Using Richer Models for Articulated Pose Estimation of Footballers

Vahid Kazemi
vahidk@nada.kth.se
Josephine Sullivan
sullivan@nada.kth.se

CVAP
KTH, The Royal Institute of Technology
Stockholm, Sweden

Abstract

We present a fully automatic procedure for reconstructing the pose of a person in 3D from images taken from multiple views. We demonstrate a novel approach for learning more complex models using SVM-Rank, to reorder a set of high scoring configurations. The new model in many cases can resolve the problem of double counting of limbs which happens often in the pictorial structure based models. We address the problem of flipping ambiguity to find the correct correspondences of 2D predictions across all views. We obtain improvements for 2D prediction over the state of art methods on our dataset. We show that the results in many cases are good enough for a fully automatic 3D reconstruction with uncalibrated cameras.

1 Introduction

This work tackles the problem of automatically reconstructing the 3D pose of a person, in particular a football player, from multiple images taken from uncalibrated affine cameras. We adopt a bottom up approach, summarized as, localize the skeletal 2D joints in each image independently and then perform factorization with limb length constraints to estimate the 3D pose. The joint localization task is the more challenging part and is the paper's main focus.

Localization of a person's limbs in an image is very difficult for a myriad of reasons, most notably the range of articulations of the person (especially true in sports footage), self-occlusion, foreshortening of limbs and motion blur. However, in recent years significant progress has been made with the introduction of pictorial structure type models using discriminatively learned parts [3, 6, 15]. These models compromise between accurate modeling of the underlying flexibility in the appearance and spatial configuration of the person's limbs and computational concerns to make the parameter learning and the inference tractable.

Despite this progress, though, the results are far from perfect in real world scenarios. Figure 1(a) shows the results from the state-of-the-art *Flexible Mixture of Parts* (FMP) model [15] on images from KTH multiview football dataset. The right of figure 1(a) shows an example of a common failure. The problem is partly due to the simplifications made in the modeling. However, the main observation exploited in this paper is that while the *true configuration* might not always correspond to the global optimum of the FMP's cost function, it frequently gets a high score. One can observe this by examining figure 1(b). It shows that

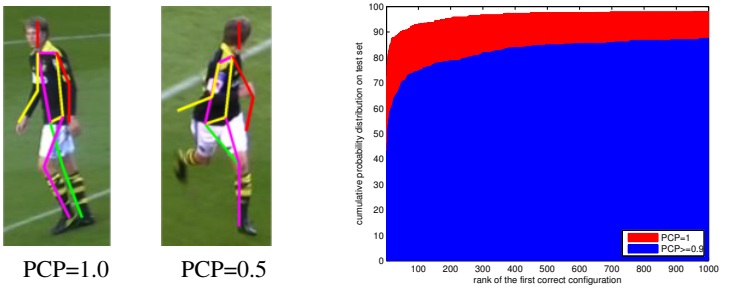


Figure 1: **(a)** Shown is the top scoring configuration returned by the FMP model and its PCP score for two images. The PCP score is the proportion of correctly localized limbs. **(b)** This is a cumulative histogram of the rank of the first correctly predicted pose by the FMP model. In 36% of the test cases the top scoring configuration has PCP=1. While 88% of the time there exists a configuration with PCP=1 in the top scoring 1000 configurations. These percentages change to 68% and 98% when the definition of a correct configuration is lowered to having PCP ≥ 0.9.

on our football dataset a correct configuration - all the parts are localized correctly - is in the top 1000 scoring configurations w.r.t. the FMP cost function 88% of the time, while the top scoring configuration is a correct configuration only 36% of the time.

As a correct configuration is frequently in the set of the top n scoring configurations w.r.t. the simplified (FMP) scoring function and it is straightforward to obtain these configurations [9], we only need to evaluate a more accurate and arbitrarily complex scoring/re-ranking function on this small set. This is the general strategy we adopt. In this paper we learn this re-ranking function and describe the components it includes. While the latter part of the paper presents a road map of how to put the arms and legs in correspondence (solving the left/right ambiguity) across the multiple images in order to allow a 3D reconstruction.

Our main contributions are: **1)** We introduce a new model which is an extension to [15]. It utilizes a global segmentation score, extra pairwise terms, and an exclusion principle to avoid double counting the score of overlapping parts. The overhead of our model over the FMP model is very small as our search space is a relatively small constant number. **2)** We present an effective parameter learning procedure based on the SVM-Rank formulation [7] to calibrate the factors included in our re-ranking function. **3)** We present a first attempt to automatically and accurately solve the 3D reconstruction from multiple view images in a non-studio environment. **4)** We present a new dataset of 771 images of football players taken from 3 views at 257 time instances, which will be publicly available on the author's website.

1.1 Related Work

By imposing a few assumptions on the pictorial structure model - independent appearance scores, quadratic deformation function - [3] developed an algorithm that finds the global optimum of the pictorial structure energy function in linear time complexity to the number of locations on the image. Using discriminatively trained parameters [1, 2] within this model produces very good results. There has been a few attempts on extending the model to handle inaccurate annotations using latent parameters [2, 8]. [10] tries to improve the pose priors by using a local kernel regression model. [11] proposes a cascade model for enabling the use

of more sophisticated appearance models. [12] uses a more complicated graphical model to model extra dependencies between parts, and utilizes an approximate belief propagation algorithm to do the inference. Flexible Mixture of Parts (FMP) model [15] uses multiple linear models to represent the appearance of the object. We use the FMP model as the base of our work which has outperformed all the previous work by a significant margin. The paper [9] describes an efficient algorithm to approximately compute a set of high scoring configurations with almost no extra cost. Commonly automatic 3D pose reconstruction is performed by tracking with a 3D model [2] or applying a learnt regression function which maps an extracted image feature to a 3D pose [13]. However, due to the developments in 2D pose estimation it has allowed us to explore in this paper the automatization of previously semi-manual based algorithms using 2D joints [14].

2 Components of a more accurate scoring function

Given the n-best configurations returned by the FMP model, the challenge is to re-score them in order to identify the ones which are closest to a correct configuration. The re-ranking function we learn is a linear combination of different features which indicate - weakly or strongly - the plausibility of a hypothesized configuration. In this section we describe the features and measurements which are extracted. These include a global segmentation score measuring how compatible a hypothesized configuration is with a segmentation of the image into foreground and background based on colour and a re-weighting of part appearance scores to impose an exclusion principle to avoid double counting the score of overlapping parts. First, though, we review the scoring function of the FMP model [15]. Many of its individual components are included in our re-ranking function but computed on a graph defining the dependency structure which includes loops.

2.1 Review of the flexible mixture of parts model

In the flexible mixture of parts (FMP) model [15] the object is divided into multiple parts, and each part is modelled by a set of templates. A graph structure, $G = (V,E)$, represents the dependencies used when fitting this model. V is the set of parts and E is the set of edges indicating which parts are linked. The coordinates of the centre of the ith part is denoted by p_i and $p = (p_1,\ldots,p_K)$ is the vector of all the part centres. Each part is also assigned a template t_i where each $t_i \in \{1,\ldots,T\}$ and let $t = (t_1,\ldots,t_K)$. The FMP model then scores a configuration p and its associated part types t with

$$S_{\text{fmp}}(p,t) = S_{\text{a}}(p,t) + S_{\text{d}}(p,t) + S_{\text{c}}(t). \tag{1}$$

which has three distinct components. $S_{\text{a}}(p,t)$ is a weighted sum of appearance scores for each part

$$S_{\text{a}}(p,t) = \sum_{i\in V} s_{\text{a}}(p_i,t_i) = \sum_{i\in V} w_i^{t_i} \cdot \phi(I,p_i), \tag{2}$$

where $\phi(I,p_i)$ is a HOG descriptor of the image patch centred at p_i and $w_i^{t_i}$ is the template for ith part of type t_i. $S_{\text{d}}(p,t)$ is the deformation score

$$S_{\text{d}}(p,t) = \sum_{e\in E} s_{\text{d}}(p_e,t_e) = \sum_{e=(i,j)\in E} w_{ij}^{t_i,t_j} \cdot \psi(p_i - p_j), \tag{3}$$

and is a sum of quadratic functions ($\psi(dx,dy) = [dx\ dx^2\ dy\ dy^2]$) modeling the deformation between connected parts. While $S_c(t)$ is the score which consists of a prior for each part type and a compatibility score between the types of connected parts

$$S_c(t) = \sum_{i \in V} s_c(t_i) = b^{t_{\text{root}}}_{\text{root}} + \sum_{i \in V \setminus \text{root}} \left(b^{t_i}_i + b^{t_{i,\text{parent}(i)}}_{i,\text{parent}(i)} \right). \tag{4}$$

Using the generalized distance transform and assuming G is a tree one can efficiently find the configuration, $(p_{\text{fmp}}, t_{\text{fmp}})$ which maximizes $S_{\text{fmp}}(p,t)$ and the configurations corresponding to the n-best scores of $S_{\text{fmp}}(p,t)$. The top scoring configuration frequently has a high PCP score and in general the head, torso and one leg are reliably detected. The problem of *double counting*, though, is prevalent. To help combat this issue, we include in our re-ranking function the same individual deformation scores, defined in equation (3), but augment these with deformation scores between pairs of left and right parts, see figure 4.

2.2 Modelling the correlation between parts

As we only focus on the n-best configurations returned by the FMP model we are at liberty to exploit more complicated and computationally expensive scoring of a configuration. Here we describe the scores we compute that are not facsimile of those in the FMP model. The first is a re-weighting of the individual appearance scores in equation (2) to prevent the double counting of evidence. The second is one based on performing crude segmentation. The crucial factor in both is that we allow ourselves to consider the global configuration p simultaneously as opposed to only considering pairs of parts at a time.

2.2.1 Enforcing an exclusion principle

Double counting occurs frequently in the football data, for instance when one of the legs is in motion and appears blurry while the other is stationary. In this situation the FMP or any pictorial structure model commonly double counts the strong evidence (usually the stationary limb) due to the independence assumptions they make. It is necessary to take the visibility of each part into account to allow for a more accurate modeling of the underlying situation and to implicitly enforce an exclusion principle. We employ probabilistic reasoning to do this modeling. Let sets $\mathcal{S}_{p,1}, \ldots, \mathcal{S}_{p,L}$ partition the set of K parts. Each $\mathcal{S}_{p,l}$ either contains the left and right versions of a part or just one single part for the parts associated with the head and torso. Let $p_{\mathcal{S}_{p,l}}$ denote the positions of the parts in $\mathcal{S}_{p,l}$, similarly for $t_{\mathcal{S}_{p,l}}$ and $I_{\mathcal{S}_{p,l}}$ is the region of the image I which corresponds to where the parts in $\mathcal{S}_{p,l}$ occur. If the parts in $\mathcal{S}_{p,l}$ do not overlap then the likelihood of $I_{\mathcal{S}_{p,l}}$ is

$$p(I_{\mathcal{S}_{p,l}} \mid p_{\mathcal{S}_{p,l}}, t_{\mathcal{S}_{p,l}}) = \prod_{k \in \mathcal{S}_l} p(I_{p_k} \mid p_k, t_k) \tag{5}$$

However, if the parts in $\mathcal{S}_{p,l}$ overlap then the likelihood is calculated differently. As we do not know which part is the closest to the camera, we cycle through the different possibilities to get

$$p(I_{\mathcal{S}_{p,l}} \mid p_{\mathcal{S}_{p,l}}, t_{\mathcal{S}_{p,l}}) = \sum_{k \in \mathcal{S}_l} p(I_{p_k} \mid p_k, t_k)\, P(\text{part } p_k \text{ is the most visible part in } \mathcal{S}_{p,l}) \tag{6}$$

where for simplicity it is assumed that only one of the parts in $\mathcal{S}_{p,l}$ is visible at a time. If it is assumed that each $p(I_{p_k} \mid p_k, t_k) \propto \exp(s_a(p_k, t_k))$ and each part in $\mathcal{S}_{p,l}$ is equally likely to be the one visible, then we can define scores which mimic $p(I_{s_{p,l}} \mid p_{s_{p,l}}, t_{s_{p,l}})$:

$$s_{l,\text{joint}}(p,t) = \begin{cases} \log\left(\frac{1}{|\mathcal{S}_l|}\sum_{k\in\mathcal{S}_l}\exp(s_a(p_k,t_k))\right) & \text{if parts in } \mathcal{S}_l \text{ overlap} \\ \sum_{k\in\mathcal{S}_l} s_a(p_k,t_k) & \text{otherwise} \end{cases} \tag{7}$$

2.2.2 Segmentation score

A configuration p produces a segmentation of the image into background and foreground pixels. One can then measure the plausibility of configuration p by comparing this segmentation to one produced by another independent process. In our case this independent process segments based on comparing the colour of each pixel to learnt distributions of the colour for background and foreground pixels. We learn these foreground and background distributions for each test image with the following procedure. The high scoring configurations returned by the FMP model are used to create an initial estimate of the segmentation into foreground and background, see figure 2. This is done simply by averaging the foreground masks created from the boxes representing the parts in each configuration. The result is a rough estimate of the probability of a pixel belonging to the foreground. Thresholding these probabilities with separate criteria gives an under and over-segmentation. The foreground pixels from the under-segmentation are used to fit a GMM distribution for foreground pixels

$$p(c_x \mid l_x = f) = \sum_{i=1}^{M_f} \alpha_i^f \mathcal{N}(c_x \mid \mu_i^f, \Sigma_i^f) \tag{8}$$

where c_x is the RGB colour of a pixel at location x and l_x is the pixel's label as foreground or background. Similarly the background pixels from the over-segmentation are then used to fit a GMM distribution representing $p(c_x \mid l_x = b)$. Assuming a uniform prior probability, the posterior probability of pixel being foreground given its colour is

$$P(l_x = f \mid c_x) = \frac{p(c_x \mid l_x = f)}{p(c_x \mid l_x = f) + p(c_x \mid l_x = b)} \tag{9}$$

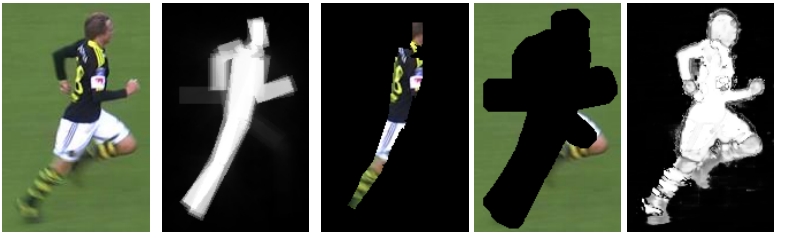
We aggregate these individual posterior probabilities into a plausibility score of p based on its agreement with the segmentation

$$s_{\text{seg}}(p) = \frac{1}{N}\left(\sum_{x\in\mathcal{F}_p} P(l_x = f \mid c_x) + \sum_{x\in\mathcal{B}_p} P(l_x = b \mid c_x)\right) \tag{10}$$

where N is the total number of pixels, $\mathcal{F}_p$ is the set of pixels labeled as foreground according to p and similarly $\mathcal{B}_p$ is the background set.

3 Learning the parameters of the re-ranking function

In the previous section we introduced scores which indicate the plausibility of the person's hypothesized 2D pose. The next task is to combine these within one single function which can be used to re-rank the n-best configurations output by the FMP model. To this end



(a) input (b) initial mask (c) under seg. (d) over seg. (e) $p(l_x = f \mid c_x)$

Figure 2: To estimate the initial segmentation of an image (a) we use the top scoring configurations from the FMP model to get an initial estimate of the probability of a pixel belonging to the foreground (b). The results are then used to create under (c) and over (d) segmentation masks. A GMM is fit to both the colours of the foreground pixels and the background pixels. These distributions are then used to compute the posterior probability of each pixel being foreground (e).

we construct a feature vector $x_{p,t}$ for each (p,t) by concatenating the different components already described:

$$x_{p,t} = (s_{\text{seg}}(p), s_{1,\text{joint}}(p,t), \ldots, s_{L,\text{joint}}(p,t), s_{\text{d}}(p_{e_1}, t_{e_1}), \ldots, s_{\text{d}}(p_{e_l}, t_{e_l}), s_{\text{c}}(t_1), \ldots, s_{\text{c}}(t_K)) \tag{11}$$

where the edges $e_i \in E$ are now taken from a graphical model of the pairwise dependencies between parts with loops, see figure 4. We let the final scoring function take the form of a weighted sum of the individual components of $x_{p,t}$:

$$\text{score}(p,t) = w \cdot x_{p,t} \tag{12}$$

Our objective is to learn the linear weights w such that configurations closer to the ground truth are scored more highly. Closeness to the ground truth can be measured by the PCP score [5]. This measure returns 1 if each part of the hypothesized configuration overlaps significantly with its corresponding part in the ground truth configuration. Our training data consists of N training images. For each training image I_k we calculate the n-best configurations returned by the FMP model. Each of these configurations generates a feature vector x_{ki} and let y_{ki} denote its PCP score. Let r_k be a subset of the pairwise constraints imposed by the ranking of x_{ki}'s based on y_{ik}:

$$r_k = \{(x_{ki}, x_{kj}) \ : \ y_{ki} > .9 \ \text{ and } \ y_{kj} \leq .9\} \tag{13}$$

Then we find the optimal w by minimizing the SVM-Rank[7] objective function:

$$\arg\min_{w,\xi_{ijk}} \frac{1}{2}\|w\|^2 + C\sum_{i,j,k}\xi_{ijk} \tag{14}$$

subject to for $k = 1, \ldots, N$

$$w \cdot x_{ki} \geq w \cdot x_{kj} + 1 - \xi_{ijk} \quad \forall (x_{ki}, x_{kj}) \in r_k \quad \textbf{and} \quad \xi_{ijk} \geq 0 \quad \forall (i,j,k) \tag{15}$$

Note the formulation is similar to that of the SVM, but the set of constraints has been extended to enforce the correct ordering between all pairs of configurations within each r_k. The main reason for using the SVM-rank model instead of a regular SVM is that the absolute value of our target function is not an accurate quantitative measure, but we assume the measure is accurate enough for comparing two configurations from the same image. To do the optimization we used the publicly available cutting-plane solver from [7].

4 3D Reconstruction

To estimate a player's 3D pose we must put his arms and legs in correspondence across the three views. This is because the current 2D pose model cannot distinguish between the real left and right limbs. There are 32 possible correspondences, ignoring mirrored configurations. We reconstruct the position of the skeletal joints in 3D for each of these combinations and the 2D joint locations highlighted by our re-ranking function. The correspondence which results in a plausible 3D pose - estimated 3D skeleton has limb lengths similar to those estimated during training - and gives the smallest re-projection error is then chosen. To do the reconstruction, first we compute an initial estimate of the 3D pose, $\tilde{X}$, and camera matrices, $\tilde{M}$, using the affine factorization algorithm. These quantities must then be rectified and therefore we seek an affine transformation, A, which transforms $\tilde{X}$ and $\tilde{M}$ to the true 3D locations and camera matrices. A is estimated by minimizing a cost function which softly enforces that each limb of the rectified 3D skeleton has the same length as observed in the training data. We use MATLAB's standard nonlinear optimization toolbox to perform this.

5 Results

We have annotated a total of 771 images of football players, which includes images taken from 3 views at 257 time instances. We used 180 of the images for training our model and the rest for testing. Figure 3 shows three annotated examples from our football dataset.

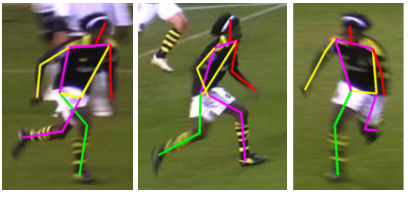


Figure 3: Three annotated examples from our football dataset which are taken at the same time instance.

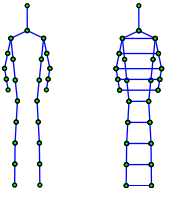


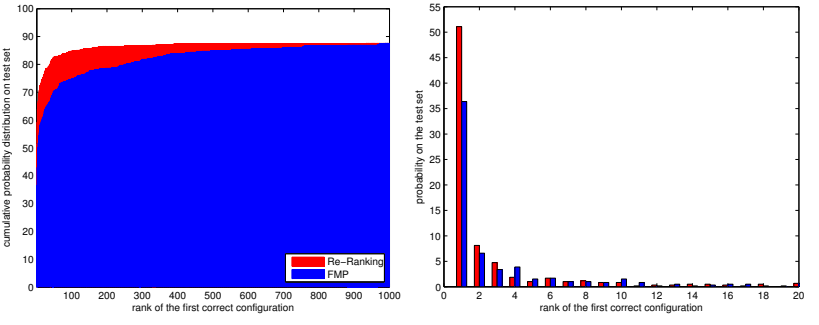
Figure 4: The pairwise dependencies in the FMP model (left), compared to the one used in the re-ranking function (right).

Table 1 summarizes the results on our dataset with and without using the re-ranking function, as well as the results of picking the closest configuration to the ground truth between top 1000 configurations. In addition to the standard PCP score, we have provided the PCP scores ignoring the left/right limb assignments. This criteria is more accurate for our dataset since the limbs annotated as left/right on 2D images do not represent the real left/right limbs of the person. The results are improved by 3.3% with the PCP score criteria and 4.1% if we ignore the flipping. Figure 5(a) shows the cumulative probability distribution of rank of the true configuration across the top 1000 configurations given by the FMP model, in comparison with the results with our model. Figure 5(b) shows the same results on a finer scale. We can observe that the probability of the true configuration getting the top score based on FMP model is 36%, while this probability is increased to 51% using our model (an oracle ranking function in this case could improve the results up to 88%).

Figure 6 shows some qualitative results from our experiments on our football dataset. We observed that in many cases the double counting problem is fixed using our model (1-2nd

Ranking function	left/right flips **not** ignored	left/right flips ignored
Flexible Mixture of Parts	0.884	0.895
Re-ranking SVM-Rank	0.917	0.936
Oracle re-ranking	0.982	0.982

Table 1: Summary of the results on our football dataset with and without the re-ranking function. The first column of numbers displays the average PCP score of the top scoring configuration returned by the FMP model, our learnt re-ranking function and an oracle re-ranking function. The second column is the average PCP score when the left and right labels for the arms and legs are ignored.



(a) Cumulative histogram (b) Histogram on finer scale

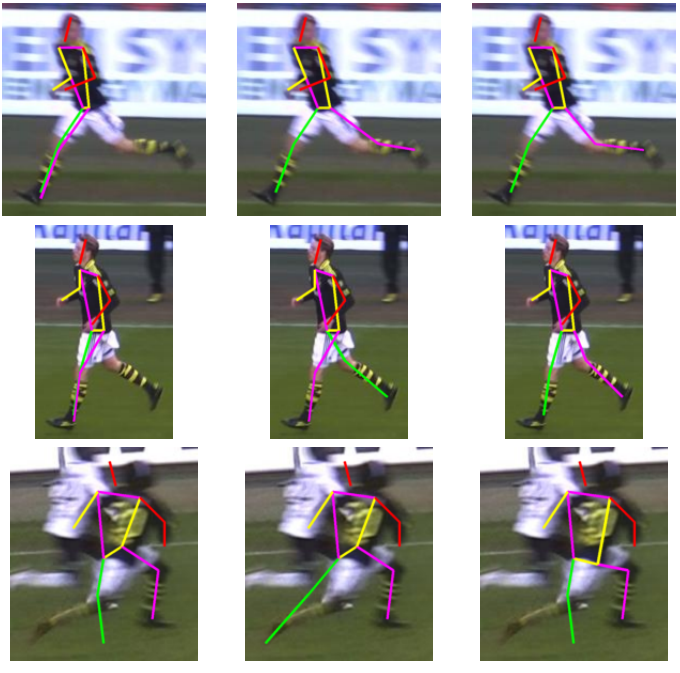
Figure 5: (a) The cumulative histogram of the rank of the first correctly predicted pose by the flexible mixture of parts model before (blue) and after reranking (red). (b) The same histogram on a finer scale.

rows). While in some cases the predicted flip is not compatible with the ground truth (2nd row) and this is the reason for the additional improvements if we ignore the flipping. The measurements in some cases are too noisy for our model, and we do not observe much of an improvement in these cases (3rd row). Finally, we have used the 2D estimates from our model to reconstruct the configuration of the player in 3D. With no assumptions about the pose of the player this is an extremely difficult task. However, when we have fairly good 2D estimates across all views we are able to get reasonable results. Figure 7 shows a stick figure of the 3D reconstruction of the top scoring 2D configurations, along with the back projected 2D estimates.

5.1 Conclusions

We described a simple and efficient way of improving the performance of part based models by evaluating a more complicated scoring function to reorder a set of high scoring configurations. With good enough predictions of the location of a set of body joints across three images, we can obtain fairly accurate estimation of camera parameters and 3D joint positions. We believe by enforcing the temporal continuity constraints over sequences of images we can expect a boost in robustness and accuracy of our 3D predictions, which will be the subject for a future work. We would also like to exploit a multi-modal ranking function as opposed to a linear model which we have utilized in this work.

Acknowledgement: This work has been funded by the European Commission within the project FINE (Free Viewpoint Immersive Networked Experience).



(a) FMP (b) Reranking (c) Closest to G.T.

Figure 6: This figure shows (a) the result of FMP compared to (b) our reranking function, in addition to (c) the results of picking the closest configuration to the ground truth from a set top 1000 configurations. In many cases (row 1-2) we can solve the double counting problem, but sometimes (row 2) we have problem with the flipping ambiguity. In the last case the measurement is too noisy for our model and we are not able to improve the results.

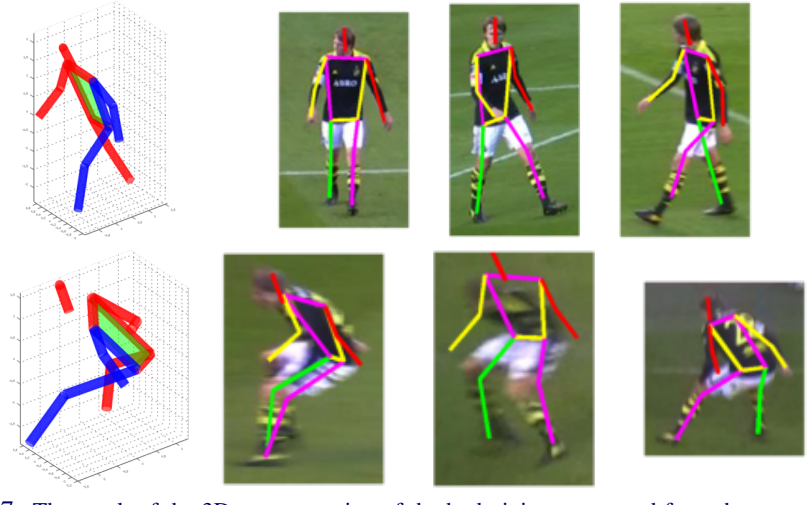


Figure 7: The result of the 3D reconstruction of the body joints computed from the top scoring 2D configurations, along with the back projected 2D estimates.

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] J. Deutscher and I. Reid. Articulated body motion capture by stochastic search. *International Journal of Computer Vision*, 61:185–205, 2005.

[3] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

[4] 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.

[5] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proceedings of the Conference on Computer vision and Pattern Recognition*, 2008.

[6] O. Firschein and M. A. Fischler. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.

[7] T. Joachims. Optimizing search engines using clickthrough data. In *Knowledge Discovery and Data Mining*. ACM, 2002.

[8] 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.

[9] D. Park and D. Ramanan. N-best maximal decoders for part models. In *Proceedings of the International Conference on Computer Vision*, 2011.

[10] B. Sapp, C. Jordan, and B. Taskar. Adaptive pose priors for pictorial structures. In *Proceedings of the Conference on Computer vision and Pattern Recognition*, 2010.

[11] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *Proceedings of the European Conference on Computer Vision*, 2010.

[12] L. Sigal and M. J. Black. Measure locally, reason globally: Occlusion-sensitive articulated pose estimation. In *Proceedings of the Conference on Computer vision and Pattern Recognition*, 2006.

[13] A. Thayananthan, R. Navaratnam, B. Stenger, P. Torr, and P. Cipolla. Pose estimation and tracking using multivariate regression. *Pattern Recognition Letters*, 29(9):1302–1310, 2008.

[14] P. Tresadern and I. Reid. Uncalibrated and unsynchronized human motion capture: A stereo factorization approach. In *Proceedings of the Conference on Computer vision and Pattern Recognition*, 2004.

[15] 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.