Learning Human Pose in Crowd

Shaogang Gong
School of EECS
Queen Mary Univ. of London
sgg@eecs.qmul.ac.uk

Tao Xiang
School of EECS
Queen Mary Univ. of London
txiang@eecs.qmul.ac.uk

Somboon Hongeng
School of EECS
Queen Mary Univ. of London
sxh@eecs.qmul.ac.uk

ABSTRACT

In a crowded public space, body and head pose can provide useful information for understanding human behaviours and intentions. In this paper, we propose a novel framework for locating people and inferring their body and head poses. Human detection and pose estimation are two closely related problems but have been tackled independently in previous studies. In this work, we advocate joint detection and recognition of both head and body poses. Our framework is based on learning an ensemble of pose-sensitive human body models whose outputs provide a new representation for poses. To avoid tedious and inconsistent manual annotation for learning pose-sensitive models, we formulate a semi-supervised learning method for model training which bootstraps an initial model using a small set of labelled data, and subsequently improves the model iteratively by data mining from a large unlabelled dataset. Experiments using data from a busy underground station demonstrate that the proposed method significantly outperforms a state-of-the-art person detector and is able to yield extremely accurate head and body pose estimation in crowded public spaces.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object recognition*

General Terms

Algorithms, Performance

1. INTRODUCTION

One of the key objectives of automated video surveillance is to perform human behaviour profiling and monitoring in busy public space. Body and head orientations (referred to as poses in this paper) can provide useful information for understanding human behaviours and intentions in videos captured from crowded public spaces (see Figure 1). In general, video data of a public space is often poor for either reliable analysis of facial expression due to low resolution (long distance from cameras) or consistent recognition of body action and body configuration due to severe occlusions. Under such viewing conditions, body and head pose profiles over space and time provide more reliable measurements for inferring and interpreting individual's intensions. For instance, a person facing the train with his head turning towards the camera direction may indicate he is checking the information board to make sure the train is the right one to board; two people slightly tilting their heads towards each other may suggest they are chatting. Furthermore, body pose and head pose of the same person often differ although they are intrinsically correlated due to physiological and behavioural constraints. Both need to be visually estimated and analysed collaboratively for better accuracy and robustness.

There is a large amount of research on human detection and estimation of body and head poses. However, most existing techniques simply do not work well with low-quality CCTV videos of crowded public scenes [11, 3, 10]. For those which were aimed for this purpose, the problems of human detection and pose recognition are treated as separate problems solved independently. For pose estimation, it is assumed that detection was made available elsewhere [13, 1], whilst detection is based on building pose-specific detectors. During detection, a model requires to first recognise the pose before selecting a pose-specific detector for detection [14, 16]. This clearly presents a chicken-and-egg scenario if a holistic model is to be developed – without locating people accurately a pose classifier will give false pose estimation. Without estimating pose accurately, detection becomes extremely difficult. Our model is designed to tackle both problems simultaneously in a holistic framework. Despite a few recent attempts on simultaneous object detection and pose estimation [15, 2], to the best of our knowledge, none of the existing methods is designed for joint human detection and estimation of body and head poses simultaneously.

Figure 1: Examples of crowded public scenes.

To that end, we propose an ensemble of pose-sensitive body models as a holistic detection model whilst individual detection outputs from the ensemble are also used for

MPVA'10, October 29, 2010, Firenze, Italy.

pose estimation collectively, thus solving the two problems jointly and simultaneously. Each pose-sensitive model in our ensemble model is based on a recently proposed discriminatively trained deformable part-based model (DPM) [7], with a number of important modifications to make it sensitive to human pose and more robust to noise, scale changes, and occlusions. Learning this ensemble of pose-sensitive human body models requires manual annotation of both image locations and body poses of people in a large amount of video image frames. To address this problem, we propose a novel semi-supervised learning method which bootstraps an initial model using a small set of labelled (annotated) data, and then subsequently improves the model iteratively by automatic data mining (self-sampling) from a large unlabelled dataset. This method is particularly suitable for a busy public scene where there are abundant image frames with large number of people of different body and head poses. Manual annotation of such data is both tedious and inconsistent, which can lead easily to biased data sampling for model training resulting in sub-optimal detection and pose estimation.

To validate our approach, extensive experiments are carried out using data from both the PASCAL VOC2008 dataset [6] and the i-LIDS dataset [9] featured with two busy scenes in an underground station (see Figure 1). Our results suggest that (1) on person detection, our ensemble of pose-sensitive models significantly outperform the state-of-the art DPM model of [7], particularly when automatic mining (self-sampling) of unlabelled data are utilised for model training using the proposed semi-supervised learning method. (2) Both body and head poses can be accurately estimated using our method under very challenging conditions.

2. ENSEMBLE OF POSE-SENSITIVE MODELS

2.1 Multi-scale Deformable Part Models

Let us first briefly describe the multi-scale deformable part models (DPM) proposed by [7] which is used as a base component of our model, before we detail some important modifications to DPM required for formulating an ensemble model. A DPM model of an object with n parts is defined as a $(n+2)$-tuple $\beta = (F_0, P_1, \ldots, P_n, b)$, where F_0 is a coarse-scale global "root" filter covering an entire object, P_i is a model for the i-th part and b is a bias term. Each part model P_i is defined by a 3-tuple (F_i, v_i, d_i) where F_i is a fine-scale "part" filter computed at twice the resolution of the root filter. Term v_i is a vector specifying an anchor position for part i relative to the root position, and d_i is a vector specifying the coefficients of a function defining a deformation cost for the misalignment of the part. The spatial distribution of parts is thus specified by both v_i and d_i. Both root and part filters are applied to a feature map extracted from image. A variant of Histogram of Gradient (HOG) features are employed [5], which have shown to be robust against noise, scale changes and occlusions for object detection. Figure 2 shows an example of DPM model and HOG features.

For learning the parameters β of a DPM, images containing the object of interest only need to be annotated in the form of bounding boxes indicating object locations. Since no annotation at the part level is required, the model is trained by multi-instance learning using a latent-SVM (LSVM). For detection, to deal with object scale variations, a test image is repeatedly smoothed and subsampled to form an image pyramid and features are extracted at each pyramid level to form a feature pyramid. At each level of the pyramid, for a hypothesis of an object at location $z = (p_0, \ldots, p_n)$ (where $p_i = (x_i, y_i, l_i)$ specifies the position and the pyramid level of the i-th filter), a detection score is computed as:

$$sc(p_0, \ldots, p_n) = \sum_{i=0}^{n} F_i \cdot \phi(H, p_i) - \sum_{i=1}^{n} d_i \cdot \phi_d(dx_i, dy_i) + b \quad (1)$$

where $\phi_d(dx_i, dy_i) = (dx_i, dy_i, dx_i^2, dy_i^2)$ are deformation features computed from the displacement of the i-th part. A detection window is scanned over the feature pyramid and at each root location p_0, the detection score is computed according to the best possible placement of the parts:

$$sc(p_0) = \max_{p1,\ldots,p_n} sc(p_0, \ldots, p_n). \quad (2)$$

Objects are then detected by thresholding these scores.

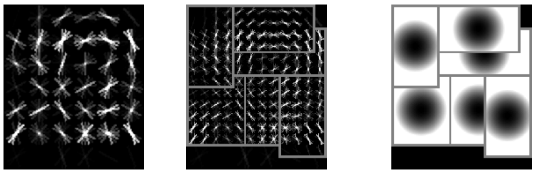

(a) Root-filter (b) Part-filters (c) Spatial models

Figure 2: An example deformable part-based model for the head-shoulder region of a right-facing person. (a) The root filter specifies a global detection window. (b) Six part-filters for finer detection of parts. (c) Spatial distributions of the 6 part-filters.

2.2 Ensemble of Pose-Sensitive Mixture Models

The DPM model described above was designed as a generic object detection model. To make it more suitable for person detection in a crowded scene and more importantly, to make it pose sensitive so that detection and pose estimation can be solved jointly, the following key modifications are introduced:

1. In busy scenes like those in Figure 1 it is noted that detection score of a DPM-based detector is biased towards higher levels in the feature pyramid. As a result, most detections are obtained from those higher-resolution feature maps which make it sensitive to image noise and cluttered background. In order to offset the bias, we normalize the matching scores of the root and part-filters at level l_i in Eq. (1) by the size of the feature map at that level. Our experiments in Section 5 demonstrate that this normalization consistently improves the detection results.

2. Both the root and part filters are symmetric models in the original DPM model. This is not a problem for modelling the front and back view of a person but it cannot capture poses from other orientations. Asymmetric models are thus adopted in our model.

3. In [7], a mixture of DPMs is proposed to deal with the pose variation of an object. Each mixture component is learned using object samples of a certain aspect ratio. Their mixture model is thus for modelling the variation of object appearances due to change in the aspect-ratios of the bounding boxes (e.g., longish for a full body vs. squarish for a half body), rather than variations in orientation. To make the model pose-sensitive, we formulate an ensemble of pose-sensitive DPM mixtures as described next.

Suppose we are considering V possible body poses, an ensemble model with V pose-sensitive DPM mixtures is built and denoted as a V-tuple $EM = (M^1, \ldots, M^v, \ldots, M^V)$, where M^v is a pose-sensitive DPM mixture model for the v-th pose. To learn M^v, a set of training images belonging to the v-th pose are required. The issue of how to obtain these pose-specific training samples without manual annotation is addressed in Section 3. Figure 3 shows an ensemble of four DPM mixture models for upper-body detection. The four mixture models correspond to the four classes of body poses that we want to classify: frontal, rear, left, and right. For simplicity, we set the number of components of each mixture model to two, which roughly correspond to two types of bounding box aspect ratios: one covers from the top of the head to the upper-torso and the other from the top of the head to the upper-chest.

2.3 Joint Human Body Detection and Pose Estimation

Now given an ensemble model with V DPM mixture models and a test image, at location p_0, V detection scores can be computed using the V mixture models respectively which forms a score vector $S(p_0) = (sc^1(p_0), \ldots, sc^V(p_0))$, where $sc^v(p_0)$ is the score of the mixture model M^v computed by Eq. (2) using the mixture component that gives the highest score. $sc^v(p_0)$ is normalised to have a value range between -1 and 1 via logistic regression. A straightforward way of performing detection and pose estimation is to take the maximum score as the final score of the ensemble and recognise the pose according to which DPM model gives that maximum score. However, due to the large cross-pose similarity a person image often yields large responses from multiple mixture models which makes the pose estimation unreliable. To overcome this problem, we treat the score vector as a new representation of both human and its pose and train another discriminative classifier for both detection and pose estimations. Specifically, we use a multi-class SVM with radial basis kernels (MC-SVM) as an ensemble classifier. The MC-SVM is built using one-against-rest strategy. It takes the score vector $S(p_0)$ as input and produces $V + 1$ outputs, one for each body pose and one for the non-object class. With this ensemble classifier, a joint human body detection and pose classification can thus be performed in one shot.

3. SEMI-SUPERVISED LEARNING OF ENSEMBLE MODEL

We now describe a semi-supervised learning method for training the ensemble model $EM = (M^1, \ldots, M^v, \ldots, M^V)$ using a small set of annotated image (I^a) and a large set of unannotated images (I^u). First, we construct an initial training set D_0 consisting of only annotated samples from I^a. D_0 is used to learn an initial model EM_0 using LSVM, which scores all the annotated training samples to form the score vectors which are then used to learn the ensemble detector and pose classifier using a SVM, denoted as EM_{svm_0}. Next, EM_{svm_0} is applied to both I^a and I^u to obtain a set of detections. Based on these detection, a set of 'good' training samples, both positive and negative are selected (supervised from I^a and unsupervised from I^u) and added to the initial training set D_0 to form a larger training set and learn a stronger model. Finally, we run the updated ensemble model on a validation test set and compute an average precision λ. The whole "detection-mining-updating" procedure continues iteratively and in each iteration, the ensemble model is retrained and used to mine more 'good' samples, until λ stops improving; we then have the final model for body detection and pose estimation.

Let us now describe in detail how to mine 'good' samples from I^a and I^u for learning a stronger model in the next iteration of semi-supervised learning. For clarity, we drop the iteration index and the superscript v for different pose-sensitive DPM mixtures. Let $D_0 = \{D^+, D^-\}$ be a set of manually labelled samples containing the positive samples D^+ and negative samples D^- from the manually annotated images I^a. Using D_0, an initial detector and pose estimator EM_{svm_0} is obtained and all positive samples in D^+ are scored with a normalised score ranging between -1 and 1, and ranked according to their scores. A detection threshold τ_d is set so that the top 95% ranked positive samples have a score greater than τ_d. This threshold is then used to obtain a set of detections. Based on these detections, we mine four types of samples to add to D_0:

- $\mathbf{D_a^+}$: positive samples from I^a – detections that overlaps with a ground truth bounding box
- $\mathbf{D_a^-}$: hard negative samples from I^a – detections that do not overlap with a ground truth bounding box
- $\mathbf{D_u^+}$: positive samples from I^u – detections with a score greater than τ_p
- $\mathbf{D_u^-}$: negative samples from I^u – candidate windows with a score less than τ_d

The way we mine $\mathbf{D_a^+}$ and $\mathbf{D_a^-}$ is in line with most existing methods for learning a discriminative detector [5, 7]. Mining $\mathbf{D_u^+}$ and $\mathbf{D_u^-}$ is new as no one has attempted training an object detector using unlabelled data. Obviously without knowing where the objects are and even whether there are any objects in each image in I^u, one must take a much more conservative approach in mining both positive and negative samples to avoid model drifting. To that end, we set a very high value of τ_p (0.9 in this work) to make sure that all unsupervised mined positive samples indeed contain the object of interest. As for negative samples from I^u, $\mathbf{D_u^-}$ would correspond to candidate windows that either contain no human or only part of human body. Note that among the large number of candidate windows in I^u (thousands in each frame), most of them will fall into $\mathbf{D_u^-}$. To select the hard negatives (i.e. those the current model is likely to produce a false positive), we rank all scores in $\mathbf{D_u^-}$ and filter out those with lowest scores (corresponding to easy negatives).

4. HEAD DETECTION AND POSE ESTIMATION

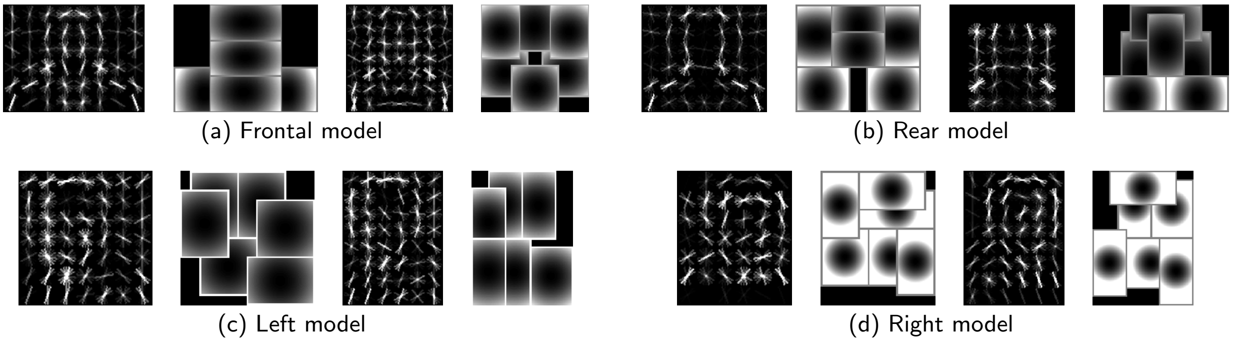


Figure 3: Mixture models for the frontal (a), rear (b), left (c) and right (d) poses of the upper body. Each mixture model has two DPM components. Compared to Figure 2, we show only the root filter and the spatial models of parts of each DPM.

The method described in Section 2 can be applied to head detection and pose estimation provided that images of head regions of different poses are available. However, in a typical busy public scene captured by surveillance camera, head regions can be as small as 10 × 20 pixels (see Figure 1). HOG features thus become less discriminative due to the low resolution. In this section a method is formulated which relies on the body detection model output for head detection and a more robust feature for head pose representation.

Estimating Head Location – It can be seen in Figure 3 that although there is not a single part of DPM that correspond accurately the location of head, the locations of the automatically inferred 6 parts do contain information that, although very coarse, can be used for head detection. Specifically the problem of head localisation given the 6 detected body part location is treated as a regression problem and solved through Canonical Correlation Analysis (CCA) [8]. Suppose h (4-dimensional) and b (24-dimensional) are two multivariate random variables representing the bounding boxes of a head and six body parts inferred from our ensemble of DPM models. Assume we have a set of K observations of b_k and h_k, the former being obtained automatically the latter manually, we aim to learn a regression matrix $R_{(h|b)}$ so that during testing given a detection of human body and its 6 parts, its head location can be estimated. CCA solves this problem by finding basis vectors $\mathbf{w}_h$ and $\mathbf{w}_b$ for two sets of variables h and b such that the correlation between the projections of the variables onto these basis vectors are mutually maximised. We refer to [8] for an algorithm for deriving $\mathbf{w}_h$ and $\mathbf{w}_b$ from the covariance matrix constructed from h_k and b_k. Let H be a $4 \times K$ matrix whose column k is h_k, B be a $24 \times K$ matrix whose column k is b_k, and W_b be a matrix whose columns are four basis vectors from $\mathbf{w}_b$ that correspond to the four largest canonical correlations. Given W_b, B, H and the parts vector b from a human body detection in a test image, we can compute a regression matrix $R_{(h|b)}$ and give an estimate of h as follows:

$$\begin{aligned} R_{(h|b)} &= HB^T W_b (W_b^T B B^T W_b)^{-1} \\ h &= R_{(h|b)} W_b^T b \end{aligned} \tag{3}$$

Head Pose Estimation – The head regions that we detect are often small, unclear, poorly illuminated and contain various backgrounds. Under these conditions, most existing head pose estimation methods [12] cannot be applied. We thus adopt a head feature map proposed in [13] which captures the skin and hair region distribution for pose classification. To compute a feature map, we only need to compute similarity distance maps between the head image and the mean head images for different poses (shown in Figure 4(a)). Pixel i of the feature map is then computed as the maximum of pixels "i" of all mean pose maps. Figure 4(b) shows a head image and its features map. It can be seen that this "image-to-feature map" transformation performs implicit segmentation so that the background pixels are removed, the skin regions represented as dark areas and the hair regions represented as bright areas. A MC-SVM classifier is then trained using these feature maps for pose classification.

5. EXPERIMENTS AND DISCUSSIONS

Datasets – We evaluated our method using two challenging public datasets PASCAL VOC2008 [6] and i-LIDS [9] and followed the VOC protocol for detection evaluation. *VOC2008 person dataset:* There are 2002 images with 4168 people in the training and validation sets. The test set is being used for the VOC2010 challenge and not available; we thus used the Taster set for testing which contains 245 images with 367 people. *i-LIDS dataset:* the i-LIDS database contains extensive CCTV footages of two busy underground station scenes (see Figure 1). These videos were captured at 25 fps, 720×576 resolution, under uncontrolled conditions. The scenes are much more crowded than those in VOC2008 dataset (on average dozens of people per image in i-LIDS vs. 2 in VOC), and thus are more challenging. Due to severe occlusions, upper-bodies are visible mostly from chest to head. Head regions are small (10 × 20 to 40 × 60 pixels). We created three datasets from i-LIDS for training, validation and testing. The labelled training set consists of 150 frames. An annotated sample of a human consists of an upper-body's bounding box, a head's bounding box, body pose and head pose. There were 1218 frontal, 852 left, 852 right and 450 rear samples in the labelled training set and 100 background images were used as negative training images. Similarly a small validation set of 100 frames has been used to determine when to terminate our semi-supervised learning interactions. The unlabelled training set contains 10,142 un-annotated frames. The test set contains 100 annotated frames with 325 frontal, 114 left, 114 right and 152 rear samples.

Detection on VOC2008 – the dataset contains mostly the frontal views of full-body and half-body humans. It is thus not suitable for pose estimation evaluation and was only

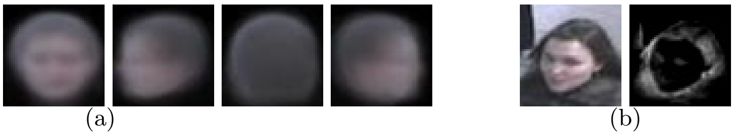
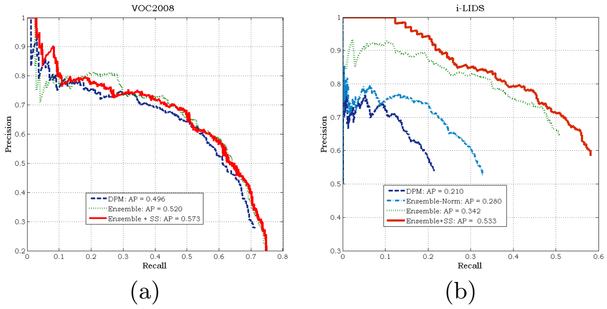

(a) (b)

Figure 4: (a) Mean-images: frontal, left, rear, right. (b) Frontal face and the feature map.

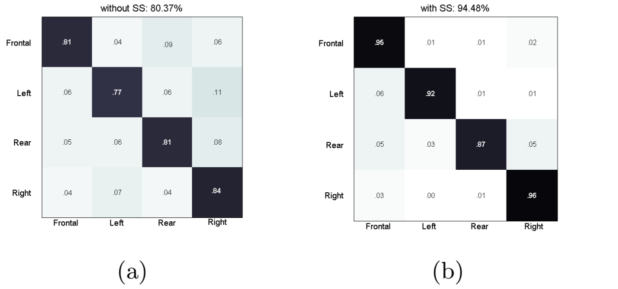


(a) (b)

Figure 5: Performance comparison on person detection.

used for detection. Without pose-sensitive DPM mixture models learned from samples of different poses, our ensemble model contains two mixture components corresponding to full body and half body. It is thus identical to the original DPM model in [7] apart from the modifications on asymmetric model and score normalisation (see Section 2.2). Our experiments on VOC2008 were designed to demonstrate the effectiveness of the modifications and the semi-supervised learning method. In particular, for semi-supervised learning of our ensemble model, we used 80% of the training and validation sets as labelled data, and the rest 20% as unlabelled. In other words, the same amount of data were used for learning but with less annotation. Figure 5(a) shows the detection results represented as precision-recall curve with averaged precision (AP) value marked. It shows that with the introduced modifications, the detection performance measured using AP is improved from 0.496 to 0.520. With the semi-supervised learning (SS), it is further increased to 0.573, a 16% increase over the original DPM model which won the VOC challenge on person detection in 2008 and 2009. This suggests that even with the same amount of training data but with less annotation, our semi-supervised learning method can mine better training samples in an unsupervised manner resulting in better detection model. This is not surprising as recent research [4] has also shown that for labelled data, manual annotation is often biased and sub-optimal and can lead to inferior learning of a detector.

Detection on i-LIDS – Figure 3 shows the ensemble of four pose-sensitive mixture models trained using our i-LIDS training dataset based on our semi-supervised learning algorithm. The detection result is shown in Figure 5(b). As expected the DPM model performs poorer on this more challenging dataset, yielding an AP of 0.210. With our ensemble model trained using only the labelled training and validation sets (Ensemble), the AP is improved to 0.342. This demonstrates the importance of learning pose-sensitive models. Although additional annotation is required on the pose of each labelled positive samples, the over 50% increase of detection performance makes our ensemble model a much more attractive solution for person detection in crowded scenes. To examine the effectivenss of our score normalisation, we also implemented our ensemble model without score normalisation (Ensemble-Norm). This gives a AP value of 0.280 which indicates that it is crucial to perform score normalisation in a busy scene (much so than for the VOC2008 data). Finally, Figure 5(b) shows that a much stronger model is obtained by semi-supervised learning using the large unlabelled dataset (Ensemble+SS) with AP of 0.533. This again validates the effectiveness of the proposed semi-supervised learning method. This increase of performance is much larger than that achieved on VOC2008 due to the much larger size of the unlabelled training set. Examples of detection using our semi-supervised trained ensemble model can be seen in Figure 6. It is clear that due to the presence of large number of people in the far end of the camera view, it is extremely difficult to achieve a high recall rate in these scenes. Nevertheless, our detector can detect most people that are close to camera even with severe occlusions.


without SS: 80.37%

	Frontal	Left	Rear	Right
Frontal	.81	.04	.09	.06
Left	.06	.77	.06	.11
Rear	.05	.06	.81	.08
Right	.04	.07	.04	.84

with SS: 94.48%

	Frontal	Left	Rear	Right
Frontal	.95	.01	.01	.02
Left	.06	.92	.01	.01
Rear	.05	.03	.87	.05
Right	.03	.00	.01	.96

(a) (b)

Figure 8: i-LIDS head pose classification confusion matrices when detection precision is 0.80. (a):Trained using only labelled data, pose estimation using SVM. (b) Trained using semi-supervised learning, pose estimation using SVM.

Pose estimation on i-LIDS – The body pose detection result is shown in Figure 7. It can be seen that with a detection rate of 0.80 (recall rate of 0.38), the estimation of the poses of these detected bodies is very accurate with an average classification rate of 94.4% (Figure 7(c)). This, compared with the 80.37% classification rate achieved when our ensemble model is trained using only the labelled dataset (Figure 7(b)), validates the usefulness of semi-supervised learning of a strong model. We also tested pose estimation based on the maximum score of ensemble members, which is the typical method used for pose estimation [14, 16]. The result in Figure 7(a) (55.59% compared to 80.37% in (b) when trained using labelled data) suggests that due to the ambiguities between different body poses, learning another discriminative pose classifier using the ensemble model outputs can significantly improve the pose estimation accuracy. Figure 8 shows the head pose estimation results, which again

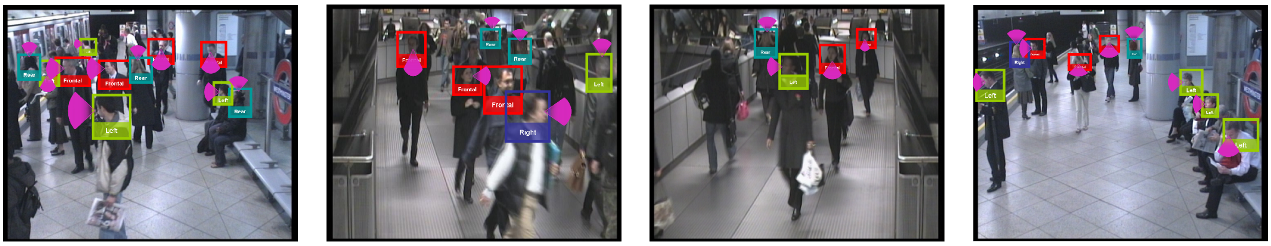
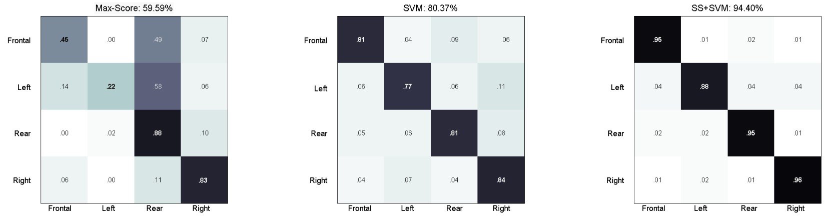


Figure 6: Examples of upper-body detection and pose estimation for body and head when the precision and recall rates are 0.802 and 0.383 respectively. Bounding box colour and text indicate body pose and magenta dials indicate head pose.



(a) (b) (c)

Figure 7: i-LIDS body pose classification confusion matrices when detection precision is 0.80. (a):trained using only labelled data, pose estimation using maximum score. (b) trained using only labelled data, pose estimation using SVM. (c) trained using semi-supervised learning, pose estimation using SVM.

shows that with semi-supervised learning, the performance of head pose estimation (Figure 8(b)) is superior to that of the same model learned using only labelled data (Figure 8(a)). This is because with semi-supervised learning, our ensemble model becomes much stronger, leading to better body and part localisation, which in turn yields better localisation of head regions for pose estimation. Examples of pose estimation results can be seen in Figure 6.

Acknowledgement

We shall acknowledge Javier Orozco's contribution towards experiments from which some of the results are presented in this paper.

6. REFERENCES

[1] J. Aghajanian, J. Warrell, S. Prince, P. Li, J. Rohn, and B. Baum. Patch-basedwithin-object classification. In *ICCV*, 2009.

[2] K. Ali, F. Fleure, D. Hasler, and P. Fua. Joint pose estimator and feature learning for object detection. In *ICCV*, 2009.

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021, Miami, USA, June 2009.

[4] M. Blaschko and C. Lampert. Learning to localize objects with structured output regression. In *ECCV*, 2008.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 2, pages 886–893, 2005.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results.

[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 2009.

[8] D. R. Hardoon, S. Szedmak, O. Szedmak, and J. Shawe-taylor. Canonical correlation analysis; an overview with application to learning methods. Technical report, 2007.

[9] HOSDB. Imagery library for intelligent detection systems (i-lids). In *IEEE Conf. on Crime and Security*, 2006.

[10] H. Jiang. Human pose estimation using consistent max-covering. In *ICCV*, 2009.

[11] M. Lee and I. Cohen. Human upper body pose estimation in static images. In *ECCV*, volume 3022, pages 126–138, Cambridge, MA, 2004.

[12] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision: A survey. *PAMI*, 31(4), 2009.

[13] J. Orozco, S. Gong, and T. Xiang. Head pose classification in crowded scenes. In *BMVC*, London, UK, September 2009.

[14] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In *CVPR*, 2009.

[15] Q. Yuan, A. Thangali, V. Ablavsky, and S. Sclaroff. Multiplicative kernels: Object detection, segmentation and pose estimation. In *CVPR*, 2008.

[16] J. Zhang, S. Zhou, L. McMillan, and D. Comaniciu. Joint real-time object detection and pose estimation using probabilistic boosting network. In *CVPR*, 2007.