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

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 [1, 2, 4]. 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 [4] on images from our 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 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 [3], we only need to evaluate a more accurate and arbitrarily complex scoring/re-ranking function on this small set.

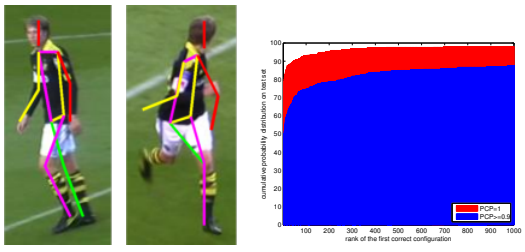

Figure 1: **(a)** Shown is the top scoring configuration returned by the FMP model and its PCP score for two images. The PCP score is the proportion of correctly localized limbs. **(b)** This is a cumulative histogram of the rank of the first correctly predicted pose by the FMP model. In 36% of the test 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.

Since we only consider the n-best configurations returned by the FMP model we are at liberty to exploit more complicated and computationally expensive scoring of a configuration. Our new model re-weights appearance scores from the FMP model to prevent the double counting of evidence. We also use the colour distribution of foreground and background to penalize configurations which do not explain all the foreground. The crucial factor here is that we allow ourselves to consider the global configuration simultaneously as opposed to only considering pairs of parts at a time.

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.

To evaluate our method we have annotated a dataset of 771 images of football players, which includes images taken from 3 views at 257 time instances. 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. It can be seen that in both case using a re-ranking function improves the performance comparing to the state of the art FMP model. The difference is much more significant if we only compare the top scoring configuration. 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%).

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 2 shows a stick figure of the 3D reconstruction of the top scoring 2D configurations, along with the back projected 2D estimates.

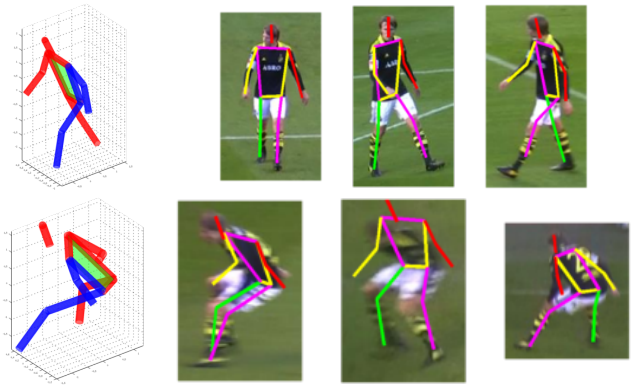

Figure 2: The result of the 3D reconstruction of the body joints computed from the top scoring 2D configurations, along with the back projected 2D estimates.

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

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

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

[4] 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.