Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation

Sam Johnson
s.a.johnson04@leeds.ac.uk
Mark Everingham
m.everingham@leeds.ac.uk

School of Computing
University of Leeds
Leeds, UK

Human pose estimation is the task of estimating the 'pose' or configuration of a person's body parts *e.g.* labeling the position and orientation of the head, torso, arms and legs in an image. In this paper we propose an extension of the pictorial structure model (PSM) approach [2]. Our method incorporates richer models of appearance and prior over pose without introducing unacceptable computational expense. We build on the idea of a 'mixture of trees' model [3]. The space of human poses is partitioned into a set of clusters such that the prior over plausible poses can be modeled with greater fidelity. Within each pose cluster we use pose-specific appearance terms which implicitly capture the dependence of a part's appearance on pose and the correlation between the appearance of parts.

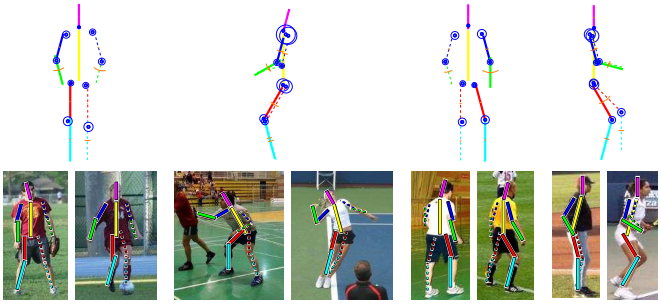
To cope with the large variation in part appearance due to factors such as clothing or varying anatomy we use state-of-the-art *nonlinear* SVM classifiers to model the appearance terms. This would typically be prohibitive in terms of computational expense, however we show that by adopting a cascaded reduced set machine formulation [6] we can exploit such strong classifiers efficiently. Current methods have been limited by the lack of available training data – to overcome this we introduce a new annotated dataset of 2,000 diverse and challenging consumer images which will be made publicly available (Figure 1). Our results show that the use of stronger appearance terms and prior model in the proposed approach results in a greater than 50% relative improvement in pose estimation accuracy on this dataset compared to a state-of-the-art method [4].

Mixture of PSMs prior model. In most previous methods a single tree-structured Gaussian prior model and set of part appearance models are learnt [1, 4, 5]. This leads to a broad, non-descriptive prior and an appearance model which cannot capture the 'multi-modal' appearance of body parts *e.g.* the different appearance of the head in frontal, profile or rear views. We propose to overcome these issues by partitioning the pose space into clusters, to give a mixture of PSMs – one PSM for each cluster. Partitioning the pose space has two positive effects: (i) the prior over pose is modeled more faithfully since the tree-structured Gaussian assumption is more realistic within a pose cluster than over the whole space; (ii) since each cluster contains parts in a tighter range of configurations the variation in appearance is reduced. This allows us to build much more successful appearance models for each part at a cluster level than globally, where the discrimination task is much harder. Correlation between the appearance of multiple parts conditioned on the cluster can also be captured.

Classifier. Previous work has typically used simple linear models to represent part appearance [4, 5]. Even within a pose cluster this is inadequate – there is simply too much variation in the possible appearance of a part. We therefore propose to use an SVM classifier with a Radial Basis Function (RBF) kernel – this enables modeling of the multi-modal appearance distribution. In order to evaluate appearance terms within the PSM framework, the classifier for each part must be applied *exhaustively* over all image positions and orientations. Using a nonlinear SVM this is prohibitively computationally expensive since evaluating the classification function can require many thousands of high-dimensional dot product evaluations. We therefore propose to use a cascade of simple to complex classifiers [6]. The idea is that most image windows can be rejected as non-parts by simple (and fast) classifiers, such that only a few windows need to be considered by the full (slow) nonlinear SVM classifier.

Results. We report results on 1,000 images from our new dataset (Table 1) and the IIP dataset [5] (Table 2). Using a global prior (as in previous work [1, 4, 5]) we first compare (A) a linear appearance classifier and (B) our proposed nonlinear SVM. The overall accuracy improves from 36.4% to 44.7%. We next introduce partitioning of the pose space, using four clusters. A large improvement in accuracy is evident using just the linear model (C) – from 36.4% to 43.6%. When we combine the clustered pose models with the nonlinear SVM appearance model (D) the overall accuracy improves further to 55.1% – a 51% relative improvement over the baseline method. Example output on our new dataset can be seen in Figure 2. On the IIP dataset [5] our proposed method gives overall accuracy of 66.2%, compared to the best result reported to date of 56.4% [4].

Figure 1: Examples from our new dataset of 2,000 images. Images were gathered from Flickr using the tags shown, and exhibit a wide range of poses with varying degrees of difficulty in challenging natural scenes.



(a) "Front" (b) "Left" (c) "Back" (d) "Right"

Figure 2: Learnt priors over pose for four pose clusters and some examples of our output for images within each. Solid lines represent limbs on the right side of the body from a person-centric viewpoint. The variance at one standard deviation of joint offsets (blue circles) and joint angles (orange arcs) is shown. The learnt clusters can be seen to approximate views from the front, left, back and right. The bottom row shows the output of our proposed method on a number of images from our dataset.

Method	Total	Torso	Legs	Arms	Head
Linear (A)	36.4	64.1	41.8	20.5	49.9
Nonlinear (B)	44.7	70.9	52.2	27.9	55.9
4 Cluster Linear (C)	43.6	74.1	51.5	24.0	59.7
4 Cluster Nonlinear (D)	**55.1**	**78.1**	**62.3**	**40.1**	**62.9**

Table 1: New dataset results. Comparison of localization rates (in percentages) for experiments with global and clustered appearance terms and prior. See text for discussion – more detailed results can be found in the paper.

Method	Total	Torso	Legs	Arms	Head
Ramanan[5]	27.2	52.1	30.0	15.6	37.5
Andriluka *et al.* [1]	55.2	81.4	59.1	39.6	75.6
Johnson & Everingham[4]	56.4	77.6	58.2	46.2	68.8
Proposed	**66.2**	**85.4**	**69.4**	**55.7**	**76.1**

Table 2: IIP dataset [5] results. Comparison of localization rates (in percentages) for previous approaches and ours.

Conclusions. We show that combining a mixture model of pose with nonlinear appearance classifiers enables modelling the high levels of variation present in natural images of human poses. Overall we achieve over 50% relative improvement in accuracy over a state-of-the-art method on a large challenging dataset introduced here, and over 17% relative improvement over the best reported results on the IIP dataset.

[1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009.

[2] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1), 2005.

[3] S. Ioffe and D. Forsyth. Mixtures of trees for object recognition. In *CVPR*, 2001.

[4] S. Johnson and M. Everingham. Combining discriminative appearance and segmentation cues for articulated human pose estimation. In *MLVMA*, 2009.

[5] D. Ramanan. Learning to parse images of articulated bodies. In *NIPS*, 2006.

[6] S. Romdhani, P. Torr, B. Schölkopf, and A. Blake. Efficient face detection by a cascaded reduced support vector expansion. *Proceedings of the Royal Society*, 460(2501), 2004.