. The 2D pose of the person is then $V = (V_1, \ldots, V_N)$. Let I be a random variable representing the image evidence. We consider part-based models that assume there is some image evidence for each part I_n and that these are conditionally independent given the position of the parts

$$P_{I|V}(i \mid v) = \prod_n P_{I_n|V}(i_n \mid v) \tag{1}$$

where lower cases are used for outcomes of the random variables. We use the response from the random forest as the 2D part appearance likelihoods

$$P_{I_n|V_n}(i_n \mid v_n) \propto f_n(i, v_n) \tag{2}$$

We infer the pose by finding the most probable state of V given the measurement data

$$\max_v P_{V|I}(v \mid i) = \max_v \left[\ln P_V(v) + \sum_n \ln P_{I_n|V_n}(i_n \mid v_n) \right] \tag{3}$$

where $P_V(v)$ describes an arbitrary 2D pose prior. This optimization can be solved in different ways, depending on the form of the 2D pose prior $P_V(v)$. In our implementation we first find the modes of the part appearance likelihoods $P_{I_n|V_n}(i_n \mid v_n)$. To make the process more efficient we first sample pixels with high probabilities to find a small set of modes. In practice we use the meanshift algorithm for this. In many cases, taking the mode with the highest probability for each joint independently leads to a valid configuration. (This corresponds to the pose prior $P_V(v) = \prod_n P_{V_n}(v_n)$ where each $P_{V_n}(v_n)$ is uniform and the same for all n.) This is because the random forest is able to aggregate information from a relatively large neighbourhood around each joint and produce confident joint hypotheses. There are, however, some cases where this approach fails. To find the estimated joints which have both a spatial configuration consistent with a valid 2D pose and high appearance scores, we search for the optimal combination of body joints from a small set of highly probable modes. This is done efficiently by using dynamic programming to minimize a cost function that incorporates a simplified shape prior. Specifically, we assume that $P_V(v)$ is factorized over a tree graph and use a mixture of Gaussians prior for the relative location of joints with respect to their parents [21]. The parameters of this prior are calculated separately based on the statistics of the training data annotations.

2.3 Appearance likelihoods in 3D

Let the 3D position of joint n be the random variable X_n and the 3D pose $X = (X_1, \ldots, X_N)$. The image evidence from view c for joint n is represented by the random variable $I_{c,n}$ and the evidence of all joints for a single view is $I_c = (I_{c,1}, \ldots, I_{c,N})$. Let $V_{c,n}$ be the 2D position of joint n in view c. Let T_c be the projective transformation of camera c. We assume the 2D position $v_{c,n}$ of joint n in view c is deterministically determined as $v_{c,n} = T_c(x_n)$. The part appearance likelihood for view c is computed by projecting X_n to that view

$$P_{I_{c,n}|X_n}(i_{c,n} \mid x_n) = P_{I_{c,n}|V_n}(i_{c,n} \mid T_c(x_n)) \propto f_n(i_c, T_c(x_n)) \tag{4}$$

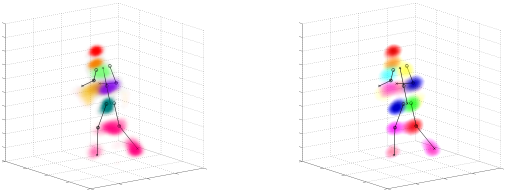


Figure 2: Overcoming ambiguities introduced by symmetric appearances. The left image shows the 3D appearance likelihoods computed from part detectors that ignore the left and right label of the parts. The right image shows the result of finding corresponding parts across views by maximizing a latent variable. The ground truth pose is shown in black.

We assume the image evidence across views is conditionally independent given x_n and thus compute the multi-view 3D appearance likelihood (see figures 1 and 2) as

$$P_{I_{1,n},\ldots,I_{C,n}|X_n}(i_{1,n},\ldots,i_{C,n} \mid x_n) = \prod_c P_{I_{c,n}|X_n}(i_{c,n} \mid x_n) \tag{5}$$

2.4 Inferring the 3D pose

Similar to 2D we estimate the pose by computing the most probable state of X given the measurement data. This equates to finding the maximum of the posterior distribution

$$\max_x P_{X|I_1,\ldots,I_C}(x \mid i_1,\ldots,i_C) = \max_x \left[\ln P_X(x) + \sum_n \sum_c \ln P_{I_{c,n}|V_{c,n}}(i_{c,n} \mid T_c(x_n))\right] \tag{6}$$

where $P_X(x)$ describes an arbitrary 3D pose prior. This optimization can be solved in different ways, depending on the choice of the state space for X and the form of the 3D pose prior $P_X(x)$. Depending on whether we have a continuous or discrete state space a solution can be found using either stochastic optimization [19] or dynamic programming [2, 4, 17].

Our 3D appearance likelihoods can be used by any multi-view part-based model. To demonstrate the performance of a full system we follow the approach of [4] and discretize the state space. We refer to [4] for an analysis of the tractability of this approach. We assume the person is within a bounding cube and create a uniform grid covering this cube. The appearance likelihoods are then evaluated for all grid points. We consider two different pose priors $P_X(x)$. The first is $P_X(x) = \prod_n P_{X_n}(x_n)$ with $P_{X_n}(x_n)$ uniform over its state space. Then the global optimum can be found by optimizing equation (5) for each joint independently. The second pose prior imposes limb length and intersection constraints as in [4].

2.5 Overcoming ambiguities introduced by symmetric appearances

In equation (4) we have assumed that the mapping between the labels for the 2D joints and the 3D joint labels is consistent across views and that it is one-to-one. However, this is not necessarily the case especially for the mirror symmetric joints, i.e. joints associated with the right and left legs (arms). For such joints, the classifier can either be trained to

- just detect the joints and ignore their label as left or right or
- recognize the left and right label *of the image*

In the latter scenario we do not know if the joints labelled as left in two views correspond to the same physical 3D joints. Therefore to match the left and right legs *of an image* with the left and right *of the person* we have two choices. If we also try to match the arms we have a total of $2^2 = 4$ choices per image. Considering all C views gives a total of 2^{2C} choices.

To handle this mirror ambiguity we introduce a discrete latent random variable $M_c = (M_{c,1}, \ldots, M_{c,N})$ which represents the mapping of the labels from the 3D joint labels to the 2D joint labels in view c. We assume M_c is uniformly distributed over its 4 states. For non-limb joints the mapping is considered unambiguous. Instead of using (4) we thus let the image evidence of each joint depend on $M_{c,n}$ as follows

$$P_{I_{c,n}|M_{c,n},X_n}(i_{c,n} \mid m_{c,n}, x_n) = P_{I_{c,n}|M_{c,n},V_n}(i_{c,n} \mid m_{c,n}, T_c(x_n)) \propto f_{m_{c,n}}(i_c, T_c(x_n)) \tag{7}$$

Then the optimum of the full posterior distribution for X and $M = (M_1, \ldots, M_C)$ assuming a uniform prior over M is given by

$$\max_{x,m} P_{X,M|I_1,\ldots I_C}(x, m \mid i_1, \ldots, i_C) = \max_m \max_x \left[\ln P_X(x) + \sum_n \sum_c \ln P_{I_{c,n}|M_{c,n},V_{c,n}}(i_{c,n} \mid m_{c,n}, T_c(x_n)) \right] \tag{8}$$

and this becomes the optimization problem we solve at inference time as opposed to (6). See figure 2. We perform the outer optimization over m by exhaustive search, independently of the method used for the inner optimization over x. This approach can therefore be applied to any part-based model. When we solve this optimization problem the joints across the views will be in correspondence, but there may still be an unresolved front/back ambiguity in 3D.

3 Experiments

To benchmark the performance of our approach in a realistic outdoor scenario we have created the publicly available KTH Multiview Football Dataset from a professional football game. The dataset consists of about 7000 images of two different players of the same team. We first annotated the 2D pose of the players for 5907 images. We used 3900 of these to train the random forest and the rest for testing the 2D pose estimation performance.

We additionally annotated two sequences where the player was captured by three moving cameras. The first sequence consists of 214×3 images and the second sequence of 175×3 images, recorded at a frame-rate of 25Hz. We used the 2D annotation to synchronize and calibrate the cameras and the human pose is reconstructed in 3D using the affine factorization algorithm [9, 10, 15, 20]. We used the 3D reconstruction of the first sequence as the ground truth for testing the 3D pose estimation performance.

3.1 Scoring and inference in 2D

What follows contains an analysis of the effect of different parameters on the performance of the random forest, as well as a comparison with the state of the art *Flexible Mixture of Parts (FMP)* model [21] trained and tested on our football dataset.

Number of trees: It is well known that decision trees are prone to overfitting and combining multiple trees can significantly help in regularizing their outcome [11]. However, we observe that in our case the improvement with more than two trees is not drastic, see figures 3(a) and 4(a) In our experiments we fixed the number of trees to 5.

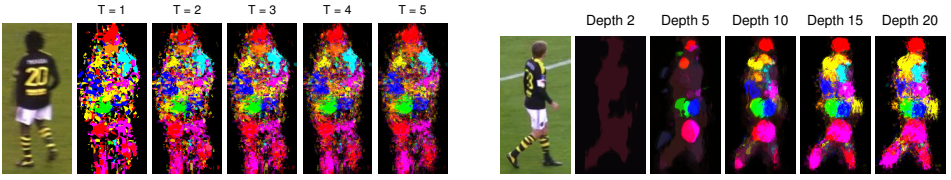
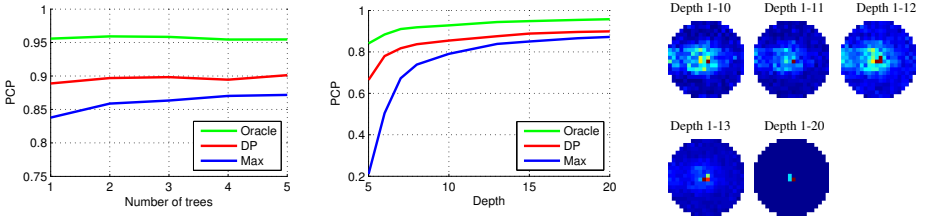


Figure 3: Left: The effect of the number of trees in the random forest on performance. The increase in performance is minimal with the addition of more trees to the forest after the first two. Right: How the depth of the tree affects the output of the random forest.



(a) Number of trees (b) Depth of tree (c) Distribution of offsets

Figure 4: (a) The effect of the number of trees on the PCP score with three different matching methods. These are taking the modes with maximum probability (blue curve), using dynamic programming with a simple shape prior (red curve), and an oracle matching method on the highest (5-10) probability modes (green curve). (b) The change in performance as a result of increasing tree depth. (c) The 2D histogram of the offsets selected at the decision tree splits at different depth levels. The initial splits use information from wider neighbourhoods.

Depth of trees: Figures 3(b) and 4(b) show how the depth of the trees affects the performance of random forest. It can be observed that with a random forest of depth 5, we can already correctly classify pixels belonging to easy to detect parts like head, hips, and knees. The depth of each tree was set to 20 in our experiments. It is worth mentioning that the resulting decision trees are not balanced. The decision trees trained on our dataset have around 10% of the nodes of a balanced tree with equal depth.

Feature pool: Decisions at each node are made by thresholding HOG [6, 9] dimensions in a neighbourhood of each pixel. To increase randomization, at each node a pool of features is created by selecting a random subset of all the available features. The optimal feature and threshold are then chosen from this pool. We set the feature pool size to 25000. Figure 4(c) shows the distribution of the offsets chosen at different depths levels of the random forest. The results show that at the earlier levels of the tree a wide exploration of the surrounding area is performed, but as we move down to the bottom of the tree most of the selected features are centered at the probe pixel. In our experiments we allow for offsets up to 50 pixels. The height of the person is about 180 pixels.

Comparison of our 2D pose estimation method to state-of-the-art: We compare our results to *Flexible Mixture of Parts(FMP) [21]* which achieves state of the art performance on general 2D human pose estimation tasks. We have trained and tested their method using the original code provided by authors on our football dataset.

Table 1 shows a summary of results on our football dataset. The results show that our method based on random forest (RF) outperforms FMP [21] on this dataset. It is also worth mentioning that our simple random forest is at least an order of magnitude faster than FMP,

Figure 5: A qualitative comparison of random forests with a state of the art pose estimation method on our dataset. The top row shows the modes of probabilities output from the random forest. A point's size indicates its certainty level. The second row is the result of inferring the configuration by imposing 2D pose priors. The last row is the result of FMP [21].

since we do not need to convolve the HOG-image with several filters for each part. Figure 5 shows some qualitative results comparing to the predictions of FMP model. The major problems seem to be caused by unseen poses, lack of strong features for parts such as lower arms, and the absence of contextual support, e.g. for outstretched limbs. The latter can be potentially solved by using higher offsets (as described in section 2.1).

We also tried our random forest on some standard datasets, which were smaller than our football dataset and had more background clutter. Under those conditions FMP still outperforms our random forest. We believe that the difficulty to deal with severe background clutter is a disadvantage of the current version of our part detectors. However, a recent work [7] shows state of the arts performance within a very similar random forest framework. Although, this approach still seems to require considerably more training data than FMP.

	Head	Torso	Upper Arms	Lower Arms	Upper Legs	Lower Legs	Average
Flexible Mixture of Parts	**.97**	**.99**	.92	.66	.94	.80	.86
RF	.94	.96	.90	.69	.94	.84	.87
RF + Pose Prior	.96	.98	**.93**	**.71**	**.97**	**.88**	**.89**
RF + Oracle Matching	.97	.99	.94	.82	.98	.97	.94

Table 1: A comparison of PCP scores of different baselines on our football dataset. The rows represent the results of the following methods. (1) FMP [21] trained and tested on our dataset. (2) Taking the optimal modes for each joint independently. (3) Taking the modes that maximise a shape prior. (4) Taking the optimal modes wrt the ground truth. For the last two baselines the matching is performed only on a few of the most probable modes (5-10).

3.2 Scoring and inference in 3D

To perform 3D pose estimation we follow the approach of [4] and discretize the search space. We assume that the person is within a bounding cube (fig. 1) and create a $64 \times 64 \times 64$ grid covering this cube. We compute our 3D part appearance likelihoods for all grid points. We perform inference with and without the pose prior discussed in section 2.4. The former imposes limb length and intersection constraints. We also perform inference with and without the latent variable handling the mirror ambiguity as discussed in section 2.5.

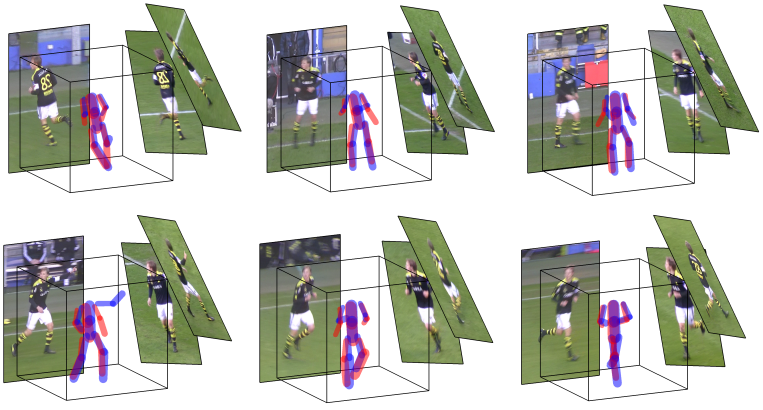


Figure 6: Final 3D poses obtained by taking, for each part independently, its most probable state over the grid. The mirror ambiguity is solved jointly. Estimation is red and ground truth is blue.

The results are summarized in table 2. The performance is measured using 3D PCP scores with $\alpha = 0.5$ [4]. The table shows that introducing the latent variable to deal with the mirror ambiguity significantly improves the final results. On this dataset it is surprisingly much more important than the pose prior. Figure 6 shows our estimated 3D poses (red) compared to the ground truth (blue), for six different frames. For this figure the inference was performed using the latent mirror variable but without any pose prior (uniform). The figure shows that our 3D appearance likelihoods accurately detect most of the body parts, even without imposing any pose prior. If we add the limb length and intersection constraints we are able to correct for some of the limited double counting that occurs for the lower legs, which is reflected by numbers in table 2.

	Upper Arms	Lower Arms	Upper Legs	Lower Legs	Average
RF	.02	.03	.86	.57	.37
RF + Pose Prior	.16	.07	.91	.87	.50
RF + Mirror Latency	.87	.68	1.00	.96	.88
RF + Mirror Latency + Pose Prior	.89	.68	1.00	.99	.89

Table 2: An evaluation of our 3D pose estimation results in terms of PCP scores. The rows represent the results of the following methods. (1) Taking the maximum probability estimates for each part independently over the 3D grid. (2) Taking the pose priors into account. (3) Handling mirror ambiguity without pose priors and (4) with pose priors.

4 Conclusion

In this paper we have discussed multi-view human pose estimation using part-based models. We have focused on the part appearance component of such models. We believe that 2D part detectors based on random forest classification are simple to implement and efficient at test-time. We achieve state-of-the-art performance on our new large football dataset. Yet, dealing with small datasets with severe background clutter can be challenging for our method which we would like to address in future work.

When combining the 2D part detectors over multiple views for 3D part detection, the similar appearance of mirror symmetric body parts is a problem. We have highlighted this and presented a simple and surprisingly accurate solution based on a latent variable formulation. Our resulting multi-view part detectors can be used by any multi-view part-based model. We have shown that they allow 3D pose estimation outside the studio, in a professional football game, without relying on strong priors for motion or 3D pose. We hope that our new football dataset will stimulate more research of 3D pose estimation in realistic outdoor environments.

Acknowledgement This work was supported by the FP7 project "Free-viewpoint Immersive Networked Experience". The authors would like to thank AIK Football Club and Hego Tracab for help with collecting the football footage.

References

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2009.

[2] M. Bergtholdt, J. Kappes, S. Schmidt, and C. Schnörr. A study of parts-based object class detection using complete graphs. *Internation Journal of Computer Vision*, 87(1): 93–117, 2010.

[3] M. Burenius, J. Sullivan, and S. Carlsson. Motion capture from dynamic orthographic cameras. In *4DMOD - ICCV Workshop*, 2011.

[4] M. Burenius, J. Sullivan, and S. Carlsson. 3d pictorial structures for multiple view articulated pose estimation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2013.

[5] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2005.

[7] M. Dantone, J. Gall, C. Leistner, and L. Van Gool. Human pose estimation using body parts dependent joint regressors. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2013.

[8] P. F. Felzenszwalb, D. A. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2008.

[9] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9), Sept. 2010.

[10] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[11] T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer-Verlag, 2001.

[12] S. Johnson and M. Everingham. Learning effective human pose estimation from inaccurate annotation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.

[13] Vahid Kazemi and Josephine Sullivan. Using richer models for articulated pose estimation of footballers. In *Proceedings of the British Machine Vision Conference*, pages 6.1–6.10, 2012.

[14] T.B. Moeslund, A. Hilton, V. Krüger, and L. Sigal. *Visual Analysis of Humans: Looking at People*. Springer, 2011. ISBN 9780857299963.

[15] Long Quan. Self-calibration of an affine camera from multiple views. *Internation Journal of Computer Vision*, 19:93–105, July 1996.

[16] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.

[17] H. Sarmadi. Human detection and pose estimation in a multi-camera system. *Master's Thesis at KTH Royal Institute of Technology, Sweden*, 2013.

[18] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.

[19] L. Sigal, M. Isard, H. Haussecker, and M. Black. Loose-limbed people: Estimating 3d human pose and motion using non-parametric belief propagation. *Internation Journal of Computer Vision*, 98(1):15–48, 2012.

[20] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *Internation Journal of Computer Vision*, 9:137–154, November 1992. ISSN 0920-5691.

[21] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 2011.