

# 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

## Abstract

We investigate the task of 2D articulated human pose estimation in unconstrained still images. This is extremely challenging because of variation in pose, anatomy, clothing, and imaging conditions. Current methods use simple models of body part appearance and plausible configurations due to limitations of available training data and constraints on computational expense. We show that such models severely limit accuracy. Building on the successful pictorial structure model (PSM) we propose richer models of both appearance and pose, using state-of-the-art discriminative classifiers without introducing unacceptable computational expense. We introduce a new annotated database of challenging consumer images, an order of magnitude larger than currently available datasets, and demonstrate over 50% relative improvement in pose estimation accuracy over a state-of-the-art method.

## 1 Introduction

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. It is an important goal since it is a fundamental part of high-level understanding of imagery depicting people, with diverse applications including content-based image indexing, surveillance, markerless motion capture or human computer interaction. The task is particularly challenging because of the wide variation in human appearance present in natural images due to pose, clothing and imaging conditions (Fig. 1(a) and 1(d)). This is especially true for the domain we investigate – unconstrained still images – where prior knowledge of the motion being performed (*e.g.* walking) or constraints on the imaging setup (*e.g.* background subtraction) cannot be exploited.

State-of-the-art methods typically use graphical models which represent the human body as a connected collection of parts corresponding to head, torso and other limbs [1, 2, 3, 4, 5]. The particular model on which we build because of its previous success and modest computational expense is the ‘pictorial structure model’ (PSM) [6]. The model consists of two components: (i) *appearance* terms, which model the probability of a part being present at a particular location and orientation given the input image; (ii) a *prior* which models the probability distribution over pose, constraining the estimated pose to be plausible in terms of human articulation. To enable efficient inference, two (unrealistic) assumptions are made: (i) the appearance of a part is assumed independent of its pose and that of the other parts;



Figure 1: Features of people in natural images. (a) Part appearance is highly dependent on pose even for the same person; (b) However, the appearance of multiple parts are highly correlated given similar pose; (c) Clothing and anatomy differences lead to a need for nonlinear appearance models; (d) Simple models of body joint angle distributions fail to capture the large variations present in poses.

(ii) the prior over pose is a Gaussian with ‘tree-structured’ covariance (see Section 2.1). Furthermore, the per-part appearance terms use simple models *e.g.* linear image patch classifiers [14, 16]. Fig. 1 highlights some of the limitations of such approaches: Fig. 1(a) shows how part appearance varies greatly with pose *e.g.* that of the leg as a function of foreshortening. Fig. 1(b) shows how appearance is correlated with pose *e.g.* similar poses yield similar appearance for all parts. Fig. 1(c) shows how the linear assumption is likely to fail – *e.g.* due to substantial variation in clothing patterns or anatomy. All images in Fig. 1 also show how a simple Gaussian prior is a poor model of human articulation – for example given such a model the angle between the torso and upper arms is effectively unconstrained. While the limitations are clear, the choice and implementation of the PSM is a pragmatic one due to (a) paucity of data available for learning; (b) constraints on computational expense. For example the popular IIP dataset [16] offers just 100 training images from which to learn appearance and prior, so models must be kept simple to avoid over-fitting. Since the classifiers used to model appearance terms must be applied in a sliding-window fashion exhaustively over position and orientation, fast linear classifiers are appropriate.

In this paper we propose an extension of the PSM approach which incorporates richer models of appearance and prior over pose without introducing unacceptable computational expense. Our method builds on the idea of a ‘mixture of trees’ model [14, 23]. We partition the space of human pose into mixture components or ‘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* support vector machine (SVM) classifiers to model the appearance terms. This would typically be prohibitive in terms of computational expense since such classifiers are thousands of times slower than simple linear classifiers. However, we show that by adopting a cascaded reduced set machine formulation [19] we can exploit such strong classifiers at negligible additional computational expense. As noted, 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. 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 [14].

**Related work.** In terms of part appearance, most work on articulated pose estimation and tracking has used simple models, often hand-crafted, for example rectangular box filters or edge templates [6, 10, 18, 20]. The original PSM model used color box filters [2] which improves specificity, but requires *a priori* knowledge of the clothing color. Skin color has also been exploited [14] but works poorly in unknown lighting and assumes short sleeves/trousers are worn. Recent work has used machine learning methods to learn part appearance from annotated training images. Ramanan learns a weighted (linear) edge template for each part in a conditional random field (CRF) framework [16], and combines this with a color model bootstrapped from the test image. Ferrari *et al.* [5, 9] also use an image-specific color model, but require an initial localization of the person from an upper-body detector, which is not applicable to unconstrained pose. Several authors have proposed to use state-of-the-art image descriptors for body part modeling. Buehler *et al.* [3] use the Histogram of Oriented Gradient (HOG) descriptor [9] which gives controlled invariance to lighting and spatial deformation; a simple ‘nearest neighbor’ scheme is used to detect the parts. Andriluka *et al.* [1] use a similar ‘shape context’ descriptor [15], learning discriminative part detectors using boosting. Kumar *et al.* use HOG and skin color cues and a linear classifier [13], and in our previous work [12] we combine HOG with local segmentation cues again using a linear classifier, reporting the best results on the IIP dataset [1, 12, 16] to our knowledge.

The most successful recent methods [1, 3, 5, 9, 12, 16, 18] have built on the PSM framework [2]. We discuss this approach in more detail in Section 2.1. Briefly, the approach enables efficient globally optimal inference and efficient sampling by constraining the form of prior over pose to a ‘tree-structured’ Gaussian, and requiring independent appearance between parts. However, the constrained prior limits the fidelity with which plausible human articulation can be modeled and prevents modeling of phenomena such as self-occlusion of parts. Work related to ours has attempted to address these issues by using a mixture of trees [11] prior or a latent ‘pose’ variable [24]. Wang and Mori [23] use multiple tree models to better constrain allowable poses, and to represent self-occlusion. However, the accuracy of their results is limited by the weak appearance terms used. We propose to use the approach of multiple trees to better model the wide range of poses present in natural images (see Fig. 1(d) for examples), but additionally incorporate strong, pose-specific appearance terms. In the related domain of deformable object detection a similar approach has proven very successful [8], though again using much simpler appearance models than proposed here.

**Outline.** Section 2 describes our proposed method, covering (i) the PSM framework on which we build; (ii) pose clustering to improve the prior model and support pose-specific appearance terms; (iii) discriminative part detectors used to model body part appearance. Section 3 introduces our new dataset and reports experiments, and we offer conclusions in Section 4.

## 2 Method

This section describes our proposed method. We start with a brief overview of the PSM approach on which we build, and then describe our proposed contributions in terms of richer models of prior and appearance.

### 2.1 Pictorial Structure Model

The PSM [2] provides a probabilistic framework for modeling appearance and pose of the human body and supports computationally-efficient inference. The body is modeled as a collection of ten major parts or ‘limbs’ – the head, torso, and upper- and lower-limbs. A tree-structured graph over the parts is defined with nodes representing the individual

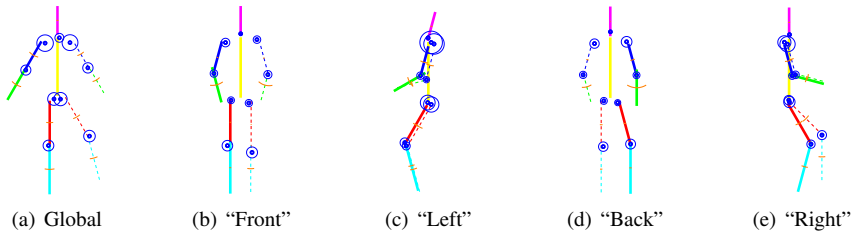


Figure 2: Learnt priors over pose for a conventional global model (one cluster) and four learnt pose clusters. Solid lines represent limbs on the right side of the body from a person-centric viewpoint. The variance at one standard deviation of joint offsets (blue circles) and joint angles (orange arcs) is shown. The learnt clusters can be seen to approximate views from the front, left, back and right.

parts and edges representing conditional probability distributions on the relative position and orientation of parts. A 2-D configuration – or pose – of the model is parametrized by  $L = \{\mathbf{l}_1, \mathbf{l}_2, \dots, \mathbf{l}_n\}$  where each  $\mathbf{l}_i = \langle x_i, y_i, \theta_i \rangle$  specifies the 2D position and orientation of a part. The posterior probability of a particular pose given an image  $I$  can be written as

$$p(L|I) \propto p(I|L)p(L) = \underbrace{\prod_i p(\mathbf{f}_i|\mathbf{l}_i)}_{\text{appearance}} \underbrace{\prod_{(\mathbf{l}_i, \mathbf{l}_j) \in E} p(\mathbf{l}_i|\mathbf{l}_j)}_{\text{prior}} \quad (1)$$

where the appearance term represents the likelihood that a part matches a descriptor of the corresponding image region  $\mathbf{f}_i$ , and the prior term represents the prior probability of a configuration.

The purpose of the *prior* term is twofold: (i) it constrains the model to configurations which are plausible given the kinematic limits of the body, and (ii) it encourages poses which are more common in the case of ambiguous image information. As shown in Eqn. 1 the tree-structured model means that the prior  $p(L)$  factorizes into a product of pairwise conditional distributions on each edge  $(\mathbf{l}_i, \mathbf{l}_j) \in E$ . This is crucial in enabling efficient inference using a dynamic programming approach [2]. Each pairwise conditional probability  $p(\mathbf{l}_i|\mathbf{l}_j)$  models the compatibility of the relative location and orientation of the parts. We model the distribution of relative location and orientation as Gaussians in a transformed space [2, 17]. This choice is necessary to enable efficient inference using the distance transform [2]. Allowing the parts to drift apart in a controlled ‘loose-limbed’ manner [17] improves robustness. The *appearance* term  $p(I|L)$  captures the agreement between a hypothesized pose and the input image. The PSM assumes independence between all parts such that the appearance term can be evaluated tractably for each part in isolation. As described in Section 2.3 we model the appearance terms *discriminatively* using a classifier for each part.

## 2.2 Mixture of PSMs prior model

In most previous methods a single tree-structured Gaussian prior model and set of part appearance models are learnt [2, 17, 16]. 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 (Fig. 1(b)). 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. Clustering also implicitly captures the *correlation* between the appearance of multiple parts, conditioned on the cluster.

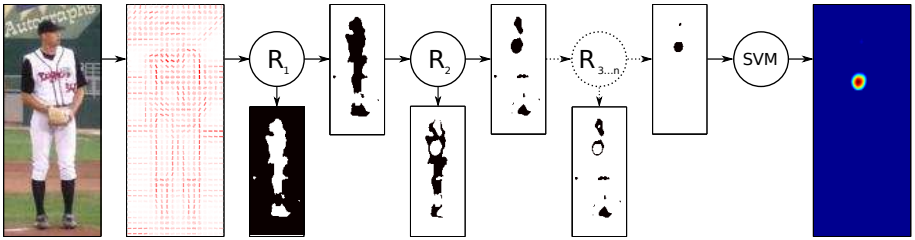


Figure 3: Part (torso) detection using a cascaded reduced set machine (RSM). Given an image we extract HOG features and pass them to the cascade. At each cascade stage a more powerful RSM classifies each possible part location. Locations which are considered to not contain parts are rejected (shown in black underneath each classifier). Locations which are classified as potential part locations are passed to the next – more powerful – stage. Finally the full SVM is applied to each remaining candidate location to give the ‘probability’ that the part is at any given location (red is high probability).

**Pose clustering.** Partitioning the pose space is formulated as a maximum-likelihood clustering problem – we seek a set of  $K$  pose clusters such that each pose in the annotated ground truth data has high probability under at least one of the clusters:

$$\hat{\Theta} = \arg \max_{\Theta} \prod_{i=1}^N \max_{j=1}^K p(\mathbf{L}_i | \Theta_j) \quad (2)$$

where  $L_i$  is the ground truth pose for one of the  $N$  training images, and  $\Theta_j$  denotes the parameters for one of the PSMs (one per cluster). The pose likelihood given one of the PSMs  $p(\mathbf{L}_i | \Theta_j)$  is defined as in Eqn. 1. Note that the cluster assignment for each image *i.e.* which cluster is responsible for the corresponding pose is treated as a latent variable, determined by maximization over the clusters.

**Algorithm.** To solve Eqn. 2 for the cluster parameters  $\hat{\Theta}$  which best explain the training data we adopt an alternation approach in the style of the Expectation Maximization (EM) or K-means algorithms [2]. Given an initial assignment of training images to clusters we compute the maximum likelihood parameters for each cluster – the distributions over relative position and orientation of connected parts. Each image is then re-assigned to the cluster under which it has the maximum likelihood. The cluster parameters are updated, and the process repeats until convergence, when no images are re-assigned to different clusters. This is a greedy process and thus the algorithm finds a local maximum of Eqn. 2. We therefore run the algorithm several times from randomized starting points: Each training pose is represented as a vector of horizontal and vertical offsets between each joint and the neck joint (to give translation invariance). K-means with random initialization is applied to these pose descriptors to obtain the initial cluster assignments.

Fig. 2 shows an example of the resulting pose clusters. The choice of the number of clusters  $K$  is discussed fully in Section 3.

### 2.3 Appearance modeling by cascaded SVM part detectors

To model the appearance of each part ( $p(\mathbf{f}_i | \mathbf{l}_i)$  in Eqn. 1) we use discriminative training to learn a sliding-window detector. As noted, a different set of detectors is used for *each* cluster. This allows us to implicitly capture the correlation in appearance between parts conditioned on the pose cluster, and reduces the variation in appearance compared to a global model.

**Descriptor.** As in some recent promising work on pose estimation [8, 20] we adopt the HOG descriptor [2] to describe the appearance of an image window. This gives a controlled level of invariance to lighting and local spatial deformation while being efficient to compute.



Figure 4: Examples from our new dataset of 2,000 images (see also Fig. 1 & 5). Images were gathered from Flickr using the tags shown, and have been annotated with 14 joint positions.

Given a hypothesized part position and orientation we compute a HOG descriptor on a part-axis oriented rectangular grid. The image gradient  $\langle \frac{d}{dx}I, \frac{d}{dy}I \rangle$  is estimated at each pixel in the image by convolution using simple derivative filters  $([-1, 0, 1])$  [4] and the magnitude and orientation are computed. The gradient orientation is quantized into a fixed number of discrete ‘bins’ and then regions of neighboring pixels are pooled into a set of spatial ‘cells’. For each cell a histogram of quantized orientations is calculated, with each pixel casting a weighted contribution according to the corresponding gradient magnitude. Gradient orientations are computed unsigned (dark/light vs. light/dark transitions are considered equal) since we do not know *a priori* whether parts will appear light or dark against the background. To avoid quantization artifacts we apply linear interpolation between neighboring orientation bins and use Gaussian weighting when pooling orientations into each cell [4].

**Classifier.** Previous work has typically used simple linear models to represent part appearance [12, 16, 17]. 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 nonlinear kernel – this enables modeling of the multi-modal appearance distribution. We define the likelihood of an image window conditioned on a part position and orientation as proportional to the output of the SVM (the signed distance from the discriminant hyperplane):

$$p(\mathbf{f}|\mathbf{l}) \propto f(\mathbf{x}) = \sum_{i=1}^N \alpha_i k(\mathbf{x}, \mathbf{x}_i) + b \quad (3)$$

where  $N$  is the number of training examples  $\mathbf{x}_i$  (image windows) and  $\alpha_i$  and  $b$  are the learnt weights and bias. We use a radial basis function (RBF) kernel  $k(\mathbf{x}, \mathbf{x}_i) = \exp(-\gamma \|\mathbf{x} - \mathbf{x}_i\|^2)$  to enable arbitrary nonlinear discriminants to be learnt.

**Cascaded reduced set machine.** 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 (Eqn. 3) 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 [19, 20] of simple to complex classifiers (see Fig. 3). 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.

To build classifiers in the cascade with controllable complexity we adopt the reduced set machine (RSM) of Romdhani *et al.* [19], which we briefly summarize here. A nonlinear SVM can be considered as a linear SVM operating on a ‘lifted’ vector in the span of the  $n$  support vectors  $\Psi = \sum_{i=1}^n \alpha_i \Phi(\mathbf{x}_i)$  where  $\Phi$  is the mapping into feature space, by a reduced set expansion: We seek a *reduced set* of  $m$  vectors with which to approximate the input  $\Psi' = \sum_{i=1}^m \beta_i \Phi(\mathbf{z}_i)$  where  $m \ll n$ . Essentially we rewrite the SVM function (Eqn. 3) in terms of a much smaller set of general vectors which are not necessarily a subset of the training vectors. A suitable reduced set is found by minimizing the norm  $\|\Psi - \Psi'\|^2$ . Note that while



Clusters:	1	2	4	6	8
<b>Linear</b>	23.0%	17.9%	24.0%	24.8%	27.8%
<b>Proposed Nonlinear</b>	34.4%	38.2%	<b>49.2%</b>	45.8%	45.7%

Table 1: Part localization accuracy as a function of the number of pose clusters for a linear appearance model and our proposed method (cascaded nonlinear SVM). Results shown are for an equal split of the training data into training and test sets.

$\Phi$  cannot be computed explicitly in the RBF case, the ‘kernel trick’ applies [19] to make the minimization possible.

Using the RSM approach we train a cascade of classifiers with increasing complexity. The  $m$ th stage of the cascade (see Fig. 3) is a nonlinear SVM with  $m$  reduced set vectors, including the  $m - 1$  vectors from the previous stage. We use a total of 50 reduced set machines, and a final cascade stage consisting of the full SVM. Note that there is insignificant computational expense in re-using vectors from previous stages since the kernel values have already been computed. For each stage we learn the bias term from a validation set. At test time windows for which the output of a cascade stage is below zero are discarded (shown below each classifier in Fig. 3). Therefore the number of windows passed to subsequent (more expensive) stages is reduced. At each cascade stage we increase the allowed proportion of false negatives exponentially towards a final target value. This value offers direct control over the amount of windows reaching the final SVM and thus allows a trade-off between accuracy and efficiency.

## 2.4 Combining multiple PSMs

Given a test image, MAP inference for each PSM is carried out in parallel using an efficient dynamic programming algorithm [9]. This yields one pose estimate and corresponding probability per cluster. An aggregation scheme is required to select which pose estimate is most likely – note that simply selecting the estimate with highest probability according to Eqn. 1 performs poorly since the probability is only defined up to an unknown constant, due to the use of discriminative appearance terms, and since the SVM does not output true (log) probabilities. Instead we learn a weight  $w_c$  for each cluster  $c$  such that the pose estimate from the cluster with maximum weighted probability is selected:  $\hat{c} = \arg \max_c w_c p(L_c | I)$ . In order to learn the weights for each cluster we use the maximum likelihood clusters assigned to each training image as ground truth, and a constrained form of multinomial logistic regression [2].

## 3 Experimental Results

In this section we report experiments on the proposed and baseline methods. First we describe the new dataset we have prepared and the evaluation protocol. We then report experiments on the effect of parameter choices and final results.

**Dataset and evaluation protocol.** To support experiments in modeling appearance with a realistic amount of training data we have collected and annotated a new dataset of 2,000 images of full body poses which we will make available publicly. This compares to the 305 images (100 for training) in the IIP dataset which has been used in previous work [11, 12, 16]. To collect the data we queried Flickr for images with the following tags: athletics, badminton, baseball, gymnastics, parkour, soccer, tennis, and volleyball. These image classes contain a range of highly challenging poses. For testing, images were cropped and scaled so that the labeled person is roughly 150 pixels in length – the same scale as the IIP dataset [16].

The dataset is partitioned into two subsets of 1,000 images for training and testing respectively. We strictly conduct all training and parameter selection on the training subset alone. Final results are reported on the test set. We adopt the evaluation criteria proposed

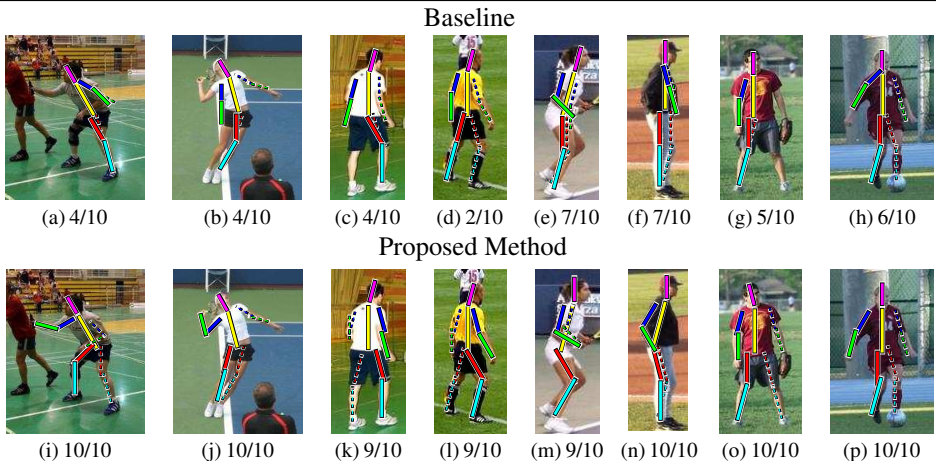


Figure 5: Example results. The top row show results from the baseline method. The bottom row shows results from our proposed method. For each the fraction of correctly localized parts is shown.

by Ferrari *et al.* [10] which has been adopted in recent state-of-the-art approaches [10, 11]: for a given pose estimate a part is considered correctly localized if its predicted endpoints are within 50% of the part length from the corresponding ground truth endpoints.

**Parameter selection.** Our method has essentially a single parameter – the number of pose clusters. The parameters of the SVM part detectors (regularization and bandwidth), and the bias terms for each cascade stage, are selected by cross-validation on the training set. Table 1 reports the results as a function of the number of clusters. Results are shown for linear and nonlinear SVM part detectors. Note that here we train and test on an equal split of the *training* data to avoid optimization on the test data. Using a ‘global’ model (1 cluster) and linear classifier as in previous work gives an overall part localization accuracy of 23%. Replacing the linear part detection stage with our proposed cascaded nonlinear SVM classifier gives a relative performance improvement of nearly 50% to over 34% of parts correctly localized. Partitioning the pose space into four clusters gives a 114% relative improvement in performance over the baseline to over 49%. Fig. 2 visualizes the prior over pose for a global model and four clusters. The learnt clusters clearly approximate poses imaged from the front, left, back and right sides. The left and right side clusters contain a much larger degree of variance in part appearance due primarily to self-occlusion.

**Final results.** Table 2 reports per-part localization results for our 1,000 image test set – all parameters were strictly selected on the training subset in all cases. The first column of Table 2 gives the overall accuracy. We use letters (A)-(E) to refer to rows of the table here.

We first compare the use of a global prior with (A) a linear appearance classifier (as in previous work [11]) and (B) our proposed nonlinear SVM. The overall accuracy improves from 36.4% to 44.7% using our proposed nonlinear classifier. The improvement in some parts is substantial, validating the use of a stronger nonlinear classifier. We next introduce partitioning of the pose space, using 4 clusters as determined to give the best results on the training subset. Recall that the clusters give us both: (i) a more accurate prior; (ii) pose-specific appearance models (one per cluster). Using the linear appearance model (C) the accuracy of many parts improves substantially over the global model (A) with the overall accuracy improving from 36.4% to 43.6%. When we combine the clustered pose with the nonlinear appearance model (E) the overall accuracy improves further to 55.1%. This demonstrates that partitioning the pose space alone does not reduce the variation in appear-



	Method	Total	Torso	Upper Leg	Lower Leg	Upper Arm	Forearm	Head				
	Linear (A)	36.4	64.1	42.4	43.1	41.2	40.7	26.2	23.7	16.5	15.7	49.9
	Nonlinear (B)	44.7	70.9	53.5	58.7	49.3	47.4	37.1	29.1	26.8	18.8	55.9
	4 Cluster Linear (C)	43.6	74.1	54.4	53.6	49.0	49.2	30.5	30.9	17.0	17.7	59.7
4 Cluster Nonlinear (Prior Only) (D)		50.4	74.1	59.6	62.3	54.6	53.5	43.1	40.7	31.7	26.6	58.2
	4 Cluster Nonlinear (E)	55.1	78.1	64.8	66.7	60.3	57.3	48.3	46.5	34.5	31.2	62.9

Table 2: Comparison of part localization rates (in percentages) for experiments on appearance terms, prior and pose-specific appearance. See text for discussion.

	Athletics	Badminton	Baseball	Gymnastics	Parkour	Soccer	Tennis	Volleyball
Part localization %	45.9	59.6	63.2	28.0	32.4	66.7	60.4	59.4

Table 3: Results of the proposed method for subsets defined by activity. See text for discussion.

ance within each cluster sufficiently such that a linear classifier is successful. We finally test if the improvement in accuracy is due to (a) the improved prior alone or (b) the pose-specific appearance terms. To test this we use a single nonlinear appearance term trained on all images, and use the pose clusters only to define the prior. The results (D) show that much of the improvement is due to the improved prior – 50.4% overall compared to 44.7% using a global prior. However, using the pose-specific appearance terms improves the accuracy further to 55.1% overall, a relative improvement of around 10%.

Our final result of 55.1% overall accuracy (E) represents a relative improvement of over 50% over the baseline method. Note also that for some parts the increase in accuracy is very substantial, for example the relative improvement in accuracy of forearm localization is around 100% (34.5% vs. 16.5% and 31.2% vs. 15.7%).

Fig. 5 shows example results for the baseline method (top row) and our proposed approach (bottom row). A number of images show clear improvements due to a more detailed pose prior *e.g.* (i) and (j) have joint angles at the right shoulder and elbow respectively which would receive a low likelihood under the global prior. This can lead to catastrophic failures visible for the baseline method in (a) and (b). Results (c) and (d) show rear views, where the ambiguity between left and right limb appearance leads to the limbs being swapped under the global model. Using our clustered prior we see in (k) and (l) that this ambiguity is overcome. It is also apparent that a more detailed prior improves accuracy – the occluded left arm in (e) is successfully localized in (m) due to the influence of a stronger, pose-specific prior (Fig. 2(e)) which features a higher degree of bending at the left elbow. Finally the improvements due to our stronger appearance models are visible in (n), (o) and (p) where part localization is more accurate. In (f) the right arm is difficult to see due to the lighting direction and clothing color, yet our approach overcomes this in (n).

**Accuracy by activity.** Table 3 shows the accuracy of our proposed method for each activity (by Flickr tag) present in our dataset, giving an insight into the broad range of pose difficulties present. The gymnastics and parkour images generally contain highly challenging and variable poses (Fig. 4). For the categories which tend to feature in other datasets – namely soccer and baseball – performance is much better, indicative of the more ‘stable’ poses that players adopt.

**IIP results.** Table 4 shows results of our proposed method (using models trained on our dataset) and previous approaches on the 205 test images of the IIP dataset [16]. Our proposed method gives overall accuracy of 66.2%, compared to the previous best reported result of 56.4% [16] – a relative improvement of over 17%. This shows clearly that more training data combined with the ability to model the multi-modalities present lead to a higher level of part localization accuracy. Additionally, the result validates that our proposed method generalizes across the datasets, and that the improved results are not due to over-fitting.

Method	Total	Torso	Upper Leg	Lower Leg	Upper Arm	Forearm	Head				
Ramanan[16]	27.2	52.1	30.2	31.7	27.8	30.2	17.0	18.0	14.6	12.6	37.5
Andriluka <i>et al.</i> [10]	55.2	81.4	67.3	59.0	63.9	46.3	47.3	47.8	31.2	32.1	75.6
Johnson & Everingham[17]	56.4	77.6	64.9	58.1	57.6	52.2	55.6	50.7	42.4	36.1	68.8
4 Cluster Nonlinear	66.2	85.4	76.1	70.7	69.8	61.0	64.9	64.4	49.3	44.4	76.1

Table 4: Comparison of part localization rates (in percentages) for previous approaches and ours on the IIP dataset[17]. See text for discussion.

## 4 Conclusions

We have shown that by combining a mixture model of pose with state-of-the-art nonlinear classifiers for *pose-specific* part appearance it is possible to model the high levels of variation present in natural images of human poses. Overall we achieve greater than 50% relative improvement in accuracy over a state-of-the-art method on our new dataset and over 17% relative improvement on the IIP dataset [17]. Our results show that performance is highly-dependent on the activity (*e.g.* gymnastics) and that future work might profitably concentrate on targeting these challenging activity classes in terms of developing efficient ways to collect large training sets, further refinement of pose priors, incorporation of scene information, and even stronger appearance models.

## References

- [1] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *Proc. CVPR*, 2009.
- [2] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [3] P. Buehler, M. Everingham, D. Huttenlocher, and A. Zisserman. Long term arm and hand tracking for continuous sign language TV broadcasts. In *Proc. BMVC*, 2008.
- [4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.
- [5] J. Deutscher, A. Blake, and I. Reid. Articulated body motion capture by annealed particle filtering. In *Proc. CVPR*, 2000.
- [6] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *Proc. BMVC*, 2009.
- [7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1), 2005.
- [8] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multi-scale, deformable part model. In *Proc. CVPR*, 2008.
- [9] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proc. CVPR*, 2008.
- [10] D. Hogg. Model-based vision: A program to see a walking person. *IVC*, 1(1), 1983.
- [11] S. Ioffe and D. Forsyth. Mixtures of trees for object recognition. In *Proc. CVPR*, 2001.
- [12] S. Johnson and M. Everingham. Combining discriminative appearance and segmentation cues for articulated human pose estimation. In *Proc. MLVMA*, 2009.

- [13] M. Kumar, A. Zisserman, and P. Torr. Efficient discriminative learning of parts-based models. In *Proc. ICCV*, 2009.
- [14] M. Lee and I. Cohen. Human upper body pose estimation in static images. In *Proc. ECCV*, 2004.
- [15] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10), 2005.
- [16] D. Ramanan. Learning to parse images of articulated bodies. In *Proc. NIPS*, 2006.
- [17] D. Ramanan and C. Sminchisescu. Training deformable models for localization. In *Proc. CVPR*, 2006.
- [18] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: tracking people by finding stylized poses. In *Proc. CVPR*, 2005.
- [19] 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.
- [20] H. Sidenbladh and M. Black. Learning the statistics of people in images and video. *IJCV*, 2003.
- [21] L. Sigal, S. Bhatia, S. Roth, M. Black, and M. Isard. Tracking loose-limbed people. In *Proc. CVPR*, 2004.
- [22] P. Viola and M. Jones. Robust real-time object detection. *IJCV*, 2001.
- [23] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Proc. ECCV*, 2008.
- [24] L. Xiangyang and D. Huttenlocher. Beyond trees: Common-factor models for 2d human pose recovery. In *Proc. ICCV*, 2005.