Learning Effective Human Pose Estimation from Inaccurate Annotation

Sam Johnson and Mark Everingham
School of Computing
University of Leeds
{s.a.johnson04|m.everingham}@leeds.ac.uk

Abstract

The task of 2-D articulated human pose estimation in natural images is extremely challenging due to the high level of variation in human appearance. These variations arise from different clothing, anatomy, imaging conditions and the large number of poses it is possible for a human body to take. Recent work has shown state-of-the-art results by partitioning the pose space and using strong nonlinear classifiers such that the pose dependence and multi-modal nature of body part appearance can be captured. We propose to extend these methods to handle much larger quantities of training data, an order of magnitude larger than current datasets, and show how to utilize Amazon Mechanical Turk and a latent annotation update scheme to achieve high quality annotations at low cost. We demonstrate a significant increase in pose estimation accuracy, while simultaneously reducing computational expense by a factor of 10, and contribute a dataset of 10,000 highly articulated poses.

1. Introduction

The task of human pose estimation is to estimate the configuration of a person's body parts – or 'pose' – in an image. We tackle an unconstrained still-image scenario where we do not have training examples of the person depicted, and their activity, anatomy and clothing are unknown, as is the background of the scene. Pose estimation in natural images is an important goal as it forms a building block in enabling high-level scene understanding in both images and video. Applications of such methods include content-based image indexing and retrieval, activity understanding, automated surveillance, markerless motion capture and human-computer interaction. The task is highly challenging for a multitude of reasons including: (i) the wide variation of human appearance due to differing anatomy (large people, small people, babies *etc.*) and different clothing including loose-fitting clothes and clothing which hides the configuration of underlying body parts such as dresses; (ii) the wide range of poses people are able to take – often leading to self-occlusion such as folding the arms in front of the body or
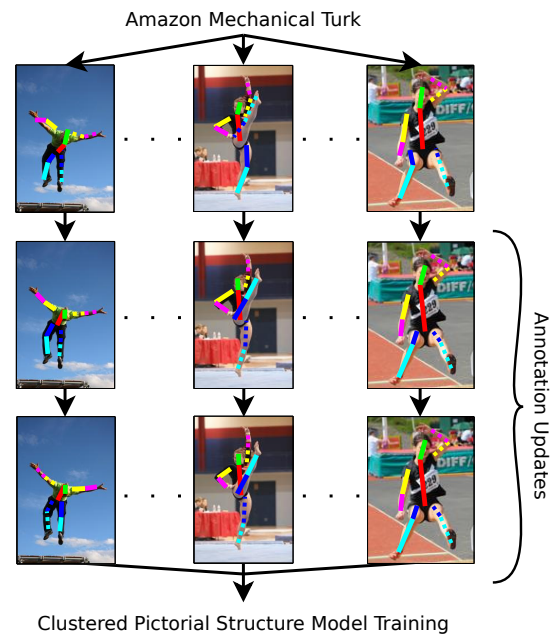

Figure 1. Improving the quality of annotations. Annotations are gathered from inexperienced Amazon Mechanical Turk workers and exhibit localization and structural errors e.g. left/right confusion. We propose an iterative annotation update scheme to improve the annotations for use in a state-of-the-art human pose estimation framework.

hiding the furthermost limbs when stood side-on to the camera; (iii) classic computer vision problems such as cluttered and widely varying backgrounds and different imaging conditions. Due to our focus on challenging, unconstrained still images we are unable to include prior knowledge of the action being performed (*e.g.* running) to constrain the pose space or prior knowledge of the appearance of a person or background, meaning techniques such as background subtraction cannot easily be exploited.

State-of-the-art pose estimation methods [1, 8, 9, 13, 16, 17, 18, 24] typically represent the human body as a graphical model composed of ten major body parts corresponding to the head, torso and upper- and lower-limbs (Figure 2a). These ten parts form the nodes in a graph, with the edges be-

tween them representing 'spring-like' kinematic links capturing inter-part spatial relationships such as their relative position and orientation. We base our approach on an extended pictorial structure model (PSM) [11, 17]. The PSM has shown previous success in pose estimation and allows efficient inference over the space of possible poses – we give a brief definition of the framework in Section 2.1.

Our 'clustered' extension to PSMs [17] effectively replaces a single PSM with a mixture model comprised of PSMs [15, 26] specialized, both in terms of pose *and appearance*, to a particular region of the pose space. By dividing the pose space, the model captures pose-specific appearance terms tied to more informative prior distributions over pose. We contribute several further enhancements over the PSM: (i) we incorporate clustering of partial poses to account for images with poses containing occluded parts; (ii) we greatly improve both accuracy and computational efficiency by modeling part appearance using a state-of-the-art mixture of linear classifiers [12].

Current pose estimation methods are limited in their applicability to general images for two reasons: (i) the models of appearance do not generalize well to arbitrary poses, which may include high levels of self-occlusion; (ii) the prior models of pose configuration tend to be restricted to mainly upright people performing relatively common actions. The root of these two issues can be traced to the reliance on small and relatively constrained datasets. Acquiring datasets for full-body pose estimation is an onerous task due to the time-consuming and detailed annotation required. As an example, consider the commonly-used 'IIP' dataset assembled by Ramanan [18]; this contains 305 images, each labeled with 14 joint locations defining a skeletal structure. The training and testing protocol for this dataset specifies 100 images to be used for training – clearly this is very few from which to learn all the possible variation in human appearance. Figure 2b shows a scatterplot of stick-men [24] representing every pose contained within this dataset normalized to the neck joint. It is clear that there is a strong bias in pose, with the majority of poses close to upright (note color-scheme in Figure 2a). We recently introduced the 2,000 image Leeds Sports Pose (LSP) dataset [17] with the aim of addressing this issue of constrained poses. With reference to Figure 2c one can see that while the range of poses is much larger there still exists a strong bias towards upright poses. This leads to models which perform well for people in typical upright poses e.g. standing or walking, but which fail for people in less common poses e.g. arising from activities such as exercise, fighting, dancing or gymnastics which might be argued to be the more salient for the pose estimation task.

We propose methods to overcome the difficulty of full-body pose dataset collection by leveraging recent 'crowd-sourcing' technologies. We dispatch images to Amazon
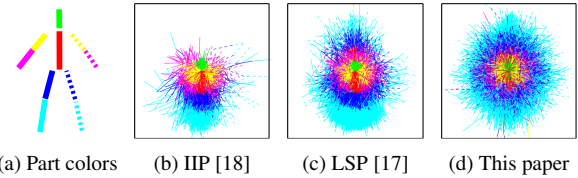
(a) Part colors   (b) IIP [18]   (c) LSP [17]   (d) This paper

Figure 2. A notion of dataset difficulty. We render (in a similar style to the upper-body visualization by Tran & Forsyth [24]) poses present in three full-body pose estimation datasets normalized so the neck is at the center of the image. Our new dataset (d) contains a much richer range of poses than those present in the IIP [18] or LSP [17] datasets.

Mechanical Turk (AMT) where a large number of inexpert users supply annotations with higher information content than present in previous datasets. We then propose a novel scheme to improve these sometimes low quality annotations, giving us a large, high-quality dataset with which to train our pose estimation technique (Figure 1). We demonstrate that our approach gives better accuracy over the entire pose space – using more training data we are able to increase the number of pose clusters while still learning meaningful prior and appearance terms. We show that our method leads to greater than 10% relative performance improvement over the state-of-the-art results.

Related work. Rather little previous work has investigated dataset provision for pose estimation. The PASCAL VOC challenge [10] includes a 'person layout' task with around 550 images annotated with bounding boxes of head, hands and feet. Eichner *et al*. have annotated images in the PASCAL VOC datasets and selected frames from 'Buffy' video [8] with upper-body pose. Tran & Forsyth [24] have recently noted the limited pose variation in this dataset. Bourdev & Malik [4] have annotated a dataset of 1,000 still images with 3-D joint positions, using an interactive approach combining a tool for manual 2-D annotation and definition of constraints with automatic 2-D to 3-D 'lifting' of the pose. However, as noted [3] the annotation process demands considerable skill, and hence their more recent work on person detection has reverted to 2-D keypoint annotation [3]. For other vision tasks, internet-based data collection has proven useful, for example the collaborative LabelMe object annotation system [21]. Sorokin & Forsyth [23] have explored dataset collection from AMT for general vision tasks. Key issues in the use of AMT lie in obtaining data of the required quality from inexperienced annotators, at an affordable cost.

For the task of object detection, several pieces of work [12, 25] have proposed to acknowledge that annotation provided for training may be inaccurate, for example an object bounding box may be imprecise, casting the learning task as one of multiple instance learning. Felzenszwalb *et al*. [12] effectively learn the 'corrected' bounding box as part of training, by an iterative scheme alternating between model learning and refinement of the annotation. We pro-

pose a similar approach to align our training data to improve the quality of part descriptors learnt from imprecise pose annotation provided by AMT workers.

In the area of articulated human pose estimation and tracking much work has focused on either improved models of body part appearance or valid configurations (pose prior). In the original work on the PSM Felzenszwalb & Huttenlocher [11] utilize color box filters, assuming that the color of each body part is known *a priori* – such an approach does not extend to unconstrained, natural images. The Iterative Image Parsing (IIP) approach of Ramanan [18] learns a (linear) weighted edge template for each part, and bootstraps a part color model specific to each image, thus exploiting general (edge) and image-specific (color) information. Ferrari *et al.* [13, 8] also use image-specific color models, based on the prior location of parts within a bounding box for upper-body pose estimation, however Tran & Forsyth [24] have shown that this method is only applicable to tightly constrained datasets such as TV footage where the majority of people face the camera in upright poses.

Recent work has used state-of-the-art image descriptors and discriminative machine learning methods to model the appearance of parts in a color-invariant fashion. A number of methods [5, 16, 17, 24] have used the Histogram of Oriented Gradient (HOG) descriptor [6] which gives a controlled degree of invariance to lighting and minor spatial deformations, applying nearest neighbor, linear or nonlinear support vector machine classifiers. Others [1, 27] have used the similar shape-context descriptor and boosted classifiers.

Most recent methods [1, 5, 8, 9, 13, 16, 17, 18, 19, 22, 24, 27] use a graphical model to capture the distribution over plausible human poses. The PSM approach of Felzenszwalb and Huttenlocher [11] has been widely adopted, due to its provision for globally optimal inference and efficient sampling, in contrast to methods using non-tree-structured graphs [24] which must resort to approximate inference methods. Our recent work has tackled some of the original issues with the PSM – its failure to handle self-occlusions and the broad, non-descriptive priors over pose – by clustering the training images in pose space [17] to form a mixture of trees. We build upon this approach – which achieves the state-of-the-art full-body pose estimation performance – contributing a number of extensions and efficiency improvements, and show how such a model can be used to improve low quality training data.

Outline. Our proposed method is described in Section 2. We cover: (i) the PSM framework and pose space clustering upon which we build; (ii) our proposed method for efficient discriminative body part detection. Section 3 describes our proposed methods for learning from low quality annotations, including (i) modeling of annotator errors; (ii) an iterative learning scheme for updating noisy anno-
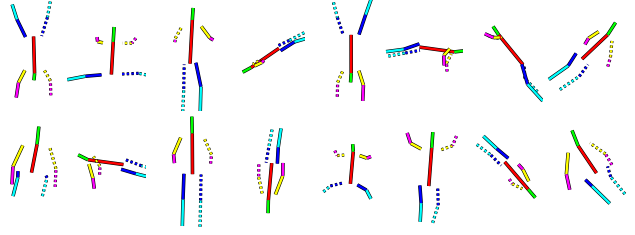

Figure 3. Pose cluster centroids for 16 clusters. We initially cluster our 11,000 training images using a variation of k-means which can handle unlabeled parts (Eqn. 3 and 4). The high degree of variability conveys the wide range of poses present in our new dataset. Solid and dashed lines represent right and left body parts respectively.

tation. Section 4 introduces our new dataset and reports experimental results. We offer conclusions in Section 5.

2. Method

In this section we describe the basis of our method and propose extensions to the state-of-the-art clustered PSM approach [11, 17] which improve both the accuracy and efficiency of body part appearance models and handle training using poses with incomplete part visibility.

2.1. Pictorial Structure Model

We define a PSM [11] which models a person as a connected collection of 10 parts – the head, torso and upper- and lower-limbs. To support efficient inference these parts are represented as nodes in a tree-structured graph with connecting edges capturing distributions over the relative layout between them. A configuration of the model is parameterized by $L = \{\mathbf{l}_1, \mathbf{l}_2, \ldots, \mathbf{l}_n\}$ where $\mathbf{l}_i = \langle x_i, y_i, \theta_i \rangle$ specifies the 2-D position and orientation of a part. Given this formulation the posterior probability of a configuration – or pose – given an image I can be written as

$$p(L|I, \Theta) \propto p(I|L, \Theta)p(L|\Theta) \qquad (1)$$

and given the tree structure of the PSM the posterior can be factorized into unary and pairwise terms acting on the individual parts

$$p(I|L, \Theta)p(L|\Theta) = \prod_i p(\mathbf{f}_i|\mathbf{l}_i, \Theta_i) \prod_{(\mathbf{l}_i, \mathbf{l}_j) \in E} p(\mathbf{l}_i|\mathbf{l}_j, \Theta_{ij}) \qquad (2)$$

where the appearance of each part is considered independent. This formulation gives two terms which must be modeled from training data: (i) an appearance term measuring how well an image region matches a model of part appearance; (ii) a prior term defining the probability of a configuration of connected parts.

Appearance term. The appearance term $p(\mathbf{f}_i|\mathbf{l}_i, \Theta_i)$ captures the compatibility between an image region $\mathbf{f}_i$ and a corresponding hypothesized part location $\mathbf{l}_i$ given the appearance model Θ_i of part i. The appearance term is com-

1467

puted exhaustively for all possible part positions and orientations in a sliding-window fashion. As in previous state-of-the-art approaches [1, 16, 17] we model the appearance of each part discriminatively. As described in Section 2.3 we propose to use a mixture of linear SVMs to capture part appearance represented by HOG descriptors.

Prior term. The prior term $p(\mathbf{l}_i|\mathbf{l}_j, \Theta_{ij})$ measures the prior probability of the configuration of two connected parts i and j. This serves two purposes: (i) encouraging relative part layouts which are more common – helping to overcome ambiguous image data; (ii) constraining the model to plausible human configurations given the learnt kinematic limits of the body. We model the relative orientations and locations of parts Θ_{ij} as Gaussians in a transformed space [11, 20], allowing efficient inference using distance transforms.

2.2. Clustered Pictorial Structure Model

The PSM model described is limited by the use of a simple Gaussian prior on pose and the assumption that part appearance is independent of pose and other parts. We extend our previous work [17] using a set of PSM models, learnt by clustering in pose space, to overcome these limitations. Clustering the images in pose space gives a number of benefits: (i) the prior within each cluster is more descriptive by capturing a tighter configuration of parts; (ii) dependencies between parts which are not connected in the tree can be captured implicitly; (iii) learning part appearance models specific to each cluster captures the correlation between pose and appearance, such as the appearance of a head from different viewpoints.

A set of clustered PSMs is sought which maximizes the likelihood of the training poses. First an initial clustering is performed using k-means on annotated joint locations normalized w.r.t. the neck joint. For each cluster a PSM prior model is then built, and images are re-assigned to the cluster whose model gives the highest likelihood for the corresponding pose. The process is repeated until no further cluster re-assignments occur.

Clustering incomplete poses. While previous datasets have been 'completely' annotated, i.e. all 14 body joints are specified for every image, this has been achieved by informed guesses of the location of occluded joints. In collecting our new dataset using inexperienced AMT workers, we specified that hidden joints should not be annotated, ensuring that all data is the result of observation rather than guessing, but resulting in annotations with some missing elements. We accommodate such missing data by modifying the k-means procedure such that clusters $\mathcal{C}_1, \ldots, \mathcal{C}_k$ (given a random sample of k initial means $\mathbf{m}_1, \ldots, \mathbf{m}_k$) are assigned poses $\mathbf{x}_1, \ldots, \mathbf{x}_N$ as follows

$$r_{ni} = \begin{cases} 1 & \text{if } i = \arg\min_j d_{mx}(\mathbf{x}_n, \mathbf{m}_j) \\ 0 & \text{otherwise} \end{cases} \qquad (3)$$

where r_{ni} is a 1-of-K coding vector with $r_{ni} = 1$ if pose vector n is assigned to cluster i. The function d_{mx} denotes the squared Euclidean distance between the *visible* points of two poses:

$$d_{mx}(\mathbf{m}, \mathbf{x}) = \sum_{p \in \mathcal{V}^x \cap \mathcal{V}^m} (\mathbf{x}_p - \mathbf{m}_p)^2 \qquad (4)$$

where $\mathcal{V}^u$ is the set of visible (hence labeled) points in vector $\mathbf{u}$. Cluster means $\mathbf{m}$ are also computed w.r.t. visibility.

Given this initial set of cluster assignments we learn a mixture of PSMs – one PSM for each cluster. This involves an EM-like alternation scheme: (i) estimation of prior distributions over relative orientation and offset for all *visible* parts in poses assigned to each cluster; (ii) re-assignment of each pose to the cluster under which it has the maximum likelihood. This process is repeated until no further cluster re-assignments take place.

2.3. Appearance Modeling

Previous work [1, 5, 16, 17, 18, 22, 27] has demonstrated the effectiveness of modeling part appearance using discriminative methods, i.e. training a sliding-window classifier for each part. Earlier work [5, 16, 18] has shown promising results using simple linear classifiers. However, the appearance of a body part can be expected to be multimodal, for example due to differences in anatomy, clothing or lighting. Our recent work [17] has shown that using state-of-the-art nonlinear classifiers (cascaded SVMs with RBF kernel), capable of capturing this multi-modal appearance, gives significantly improved results, but at the cost of a substantial increase in computational expense. We propose here to use a much more efficient classifier architecture, the mixture of linear SVMs [14], which gives improved accuracy and greatly decreased computational expense.

Clustering. Our model has two mechanisms to cope with multi-modal part appearance: (i) partitioning the pose space reduces variation in appearance due to the pose of a part; (ii) part appearance is modeled using nonlinear classifiers, by clustering part appearance *within* a pose cluster such that an individual linear classifier is responsible for each mode of appearance. Clustering is performed using a weighted form of k-means, with multiple restarts to avoid local minima. We weight elements of the descriptors by training a single linear SVM weight vector $\mathbf{w}$ from the entire set of training images assigned to each pose cluster, effectively modifying the distance used within the k-means algorithm:

$$d_{mx}(\mathbf{m}, \mathbf{x}) = \sum_i \mathbf{w}_i(\mathbf{x}_i - \mathbf{m}_i)^2 \qquad (5)$$

Using the learnt weights in this way places higher weight on the more discriminative features belonging to a part and less weight on features commonly belonging to the background. This overcomes the problem of separating the part descriptors based on the level of noise in the background rather than the appearance of the part itself.

1468

Classification. For each cluster of part descriptors (within a pose cluster) a linear SVM classifier is trained. Each classifier forms a mixture component of the overall appearance model for a given pose cluster. The number of components used by each body part is allowed to vary, and determined at training time by cross-validation. This gives optimal efficiency and accuracy by allowing parts with a higher degree of variation – such as the head and torso – to be modeled with higher fidelity. Each classifier is bootstrapped using negative samples taken from outside the body regions in the training set. Classifier responses are computed per part i as

$$p(\mathbf{f}_i|\mathbf{l}_i, \Theta_i) \propto \max_{j=1...n} \mathbf{w}_j^T \Phi(\mathbf{f}_i) \qquad (6)$$

where n is the number of mixture components, $\mathbf{w}_j$ is the weight vector for mixture component j and $\Phi(\mathbf{f}_i)$ is the feature vector extracted from image region $\mathbf{f}_i$. The $\max$ operator allows the mixture component with the highest confidence to dominate the other components which may be tuned to other appearance modes.

Part descriptors. As in recent state-of-the-art work on pose estimation [5, 16, 17] we adopt HOG-like descriptors [6] to describe the appearance of an image region. The image gradient $\langle \frac{d}{dx}I, \frac{d}{dy}I \rangle$ is estimated at each pixel using $[-1, 0, 1]$ derivative filters [6] and the gradient magnitude and orientation are computed. The image region is divided into a number of spatial 'cells' each containing a histogram of gradient magnitude over a set of quantized orientation bins. Gaussian spatial weighting and linear interpolation of magnitude across orientation bins are used to ensure smoothly-varying descriptors. These features give a controlled degree of invariance to slight orientation and spatial deformation and are invariant to color which, as noted, is *not* known a-priori for people present in an image.

2.4. Pose Cluster Selection

Given a novel image it is not known *which* of the learnt PSMs should be used i.e. in which pose cluster the depicted pose lies. The PSM posterior cannot directly be used to select the cluster since it is un-normalized. To overcome this a classifier is trained using multinomial logistic regression [2], effectively learning the relationship between the uncalibrated posterior (Eqn. 2) for each PSM and the corresponding best cluster assignment. Training data is gathered by estimating the pose present in each training image (with known cluster assignments) using each clustered PSM. During testing we estimate the pose in each image with *all* PSMs and select the best using the learnt classifier i.e. by taking the maximum over weighted PSM outputs.

### 3. Learning from Inaccurate Annotations

Amazon Mechanical Turk (AMT) has seen a surge in use recently for collecting large vision datasets [7, 23, 10], allowing efficient and inexpensive completion of simple tasks

such as image annotation. We wish to use AMT to collect a large dataset for human pose estimation, at least an order of magnitude larger than previously available. However, in the majority of previous applications the required annotations are relatively simple – image tags or bounding box labeling for example. We require accurate labeling of 14 distinct locations on the body along with an indication of the visibility of each. While AMT allows rapid collection of large numbers of such annotations they are sometimes inaccurate or even completely incorrect (see Figure 4). We propose to model and account for these errors in an iterative learning scheme – improving the quality of poor annotations to achieve a large amount of high quality training data.

Dataset collection. We aim to improve performance primarily on the more challenging poses humans take. In the LSP Dataset [17] we showed that the most challenging poses lie in the activities of gymnastics, parkour and to a lesser degree athletics. We queried Flickr with these tags and manually selected 10,800 images of different people in a wide range of extremely challenging poses. Distributing these images to AMT – along with a purpose-built annotation tool – we are able to acquire roughly 400 annotations per hour. However this efficiency comes at the cost of accuracy – some annotations exhibit localization errors, errors in 'body structure' (e.g. left/right confusion) or are completely unusable (examples in Figure 4).

Removing unusable annotations. In some cases users appear to have randomly annotated the image, or marked all points as invisible – these annotations we consider 'unusable'. In the latter case identification is trivial. In the former case identification requires further thought. Rejecting annotations as unusable leads to the AMT worker being unpaid – encouraging people to supply high quality annotations in later tasks. Due to this we must take care to only reject truly bad annotations. To semi-automate the process we rank the annotations by learning a Gaussian mixture model over part lengths and relative orientations from the LSP dataset. We then order the 10,800 annotations by their likelihood under this model and reject any which are unusable. For the final dataset we randomly sample 10,000 images from the accepted annotations.

Error modeling. We propose to take into account the errors present in AMT annotations by treating the true joint locations and body structure as *latent* variables, with the AMT annotations being *noisy* observations thereof. From a subset of 300 images for which we have both 'expert' ground-truth and AMT annotations we estimate two models of the kind of errors made by AMT workers on this task: (i) we assume that the joint locations annotated by AMT workers are distributed according to an isotropic Gaussian distribution over horizontal and vertical displacement with mean at the true location; (ii) the set of most common structural errors (including no error)

1469
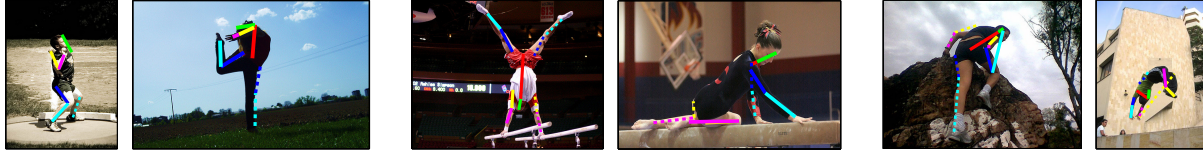
| (a) Inaccurate body point labeling | (b) Anatomical structure errors | (c) Unusable annotations |

Figure 4. Annotations received from AMT workers. AMT enables a large number of annotations to be collected efficiently at the cost of accuracy, as shown here. (a) shows poor localization of the annotated body points, (b) shows structural errors: left/right parts confused and arms/legs confused, (c) shows unusable annotations where some parts lie far from their true locations.

$\mathcal{S}$ is determined from a sample of annotations: {*No error*, *Left/right switched*, *Left/right arms switched*, *Left/right legs switched*, *Arms/legs switched*, *Arms/legs switched and left/right switched*, *Arms/legs switched and left/right arms switched*, *Arms/legs switched and left/right legs switched*}. Lacking a large amount of data with expert ground truth we assume that the AMT workers' annotation exhibits each one of these structural errors with equal probability.

Iterative learning method. Using these learnt models of annotation error we can iteratively improve the quality of the noisy AMT annotations toward what could be considered 'expert' ground-truth. In effect, the learning task is posed as one of multiple instance learning [12, 25], where the AMT annotation for an image defines a *bag* of plausible true annotations. A two-step process similar to that used by Felzenszwalb *et al*. [12] for refining object bounding box annotation is applied: (i) an initial set of part appearance models is learnt from the raw annotations; (ii) the latent joint locations and body structure are updated w.r.t. the learnt error models. This can be understood as alternating between learning the model, and selecting the 'best' annotation in each bag which is both plausible given the noisy annotation and which agrees with the current learnt model.

Updating the latent joint locations involves computing the set of all possible part locations $\mathcal{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_n\}$ where $\mathbf{l}_n = \langle x_n, y_n, \theta_n \rangle$. We assume uniform probability for joint locations lying within 2σ of the annotated locations under our learnt error model for each joint. The best location l_i^* of part i is then computed as

$$\mathbf{l}_i^* = \arg\max_{\mathbf{l}_n \in \mathcal{L}} p(\mathbf{f}_n | \mathbf{l}_n, \Theta_i) \qquad (7)$$

where $p(\mathbf{f}_n | \mathbf{l}_n, \Theta_i)$ is computed as in Eqn. 6.

Given the set of structural errors $\mathcal{S}$ we compute the most likely true part configuration L^* as

$$L^* = \arg\max_{\mathbf{L} \in \mathcal{S}} \prod_{i=1}^{10} p(\mathbf{f}_i | \mathbf{l}_i^*, \Theta_i) \qquad (8)$$

where we recompute $\mathbf{l}_i^*$ as in Eqn. 7 for each $L \in \mathcal{S}$, and as noted each structural error type is assumed equally probable.

Iterations alternating between update of the model and annotation can continue until the annotation does not change. In practice, as reported in Section 4 we find that

Pose Clusters:	1	2	4	6	8
Linear [17]	23%	18%	24%	25%	28%
Nonlinear Cascade [17]	34%	38%	49%	46%	46%
Linear Mixture	30%	36%	**50%**	47%	46%

Table 1. Part localization accuracy as a function of the number of pose clusters. We include results for a single linear classifier and a nonlinear cascaded SVM used in our previous work [17].

Pose Clusters:	1	2	4	8	16	32
Linear Mixture	34.7%	49.5%	53.4%	56.0%	**58.5%**	57.9%

Table 2. Part localization accuracy as a function of the number of pose clusters. We train using the first half of the 1,000 image training subset of the LSP dataset [17] plus our new 10,000 image dataset and test on the second 500 images of the LSP training set.

just a few iterations are sufficient. Figure 1 shows how erroneous and inaccurate AMT annotation can be improved; quantitative results are reported in Section 4.

4. Experimental Results

In this section we report results of experiments using our proposed mixture of linear part classifiers in comparison with previous work on the LSP dataset [17] along with results obtained when training with our new 10,000 image dataset and multiple rounds of annotation updates.

Dataset and evaluation protocol. As discussed in Section 3 we contribute a new dataset of 10,000 images of people in a wider range of poses than any previous full-body pose estimation dataset (See Figure 2). This training set is ten times the size of the largest such annotated dataset to our knowledge [17] and will be released publicly. Images were scaled such that the annotated person is roughly 150 pixels in length – as done in the IIP and LSP datasets [17, 18].

For our experiments we use the protocol of Johnson & Everingham [17] and split the LSP dataset into two subsets. We conduct all training and parameter selection on the training subset alone and evaluate the number of pose clusters on the latter 500 images of this subset. For our final experiments we train on the 1,000 image LSP training set combined with our new 10,000 image dataset. For evaluation we use the Percentage Correct Parts (PCP) criteria proposed by Ferrari *et al*. [13] and adopted in recent state-of-the-art approaches [1, 16, 17, 22]: a part is considered correctly localized if its predicted endpoints are within 50% part length of the corresponding ground truth endpoints.
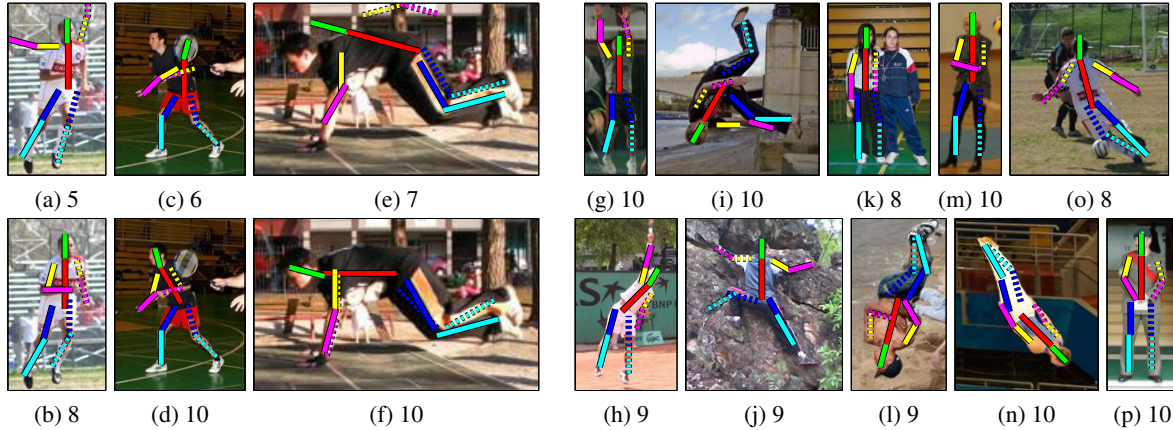
Figure 5. Output of our proposed method on images from the LSP test set [17] and number of correct parts out of 10. Images (a)–(f) show the improvements in accuracy after annotation updates. Images (g)–(p) show a selection of other results. See text for discussion.

Method	Total	Torso	Upper Leg	Lower Leg	Upper Arm	Forearm	Head
Clustered PSM [17] (1,000 training images)	55.1	78.1	64.8 66.7	60.3 57.3	48.3 46.5	34.5 31.2	62.9
Proposed (16 clusters, no updates)	59.9	85.9	71.4 71.6	63.4 63.8	50.5 52.1	33.9 35.6	70.8
Proposed (16 clusters, 1 update)	62.1	87.0	74.6 73.0	66.3 65.0	52.9 54.1	**37.1** 37.5	73.6
Proposed (16 clusters, 2 updates)	**62.7**	**88.1**	**75.2 73.8**	**66.7 66.3**	**53.0 54.4**	36.1 **38.9**	**74.6**

Table 3. Part localization rates (in percentages) on the LSP test set for a previous approach and ours. See text for discussion.

Appearance classifiers. Using a mixture of linear classifiers to model the appearance of parts has two aims: (i) we wish to capture the multi-modal nature of the intra-cluster appearance of body parts; (ii) we require an efficient classifier which can be computed exhaustively at all locations $\mathbf{l} = \langle x, y, \theta \rangle$ in the image. We first evaluate our proposed classifiers under the same protocol as Johnson & Everingham [17]. Table 1 shows the results using the first 500 images of the LSP training set [17] for training, and the second 500 for testing. Our classifiers perform extremely well, outperforming the nonlinear cascaded SVMs used previously [17] in the case of 4 and 6 clusters. Performance drops slightly at 1 and 2 clusters due to the relatively large number of training samples and our upper limit on the number of linear mixture components. Here the nonlinear SVM successfully captures the larger number of appearance modes present. The results shows that our classifier architecture is far more computationally efficient (by approximately 20 times) and is capable of equaling – or exceeding – the performance of more expensive approaches under an optimal number of pose clusters.

Optimal clustering. As can be seen in Figure 2 our new dataset contains a much higher degree of pose variation than previous datasets. This suggests that a larger number of pose clusters may be required to capture the correlations between pose and appearance present. Dividing previous datasets into a large number of clusters is counterproductive – each cluster contains too little training data to learn meaningful appearance models. Table 2 shows that by combining our new dataset with the current LSP training (giving 10,500 training images for this experiment) – we can partition the pose space into a much higher number of clusters and achieve increasing accuracy up to around 16 clusters where the accuracy improvements level off. This shows that pose estimation benefits from a larger amount of increasingly varied training data.

Annotation updates. We evaluate the effectiveness of our method for learning from noisy AMT annotation by training with 11,000 images (10,000 AMT and 1,000 LSP) and applying multiple rounds of annotation updates to the 10,000 AMT annotations. Table 3 reports the pose estimation accuracy for each round of annotation updates. We see that after a single update the accuracy improves by around 2.2% absolute accuracy. This improvement occurs across all body parts and shows that our update scheme successfully aligns the data leading to stronger appearance models. This is also shown in Figure 5(a)–(f). In all three initial images the part classifiers have been drawn to locations lying either in highly cluttered background or in the case of (c) a circular racquet with strong gradient features – similar in appearance to a face when using HOG features. In the bottom row – after performing annotation updates – all three estimated poses have improved considerably with body parts lying in their true locations. Only image (b) does not score 10/10 parts under our evaluation criteria. This is due to our method not currently searching over part length to handle foreshortening – the left arm is much shorter in the image than if it were parallel to the image plane. This can also be seen for lower right leg in (o) and lower left arm in (h).

Final results. As shown in Table 3, after a further round of updates we improve accuracy in all but one case (left forearm). In comparison to the previous state-of-the-art approach [17] we obtain a relative improvement in overall accuracy of 14% (62.7% *vs.* 55.1%). On some parts in par-

1471

Method	Total	Torso	Upper Leg	Lower Leg	Upper Arm	Forearm	Head
Ramanan[18]	27.2	52.1	30.2 31.7	27.8 30.2	17.0 18.0	14.6 12.6	37.5
Andriluka *et al.* [1]	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]	66.2	85.4	**76.1** 70.7	**69.8** 61.0	64.9 64.4	**49.3** **44.4**	76.1
Proposed	**67.4**	**87.6**	**76.1** 73.2	68.8 **65.4**	69.8 64.9	48.3 43.4	**76.8**

Table 4. Part localization rates (in percentages) for previous approaches and ours on the IIP test set[18]. See text for discussion.

ticular our improvements are substantial – left forearm accuracy is improved by around 25% and the head by around 20%. Figures 5(g)-(p) show example output. It is clear that we can now handle more 'extremely' articulated poses such as those in (i) and (l). We are also able to handle images with extremely noisy backgrounds – (j) and (n) in particular contain regions of dense noise when using gradient features.

Table 4 reports our results on the IIP dataset [18] and comparison to previous methods. We achieve a modest improvement over the best overall results reported to date [17] (67.4% *vs.* 66.2%). As noted, this dataset is small, consisting mainly of upright poses, such that the improvement in 'harder' poses is not more apparent.

5. Conclusions

We have shown that a large pose estimation dataset can be collected efficiently using AMT, but at the cost of inaccurate and sometimes grossly erroneous annotation. By incorporating simple models of such annotator error in the training process, we show that such noisy annotation can be used to substantially improve performance of pose estimation on a challenging dataset. We also show that a mixture of linear classifiers can be used to effectively model part appearance, improving accuracy while simultaneously reducing computational expense by a factor of 10. In future work we will investigate more flexible models which can capture the full range of human pose and appearance given the availability of large training sets.

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] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *Proc. ECCV*, 2010.

[4] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3D human pose annotations. In *Proc. ICCV*, 2009.

[5] 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.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, 2005.

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

[8] M. Eichner and V. Ferrari. Better appearance models for pictorial structures. In *Proc. BMVC*, 2009.

[9] M. Eichner and V. Ferrari. We are family: Joint pose estimation of multiple persons. In *Proc. ECCV*, 2010.

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2), 2010.

[11] P. Felzenszwalb and D. Huttenlocher. Efficient matching of pictorial structures. In *Proc. CVPR*, 2000.

[12] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008.

[13] V. Ferrari, M. Marin-Jimenez, and A. Zisserman. Progressive search space reduction for human pose estimation. In *Proc. CVPR*, 2008.

[14] Z. Fu and A. Robles-Kelly. On mixtures of linear SVMs for nonlinear classification. In *SSPR & SPR*, 2008.

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

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

[17] S. Johnson and M. Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. BMVC*, 2010.

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

[19] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: tracking people by finding stylized poses. In *Proc. CVPR*, 2005.

[20] D. Ramanan and C. Sminchisescu. Training deformable models for localization. In *Proc. CVPR*, 2006.

[21] B. Russell, A. Torralba, K. Murphy, and W. Freeman. LabelMe: A database and web-based tool for image annotation. *IJCV*, 77(1–3), 2008.

[22] V. Singh, R. Nevatia, and C. Huang. Efficient inference with multiple heterogeneous part detectors for human pose estimation. In *Proc. CVPR*, 2010.

[23] A. Sorokin and D. Forsyth. Utility data annotation with amazon mechanical turk. In *First IEEE Workshop on Internet Vision, CVPR*, 2008.

[24] D. Tran and D. Forsyth. Improved human parsing with a full relational model. In *Proc. ECCV*, 2010.

[25] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In *Proc. NIPS*, 2005.

[26] Y. Wang and G. Mori. Multiple tree models for occlusion and spatial constraints in human pose estimation. In *Proc. ECCV*, 2008.

[27] B. Yau and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activites. In *Proc. CVPR*, 2010.