Articulated Human Detection and Pose Estimation (CS365 Course Project)

Anant Raj
Dept. of Electrical Engineering
IIT Kanpur
Email: anantraj@iitk.ac.in

Triya Bhattacharya
Dept. of Computer Science and Engineering
IIT Kanpur
Email: triya@cse.iitk.ac.in

Mentor - Amitabha Mukerjee
Dept. of Computer Science and Engineering
IIT Kanpur
Email: amit@cse.iitk.ac.in

***Abstract*—The aim of the project is to estimate the pose of Articulated Human in static 2D images with flexible mixture of parts [1]. The main idea behind this work is that "mini part" model can approximate deformation as described in [1]. Asssuming that the co occurrence of various parts and spatial relations between them is a tree structured graph then the use of dynamic programming can efficiently optimize the performance of the proposed model. This model ultimately provides the result with highest achieved score but tracing it back from the top also results in detection of multiple humans in an static image. Experiments are done on standard datasets and also on self shooted pictures during various sports events. Along with these, experiments are also done on random images taken from the web which is not of human but looks like human (statue or paintings etc.) and analysis on those images has also been done.**

I. Introduction

Articulated pose estimation is one of the core problems in object detection and Computer Vision. Any improvement in this area will lead to solve some challenging problem and will immediatelly impact in the area of Human- Machine interface and intelligent systems. That's why people have devoted a large amount of time in doing research to correctly estimate the articulated pose.

A very early and classic work in the area of object detection is to represent object in pictorial structural framework [2]. After that there became a long tradition of using pictorial structure for human [3], [4]. In this approach the appearance of object is decomposed in local part templete and a geometric constraints is imposed on the joints. But still full body pose estimation can't be done effectively using this methos also because of the many degrees of freedom present in the human body. Along with this the appearance of limbs is also continuously changing with clothing and angle of view. These difficulties complicate inference as one must typically search images with a large number of warped (rotated and foreshortened) templates [1]. Taking these problems into account a model containing the mixture of pictorial structure with small non oriented parts is proposed [1]. This model can also be very efficient in modelling any type of articulated object apart from human body.

The rest of the document is organised as follows. A brief discussion on related work is described in the next section. After that motivation and model is discussed. Inference and learning part is discussed in the subsequent section. Last two sections are dedicated to analysis of results and conclusion.

II. Related Work

Pose estimation has been an active area of research in the field of computer vision and machine intelligence. Recently people are working on tracking human pose on video [5] and are in search of an efficient method which can initiate the tracking by initially detecting the human skelton in a static image [6]. Various amount of work are being done on different aspects of human pose estimation.

Encoding of spatial structure has been actively attracting the people to work on it. Various models have been proposed to represent the human skelton as a tree structure using probabilistic graphical model because this model allows efficient inference [1], [7]. But inspite of efficient inference this method itself has a problem of double counting [1]. Methods are proposed to overcome this demerit using loopy constraints [8], [9]. Along with this, works are being done on various aspects of learning for an efficient model learning of human pose [10], [11]. Also works are being done to come up with an efficient representation of image features.

III. Motive and Model [1]

Almost most of the work done in this area tries to approximate the human deformation as mixture of templates and so this approach also tries to represent human deformation as a flexible mixture of parts [1]. If we assume x as a 2D pixel position of a given template and $w(x) = (I + \Delta A)x + b$ is the new position of that pixel under small deformation or translation to that template. If $s(x)$ is defined as difference between $w(x)$ and x then as in the paper [1]

$$\begin{aligned} s(x + \Delta x) &= w(x + \Delta x) - (x + \Delta x) \\ &= (I + \Delta A)(x + \Delta x) + b - x - \Delta x \\ &= s(x) + \Delta A \Delta x \end{aligned}$$

So if $\Delta A \Delta x$ is very small than x and $x + \Delta x$ shifts by the same amount during transformation.The above statement holds true if any of ΔA or Δx is very small. In the previous work (pictorial model) they have taken ΔA as a small quntity and used a large number of dicretized articukated tempalte where each template differ from the other in a very small amount of forshortening and rotation. But in this case Δx is assumed to be small , so dividing a large parts in small small parts.

For any image I, $l_i = (x, y)$ is defined as the pixel location of part i and t_i is the mixture component of part i. A compatibility function (co-occurrence bias) is defined for part types which factors into a sum of local and pair wise score [1]

$$S(t) = \sum_{i\in V} b_i{}^{t_i} + \sum_{ij\in E} b_ij^{t_i,t_j} \tag{1}$$

where parameter $b_i{}^{t_i}$ favor particular type assignment for part i and $b_{ij}{}^{t_{ij}}$ favors particular co-occurrence of part type i and j as described in [1]. If two parts i and j favors consistent assignment then the value $b_{ij}{}^{t_{ij}}$ is a large positive number otherwise a large negative number. This model can force a collection of parts a particular orientation only as long as they are the part of (subtree) of connected graph $G(V,E)$ [1] and by posing this constraints this model learn the rigidity.

Full score associated with a configuration is defined as in [1]

$$S(I,l,t) = S(t) + \sum_{i\in V} w_i{}^{t_i}.\phi(I,l_i) + \sum_{ij\in E} w_ij^{t_i,t_j}.\psi(l_i - l_j) \tag{2}$$

where $\phi(I,l_i)$ is HOG feature vector extracted from pixel location l_i in image I and $\psi(l_i - l_j)$ = $[dx \;\; dx^2 \;\; dy \;\; dy^2]^T$, where $dx = x_i - x_j$ and $dy = y_i - y_j$, the relative location of part i with respect to j. In the above equation first term corresponds to the local apearance model for a particular part and the seocond term corresponds to the deformation which can occur due to various reasons. It can be observed that the proposed method here is more general so by taking special cases various models can be derived from this model only.

IV. Inference and Learning [1]

During inference we tend to maximize the total score $S(I,l,t)$ over l and t. Since already the human skelton has been modeled as a graphical tree structure. So, the score

$$S(I,z) = \sum_{i\in V} \phi_i(I,z_i) + \sum_{ij\in E} \psi_{ij}(z_i,z_j) \tag{3}$$

where $\phi_i(I,z_i)$ = $w_i{}^{t_i}$. $\phi(I,l_i)$ + $b_i{}^{t_i}$
and $\psi_{ij}(z_i,z_j)$ = $w_ij^{t_i,t_j}$. $\psi(l_i - l_j)$

can be maximized using dynamic programming starting from the leaf node. More precisely, the message part i passes to its parent j is computed by the following as in [1]

$$score_i(z_i) = \phi_i(I,z_i) + \sum_{k\in kids(i)} m_k(z_i) \tag{4}$$

$$m_i(z_j) = \max_{z_i}[score_i(z_i) + \psi_{ij}(z_i,z_j)] \tag{5}$$

”Eq.(4) computes the local score of part i, at all pixel locations li and for all possible types ti , by collecting messages from the children of i and Eq.(5) computes for every location and possible type of part j, the best scoring location and type of its child part i ” [1].

Once the message reaches to its root then the score represents thr best score for each root position. These root scores can be used to detect more than one human skeleton in the image [1]. N-best extensions of dynamic programing is very helpful in detecting multiple people in an image [13].

Supervised learning algorithm is applied to learn the model parameters. The cost function is defined as in [1] , [14] over given poitive labeled and negative examples. It can be noted easily that the score function is linear in model parameter β = (w,b). So the learned model will be of the form as in [1]

$$arg \min_{\omega,\xi_n \geq 0} \frac{1}{2}\beta.\beta + C\sum_n \xi_n \tag{6}$$

$$such \quad that \quad \forall n \in pos \quad\; \beta\phi(I_n,z_n) \geq 1 - \xi_n \tag{7}$$

$$and \quad\; \forall n \in neg, \forall z \beta\phi(I_n,z) \leq -1 + \xi_n \tag{8}$$

The above constraint states that positive examples should score better than 1 (the margin), while negative examples, for all configurations of part positions and types, should score less than -1 [1]. Structural SVM is algorithm is applied to learn the model parameters because ”the structured SVM allows training of a classifier for general structured output labels” (source wikipedia). So the output of the structural SVM is ordered tree in this case. As said in [1] dual coordinate descent method come up with a good approximation to the problem and gives a near optimal solution.

In practice human pose datbase contains the labeled joint location in the image. So to label types of part is also challenging. Relative location can be used as a signiicant cue to label the types. ”For example, sideways-oriented hands occur next to elbows, while downward-facing hands occur below elbows” [1].

V. Results and Comments

Apart from testing this model on standard datasets [15], [16], [17] we have also tested this model to a number of self developed photos which was taken during various sports event at IIT Kanpur. The results are shown in some of the figure below. The detection in BuFFy dataset is more time consuming and also more tough to achieve. The reson for that is that it is a video sequence taken from some episodes of the serial. So more complex interaction between human-human, human-objects and also the color of back ground and forground is similar upto some extent. Tough poses are still not detectable. When this model is applied to self shooted photos, the results are quite good for a single person detection in the images of football matches because these matches were played in the night so better contrast between the human body and the back ground. But results are not very good for the pictures taken during cricket matches even after changing the threshold value because players have white dress which matces with the color of pitch(ground) and also the background color is also not very seperable. Results for multiple human detection are not very good, as many false detection occurs during multiple people detection. I once got a result in which this model is detecting the background painting as an human and thats why I checked this model on some paintings and statues. Results obtained on the paintings are not shown here because those are not worth showing but results obtained on statutes are shown in some of the figure below.

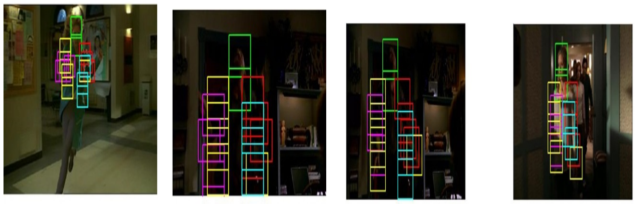

Fig. 1: Single human skeleton detection on few images from BUFFY datasets

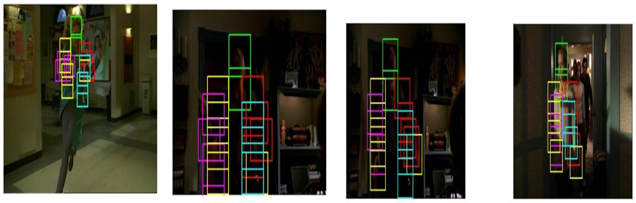

Fig. 2: Single human skeleton detection on few images from BUFFY datasets

VI. Conclusion And Our Views

This model approximate the deformation of articulated body to certain extent and works quite well for single human detection on clear surrounding. Here from clear surrounding I mean that the background color can easily be seperated out and no complex interaction is occuring between human and objects. But results are not very good for the case of multiple detections. There are so many false detection due to various reasons. Apart from this we personally think that there needs to come up with an efficient feture which can distinguish between the actual human and statues , paintings etc. This is not good that we are trying to detect humans and we end up with detecting some statues. Also rather than having a single threshold at the root we think that having multiple threshold at differeen levels of structure can provide us some good results but we are not sure about it and would like to work on it.

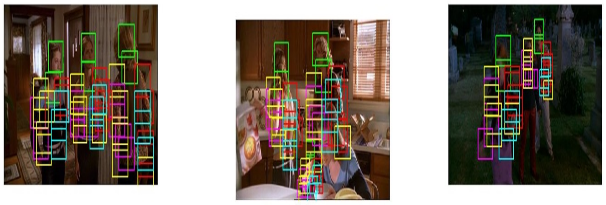

Fig. 3: Multiple detection on few images from BUFFY datasets

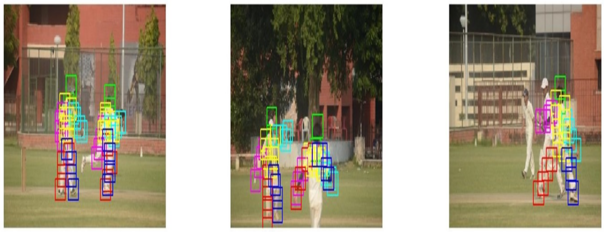

Fig. 4: Human detection on few images during a cricket match

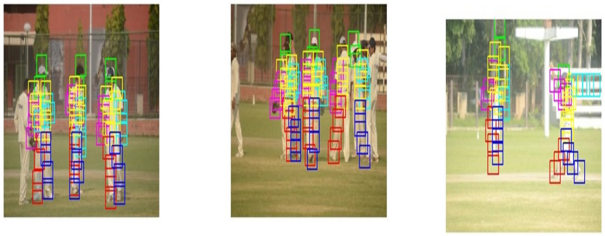

Fig. 5: Human detection on few images during a cricket match

Acknowledgment

We would like to thank Prof. Amitabha Mukerjee for their valuable comments and help. We would lso like to thank Yi Yang and Deva ramanan for providing the code for this paper. Without the code [18] it was almost impossible for us to implement and analyse the task of human pose estimation described in this paper as a course project(in a short duration of time)

References

[1] Yang, Yi, and Deva Ramanan. ”Articulated Human Detection with Flexible Mixtures-of-Parts.” (2012): 1-1.England: Addison-Wesley, 1999.

[2] M. Fischler and R. Elschlager, The representation and matching of pictorial structures, IEEE Transactions on Computers, vol. 100, no. 1, pp. 67 92, 1973.

[3] P. Felzenszwalb and D. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision, vol. 61, no. 1, pp. 55 79, 2005

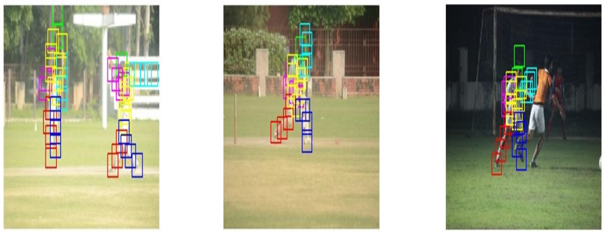

Fig. 6: Human detection during various matches

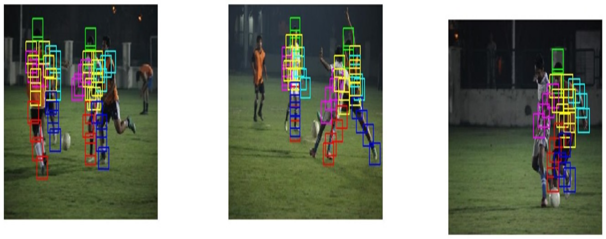

Fig. 7: Human detection on few images during a football match

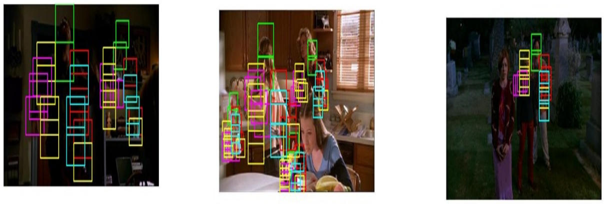

Fig. 10: False detection

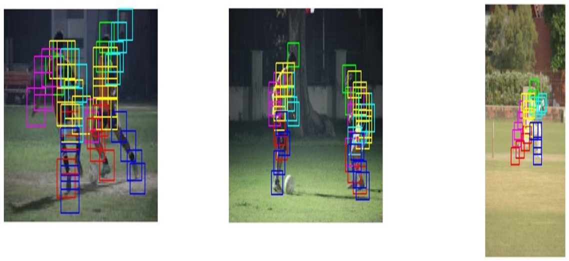

Fig. 8: Human detection during various matches

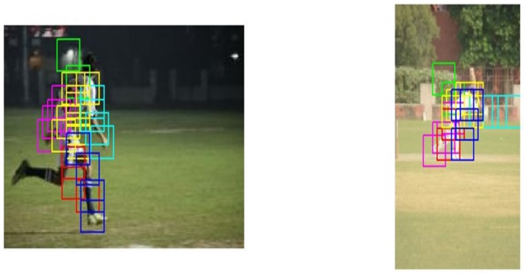

Fig. 11: False detection

[4] Ronfard, Rmi, Cordelia Schmid, and Bill Triggs. ”Learning to parse pictures of people.” Computer VisionECCV 2002 (2006): 700-714.

[5] K. Rohr, Towards model-based recognition of human move- ments in image sequences, CVGIP-Image Understanding, vol. 59, no. 1, pp. 94115, 1994.

[6] D. Ramanan, Part-based models for finding people and esti- mating their pose, pp. 199-223, 2011.

[7] S. Ioffe and D. Forsyth, Human tracking with mixtures of trees, in IEEE International Conference on Computer Vision, 2001.

[8] M. Lee and I. Cohen, Proposal maps driven mcmc for esti- mating human body pose in static images, in IEEE Conference on Computer Vision and Pattern Recognition, 2004.

[9] S. Ioffe and D. Forsyth, Probabilistic methods for finding people, International Journal of Computer Vision, vol. 43, no. 1, pp. 4568, 2001.

[10] D. Ramanan and C. Sminchisescu, Training deformable mod- els for localization, in IEEE Conference on Computer Vision and Pattern Recognition, 2006.

[11] M. Kumar, A. Zisserman, and P. Torr, Efficient discriminative learning of parts-based models, in IEEE International Confer- ence on Computer Vision, 2009.

[12] N. Dalal and B. Triggs, Histograms of oriented gradients for human detection, in IEEE Conference on Computer Vision and Pattern Recognition, 2005.

[13] D. Park and D. Ramanan, N-best maximal decoders for part models, in IEEE International Conference on Computer Vision, 2011

[14] P. Felzenszwalb, R. Girshick, D. McAllester, and R. D., Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 16271645, 2010.

[15] V. Ferrari, M. Marin-Jimenez, and A. Zisserman, Progressive search space reduction for human pose estimation, in IEEE Conference on Computer Vision and Pattern Recognition, 2008.

[16] M. Eichner, M. Marin-Jimenez, A. Zisserman, and V. Fer- rari, 2d articulated human pose estimation and retrieval in (almost) unconstrained still images, International Journal of Computer Vision, vol. 99, no. 2, pp. 190214, 2012.

[17] Sam Johnson and Mark Everingham ”Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation” In Proceedings of the 21st British Machine Vision Conference (BMVC2010)

[18] http://phoenix.ics.uci.edu/software/pose/

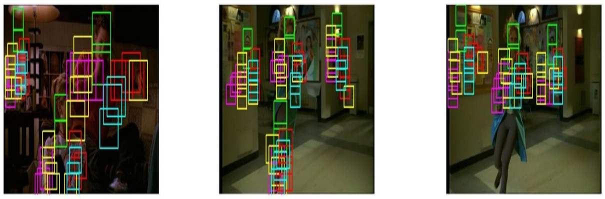

Fig. 9: False detection

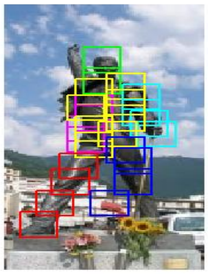

Fig. 12: Skeleton detection in statue

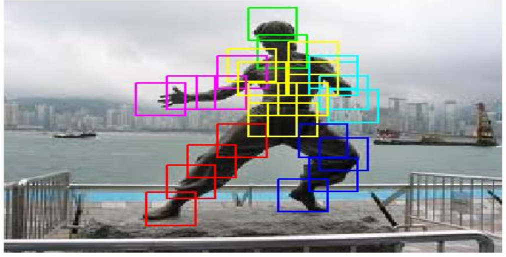

Fig. 13: Skeleton detection in statue

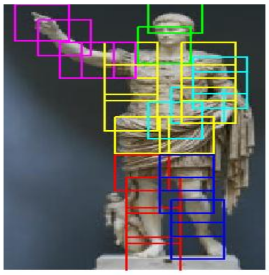

Fig. 14: Skeleton detection in statue

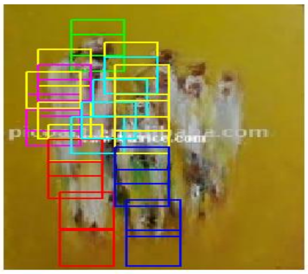

Fig. 15: Skeleton detection in painting