Human Pose Estimation Using Structural Support Vector Machines

Ke Chen, Shaogang Gong and Tao Xiang
School of Electronic Engineering and Computer Science
Queen Mary University of London
London E1 4NS, United Kingdom
{cory,sgg,txiang}@eecs.qmul.ac.uk

Abstract

This paper addresses the issue of 2D human upper-body pose estimation under cluttered environments using a discriminative structured framework. Most previous approaches focus on solving such a problem using generative models. However, a generative model has two drawbacks: a) not suitable for real-time application due to its slow inference algorithm and b) prone to over fitting given limited training data. In this work, we propose to use structured discriminative regression models for 2D human upper-body pose estimation in a model-free manner to overcome the aforementioned drawbacks. In contrast to a standard discriminative regression model, a structured regression model for human pose estimation can not only learn the relevance between image features and the presentation of human pose but also catch the inner relationship between each output. Our experimental results demonstrate the benefits brought by using structured discriminative models to articulated human pose estimation problem on cluttered images from the benchmarking Buffy the Vampire Slayer dataset and the highly challenging images from PASCAL VOC 2007 and 2008 Challenge datasets.

1. Introduction

The problem of estimating the configuration of a person's body parts have attracted more and more attentions from computer vision researchers. Human body pose esitmation has been widely used in many applications such as video surveillance [7], human-computer interface [13] and computer games [12]. However, despite the best efforts in the past decades, the human pose estimation problem estimation, especially under cluttered and uncontrolled environments, remains unsolved due to the ambiguity caused by self-occlusion, body configuration and low contrast between foreground and background.

Most existed works on human pose estimation focus on model-based methods, which specify a rough approximation of the skeleton and then use such a model in conjunction with image measurements to estimate the best-fitting pose. Those model-based techniques are characterized by a kinematic model that relates constraints between body parts including kinematic constraints of articulated human as well as other constraints such as appearance constraints [3] [14]. Pictorial Structure Model was proposed by Felzenswalb *et al* [5], which uses a prior model to measure the likelihood of the location of each limb by using appearance terms. In [14], an image parsing method based on Pictorial Structure Model employs a priori human model representing the subject and updating the model continuously with edge and colour information of still images. Recently, a method based on progressively reducing searching space by employing image parsing has been proposed which achieves superior results [6]. Based on such an image-specific color model, Eichner and Ferrari [3] use an enhanced pictorial method containing an appearance model describing hidden relationship between each body parts according to location priori within the foreground. Johnson and Everingham [10] try to add coherent appearance properties of each body parts to a Pictorial Structure Model in order to improve the results. In [11], clustering is performed to discover pose groupings in a pose space. This model is still based on the pictorial structure, thus is still a generative method. Despite its popularity, it is noted that using a generative model for estimating human pose may have some drawbacks, including a) not suitable for real-time application due to its slow inference algorithm and b) prone to over fitting given limited training data.

To overcome the drawbacks of generative methods, discriminative regression methods can be considered for human pose estimation which once trained can run very fast during testing. However, general discriminative regression methods such as Support Vector Regression (SVR) could only estimated the output pose parameters individually instead of in a global and structured manner. In other words, those non-structured regression methods ignore the important information about the relevance between each body

parts in our case. In this paper, two structured discriminative methods, i.e., Structural Support Vector Regression (SSVR) [1] [8] [9] and Latent Structural Support Vector Regression (LSSVR) [17] are adopted for 2D human pose estimation under cluttered environments. Compared to Support Vector Regression, both of aforementioned structural methods are designed to capture the dependency on structured input and structured output. Extensive experiments using public benchmarking datasets have been carried out to demonstrate that: i) During testing, our methods could run much faster than generative methods; they are thus more suitable for real-time application. ii) Our methods generate acceptable results when the size of training database is reduced dramatically. iii) Compared to non-structured discriminative methods (e.g., Support Vector Regression), structured methods achieve better performance owing to the ability to capture the important relevance information between outputs.

2. Methodology

In this section, we will present problem formulation and our methods in details. For 2D human upper-body pose estimation in still images, we wish to find out the configuration of six human upper body parts (head, torso, and upper/lower right/left arms). For unconstrained still images, we know nothing about the person's appearance (e.g. what cloth she/he wears) and it is expensive to search the whole images. For effectively estimating human pose, one preprocessing step will be taken before model learning to reduce the possible space and improve the efficiency of our approach. In particular, we will use a pre-learned upper-body detector [3] [6] to localize the human body. We detect the upper body in each frame using a sliding window approach with a Histograms of Oriented Gradients representation of human appearance [2]. By using the upper-body detector, the searching space will be reduced significantly. After the localization of the upper body, learned structured discriminative regression models are then used to estimate the body pose.

2.1. Model Input and Output

In unconstrained still images, low-contrast and diverse appearance could increase the difficulty of estimating human pose. It is therefore vital to extract informative appearance features as model input. In our model-free framework, for no kinematic model is used to constrain the estimation procedure, the features extracted from the detected upper body bounding should capture information that is useful for identifying different body parts and sensitive to body pose changes. To this end, Bag-of-word SIFT features are used as the input while the output for our regression models are structured coordinates. More specifically, for the i-th body part, the output are coordinate $[x_{1i}; y_{1i}; x_{2i}; y_{2i}]$. For using a pre-learned upper-body detector, multiple bounding boxes may exist in a single image. Note that for training, we will use ground truth location of upper bodies to extract those features as model inputs. During testing, the body location is provided by the upper body detector. After localizing the upper body, we will extract the bag-of-words SIFT within the bounding boxes for model inputs. Randomly chosen descriptors are employed by K-means to generate a codebook with 400 clusters. In order to incorporate location information of each body parts into the model inputs, each bounding box is divided into $2 \times 2 = 4$ sub-regions. A 400 dimensional feature vector is then computed from each subregion and the four feature vectors are concatenated into a 1600 dimensional feature vector as the final model input. The model output has a dimensionality of 24 (4 coordinates $\times$ 6 body parts).

2.2. Structural Support Vector Regression

The problem formulation for 2D upper-body pose estimation is as follows. For supervised learning, we have pairs of input and output x_i, y_i, where $i = 1, 2, \cdots N$ and N denotes the size of training set. x_i and y_i are feature vectors of 1600 and 24 dimensions respectively as described above. During training, each body parts are mannually annotated and the value of y_i is computed from the annotation. The objective of model learning is learn a discriminative regression function as a linear combination of joint features [1] [8] [9] [15]:

$$\text{argmin}_y f_w(x, y) = w^T \Psi(x, y), \tag{1}$$

where w is a parameter vector and $\Psi(x, y)$ is a feature vector induced by a joint kernel $K(x, y, x', y') = \Psi(x, y)^T \Psi(x', y')$. The above structured Support Vector Regression problem is thus solved by estimating the parameter vector w. This can be formulated as the following optimisation problem:

$$\begin{aligned} \min \quad & \frac{1}{2} w^T w + C\xi \\ s.t. \quad & \forall(\overline{y}_1, \cdots, \overline{y}_N) \in Y^N \\ & \frac{1}{N} w^T \sum_{j=1}^{N} [\Psi(x_j, y_j) - \Psi(x_j, \overline{y}_j)] \\ & \geq \frac{1}{N} \sum_{j=1}^{N} \Delta(y_j, \overline{y}_j) - \xi, \end{aligned} \tag{2}$$

where the loss function $\Delta(y_j, \overline{y}_j)$ and $\Psi(x, y)$ are problem-dependent. It is worth mentioning here that the above equation is an 1-slack formulation, which is more efficient than the original n-slack one. In our case, we will consider a square distance as the loss function for pose estimation. According to the dual theory, we could easily get the dual prob-

lem for the above equation, which is used for constraint generation as well as the stability analysis [9]. It is worth pointing out that the above formulation for our problem could be margin-rescaling and there also have a slack-rescaling formulation, e.g., OP3 in [9], which is not effective for our problem. For human pose estimation, the kernel function which could induce the feature vector in Equation (2) is similar to joint RBF kernel [16]. That is,

$$K((x,y),(x',y')) = exp(-\|(x,y)-(x',y')\|^2).$$

For the output loss function we use the square difference of image feature vector. More specifically, the formulation is as follows

$$\begin{aligned}\Delta(y,y') &= \|\varphi(y)-\varphi(y')\|^2 \\ &= K(y,y)+K(y',y')-2K(y,y') \\ &= 2(1-K(y,y')).\end{aligned}$$

2.3. Latent Structural Support Vector Regression

In this subsection, Latent Structural Support Vector Regression is investigated and formulated. Latent Structural Support Vector Machine, was proposed by [9] to solve the classification problem. Here the model is extended so that it can be used for regression, i.e. the model outputs become continuous rather than discrete. The difference between Latent Structural Support Vector Regression and Structural Support Vector Regression is the introduction of latent variables in the model. With latent variables the model aims to capture not only the input-output relationship but also unobserved relationships, such as relationship between different body parts [4].

The detailed formulation using Latent Structural Support Vector Machine for our problem will be presented as the following. In comparison with the aforementioned Structural Support Vector Regression model, latent variable vector will be added to the joint feature vector $\Psi(x,y,h)$:

$$\text{argmin}_{(y,h)} f_w(x,y,h) = w^T\Psi(x,y,h). \tag{3}$$

Similar to the Structural Support Vector Regression presented in the last subsection, a joint kernel $K(x,y,h,x',y',h') = \Psi(x,y,h)^T\Psi(x',y',h')$ could be induced by $\Psi(x,y,h)$. As a result, the optimisation problem of a Latent Structured Support Vector Regression model is written as:

$$\begin{aligned}\min\ & \frac{1}{2}w^Tw + C\sum_{j=1}^{N}\xi_j \\ s.t.\ & \forall(\overline{y}_1,\cdots,\overline{y}_N)\in Y^N \\ & w^T[\Psi(x_j,y_j,h)-\Psi(x_j,\overline{y}_j,\overline{h})] \geq \Delta(y_j,\overline{y}_j,\overline{h})-\xi, \\ & \text{for}\quad j=1,2,\cdots N;\end{aligned} \tag{4}$$

where $x_j, y_j, j = 1,2,\cdots N$ is the training pairs and $\overline{y}_j, j = 1,2,\cdots N$ denotes prediction results approaching y_j during training procedure. Note that, for simplifying the formulation and increasing the efficiency, the loss function will not depend on $h_i^* = \text{argmin}_h w^T\Psi(x_i,y_i,h)$ but on the predicted latent variable $\overline{h}$ for practical application [17]. Evidently, Equation (4) could be reduced into the Structural Support Vector Regression formulation by removing the latent variables. In this work, the kernel function in the above formulation the same joint RBF function as Structural Support Vector Regression. In other words, the loss function only depends on input and structured output without latent variables. The optimization problem is solved using the Concave-Convex Procedure (CCCP) [17] [18], which is guaranteed to converge to a local minimum.

Before ending this section, we have one remark about three discriminative methods. Compared to Support Vector Regression, Structural Support Vector Regression and Latent Structural Support Vector Regression have the potential to solve more complicated regression problem owing to their ability to model structured outputs. However, the price to pay is the increased model complexity, which may imply higher computational cost. As a result, the tradeoff between complexity and accuracy needs to be determined according the application at hand and the amount of training data available. In particular, the latent variables adding into Structural Support Vector Regression means that more training data are required to learn the model in comparison with SSVR. In other words, when the training data size is small, LSSVR is more likely to suffer from model overfitting resulting in worse performance.

3. Experimental Results

Datasets and Settings – Experiments were carried out to demonstrate the effectiveness and efficiency of our models for human pose estimation. We used the same databases as in Ferrari *et al.*'s paper [3] [6] including cluttered images from the TV episodes *Buffy the Vampire Slayer* and highly challenging images from PASCAL VOC 2007 and 2008 datasets.

Three experiments were conducted, each of which differs in how the training/testing dataset were organised. In the first experiment, different sizes of randomly selected images from Buffy Episodes 3&4 and VOC 2007&2008 were employed as training sets, while the test sets were the same including 276 images from Buffy Episodes 2&5&6. In the second experiment, the training and test sets were replaced by Buffy Episodes 2&3&4&5&6 and Pascal 2007 containing 91 testing images respectively. In the third experiment, a more balanced training set was used which has the same number of different categories of poses selected from Buffy Episodes 3&4 and VOC 2007&2008, while the test set was the same as that used in the first experiment.

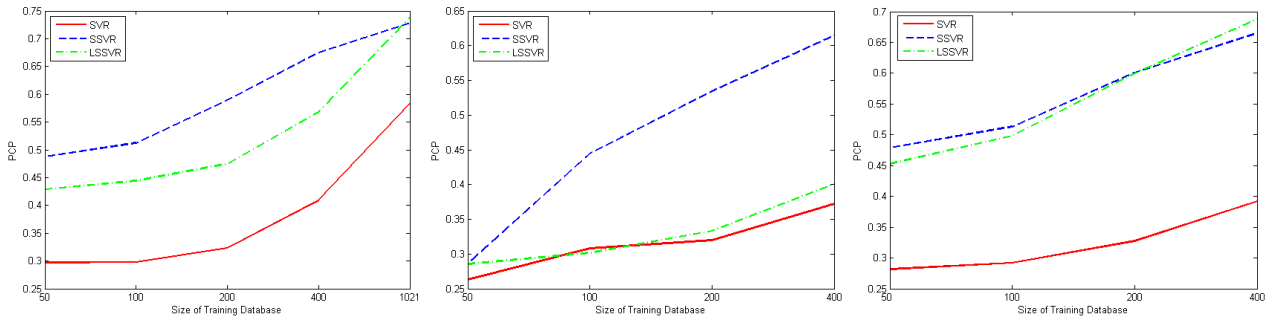

Figure 1. PCP for different testing database, where the right, middle and left plots show the results of three experiments respectively

The inputs and outputs for the three models we compared (i.e., SVR, SSVR, and LSSVR) are the feature vectors and corresponding human body configuration. In order to increase the accuracy of estimation and speed up the computation, one pre-processing step was employed, that is, the upper-body detector [6], which will search the entire image to find the rough position and scale of people. It is evident that our results rely on the good performance of upper-body detector as pose estimation will only be performed in the detected regions. The detection rate of the bounding boxes detected was relatively high on the datasets we used. Specifically we achieved a detection accuracy of 0.8043 for the Buffy database in our testing sets. When there exits multiple detections in one image, multiple poses will estimated but only one pose will be selected for comparing with ground truth. This is because as the ground truth for each image of the Ferrari *et al.*'s database [3] [6] provides only one pose. Within each of the bounding boxes, 5000 bag-of-word SIFT features was extracted and we then used k-means to create a codebook consisting of 400 clusters. In all of our experiments, we use PCA [4] to reduce the dimension of the input image feature vector from the original 1600 (4 sub-regions with 400 histogram bins in each sub-region) to 20 dimension in order to increase the efficiency. For SVR training, 24 Support Vector Regression will be trained independently, while for SSVR and LSSVR, the 24 outputs were estimated jointly. During LSSVR training, latent variables were manually labeled according to different categorizes of poses (5 in our experiments). To evaluate the performance of different models, Percentage of Correctly estimated body Parts (PCP) will be used [3] [6], i.e., an estimated body part is deemed as correct if its segment endpoints lie within 50% of the length of the ground-truth segment from their annotated locations.

Computational Efficiency of the Proposed Models – The Experimental results are shown in Tables 1, 2 and 3 as well as Figures 1, 2 and 3. Compared to the best results generated by generative models [3] (i.e., 78.1% obtained by using a training set of 1021 images and the same testing set as in our first experiment), our results are comparable. However, our discriminative models are much more efficient to compute. Specifically, using our discriminative methods, it took 5.86 seconds per image on average. On the other hand, using the generative method in [3] it took more than 70 seconds for testing one image. In other words, the testing time using the discriminative methods is more than 10 times faster in comparison with generative methods in [3,6]. Moreover, it is found that, by using less training images as Ferrari did [3], the structured methods (especially SSVR in all three experiments and LSSVR in the third experiments) which we propose for human pose estimation has a PCP not too much lower than that of the generative method.

Effect of Modelling Structured Output – Our results show that for all three experiments, the two structured regression models, particularly SSVR significantly outperform the regression model without modelling the structure of model outputs (SVR). This results show the importance of modelling output variable structure for the problem of human pose estimation. This is because the position of different body parts are typically highly correlated. Ignore the structure of them thus means that important information has been left unexplored.

Effect of Training Data Size – Figure 1 shows that, with an increasing training set size, the performance of all three regression machines improves. Moreover, in all three experiments, SSVR shows the best generalisation capability for human pose estimation among the three discriminative methods. Figure 1 also shows the disparity in performance of the LSSVR over the three datasets. It is worth pointing out that LSSVR achieves significant worse result in the second experiments. This is because in the Buffy datasets, most of the latent variables are the same (i.e. most of the poses in Buffy datasets are similar). In other words, the poor performance of LSSVR in the second experiment was due to the fact that the latent variables in the LSSVR model becomes redundant thus having a negative effect.

Effect of a Balanced Training Set – The third experiment was designed to demonstrate the importance of preparing a balanced training dataset when a LSSVR is employed. Form Table 3 as well as the right plot of Figure 1, we can see that LSSVR outperform SSVR and SVR when the size of training database is 400. In comparison, in our first ex-

Table 1. PCP with different size of randomly selected training datasets for the Buffy testing set (i.e., 276 images of Buffy Episodes 2, 5 and 6), where SoD denotes the size of training database.

SoD	50	100	200	400	1021
SVR	29.71%	29.79%	32.35%	40.93%	58.33%
SSVR	48.72%	51.21%	58.97%	67.49%	72.82%
LSSVR	42.88%	44.39%	47.47%	56.79%	73.79%

Table 2. PCP with different size of randomly selected training sets for VOC 2007 (including 91 testing images), where SoD denotes the size of training database.

SoD	50	100	200	400
SVR	26.32%	30.82%	31.98%	37.24%
SSVR	28.57%	44.42%	53.41%	61.44%
LSSVR	28.57%	30.15%	33.32%	40.06%

Table 3. PCP with different size of balanced training dataset (that is, we select equal number of images for each of five pose categories), where the testing database is the same as Table 1.

SoD	50	100	200	400
SVR	28.14%	29.22%	32.77%	39.21%
SSVR	47.91%	51.32%	60.11%	66.51%
LSSVR	45.27%	49.85%	60.03%	68.74%

periment, LSSVR could only achieve superior performance to SSVR when the size of training database is 1021. This result indicates that when the training dataset is large enough and has the balanced number of different poses for each pose category, the performance generated by LSSVR can be superior to that of SSVR and SVR.

Discussions – Firstly, it is evident from our results that discriminative methods can process test images much more efficiently. They are thus more suitable for online/real-time application, even when training database is small. Secondly, when using less training images, our methods could also achieve good results. The curves shown in Figure 1 verify that the performance of our structured discriminative methods degrade gracefully when the training dataset size decreases. Thirdly, compared to a standard discriminative method SVR, structured techniques lead to superior results, which demonstrate the importance of introducing structured learning method to 2D human pose estimation. Finally, we could benefit form adding hidden variables into Structural Support Vector Regression when the training dataset is large enough and balanced.

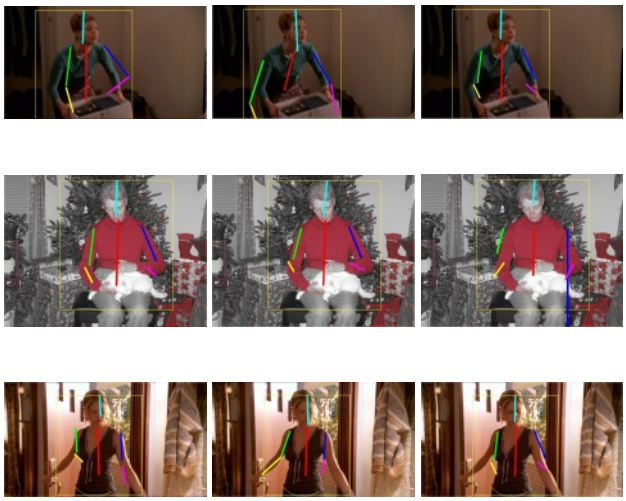

Figure 2. Illustrative results for testing Buffy and Pascal generated by LSSVR (Left), SSVR (Middle) and SVR (Right)

4. Conclusions

This paper has investigated three model-free discriminative methods for 2D human upper-body pose estimation. As seen from the results presented in the last section, our method could solve the problem effectively and more efficiently than previous works. Compared to generative model-based method, our techniques could not only achieve good performance but also high-efficiency owing to the nature of discriminative methods. Additionally, more benefits could be achieved by capturing the correlations between output variables using structured discriminative methods. We also discover that compared to Latent Structural Support Vector Regression, Structural Support Vector Regression perform well given less training data and when the training data is unbalanced. Otherwise, LSSVR is preferred.

References

[1] M. B. Blaschko and C. H. Lampert. Learning to localize objects with structured output regression. *Proceedings of European Conference on Computer Vision*, pages 2–15, 2008. 2

[2] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *Proceedings of IEEE Computer Vision and Pattern Recognition*, 1:886–893, 2005. 2

[3] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. *Proceedings of British Machine Vision Conference*, pages 1–11, 2009. 1, 2, 3, 4

[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99:1–20, 2009. 3, 4

Figure 3. Illustrative results by our model-free discriminative methods with single detection for single person (Left), multiple detection for single person (Middle) and multiple detection for multiple detection for multiple person (Right). The top image of the right column shows that our method can estimate multiple poses for one image at the same time, while the middle and bottom images of the right column illustrate that multiple wrongly-estimated poses caused by over-enlarging bounding boxes for human upper body and localization.

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

[6] V. Ferrari, M. Marin-Jiminez, and A. Zisserman. Progressive search space reduction for human pose estimation. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008. 1, 2, 3, 4

[7] W. Hu, T. Tan, L. Wang, and S. Maybank. A suvrvey on visual sureillance of object motion and behaviors. *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Application and Review*, 34(3):334–352, 2004. 1

[8] C. Ionescu, L. Bo, and C. Sminchisescu. Structural svm for visual localization and continuous state estimation. *Proceedings of IEEE International Conference on Computer Vision*, 2009. 2

[9] T. Joachims, T. Finley, and C.-N. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009. 2, 3

[10] S. Johnson and M. Everingham. Combining discriminative appearance and segmentation cues for articulated human pose estimation. *Proceedings of IEEE International Conference on Computer Vision*, 2009. 1

[11] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. *Proceedings of British Machine Vision Conference*, 2010. 1

[12] T. B. Moeslund, A. Hilton, and V. Kruger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104:90–126, 2006. 1

[13] R. Nevatia and C.-W. Chu. Body pose estimation and gesture recognition for human-computer interaction system. *Doctoral Dissertation*, 2008. 1

[14] D. Ramanan. Learning to parse images of articulated objects. 2006. Neural Info. Proc. Systems. 1

[15] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. *Proceedings of the International Conference on Machine Learning*, 2004. 2

[16] J. Weston, B. Schoelkopf, O. Bousquet, T. Mann, and W. S. Noble. Joint kernel maps. *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2004. 3

[17] C.-N. Yu and T. Joachims. Learning structural svms with latent variables. *Proceedings of International Conference on Machine Learning*, 2009. 2, 3

[18] A. Yuille and A. Rangarajan. The concave-convex procedure. *Neural Computation*, 15:915, 2003. 3