Head Pose Classification in Crowded Scenes

Javier Orozco
http://www.dcs.qmul.ac.uk/~orozco
Shaogang Gong
http://www.dcs.qmul.ac.uk/~sgg
Tao Xiang
http://www.dcs.qmul.ac.uk/~txiang

Queen Mary Vision Laboratory
School of EECS
Queen Mary University of London
London E1 4NS, UK

Abstract

We propose a novel technique for head pose classification in crowded public space under poor lighting and in low-resolution video images. Unlike previous approaches, we avoid the need for explicit segmentation of skin and hair regions from a head image and implicitly encode spatial information using a grid map for more robustness given low-resolution images. Specifically, a new head pose descriptor is formulated using similarity distance maps by indexing each pixel of a head image to the mean appearance templates of head images at different poses. These distance feature maps are then used to train a multi-class Support Vector Machine for pose classification. Our approach is evaluated against established techniques [3, 13, 14] using the i-LIDS underground scene dataset [9] under challenging lighting and viewing conditions. The results demonstrate that our model gives significant improvement in head pose estimation accuracy, with over 80% pose recognition rate against 32% from the best of existing models.

1 Introduction

Human head pose and gaze direction can provide useful information for the inference of person's intent and behaviour. The topic has traditionally been studied for expression and face recognition, and human computer interaction [10]. However, most existing techniques rely upon medium to high resolution images captured under well controlled conditions from a fairly close distance [4, 8, 12, 15]. Given high resolution images, most existing techniques deploy extensive feature extraction to capture detailed head/facial shape and texture information. Alternatively, Tian et al. [16] considered the problem of analysing coarse head pose in images captured by wide-angle overhead cameras where silhouette detection is used as the basis for head shape representation. However, this approach relies on accurate subtraction of head foreground region from the background which is not always feasible.

More recently, a few attempts have been made on head pose estimation in low-resolution images by treating the problem as a multi-class discrete pose classification problem in order to improve robustness. This is achieved by manually labelling head image textures for training different pose classifiers [2, 17, 18]. In particular, Robertson and Reid [13] proposed a combined skin and hair colour based appearance model using colour histograms for head pose classification given low resolution images. In their approach, 360° head pose in

panning angle is discretized into eight pose classes with 45° increment. Given background-foreground segmentation of an input image, pose classification is performed by matching the colour histogram of the probe image with those of eight skin-hair-colour appearance models using a probabilistic tree. They further combine the estimation of walking direction with head pose classification to stabilise head pose estimation. This approach relies critically upon good segmentation of the skin and hair texture regions of a head image.

However, images captured from most public space CCTV cameras are subject to very challenging viewing conditions and in low-resolution. Under such conditions, skin and hair textures of a head image are often not clearly distinctive in either intensity and chromaticity (see examples in Fig.1). This makes segmentation of skin and hair regions from a head image very difficult if not entirely impossible at times.

Figure 1: Typical head images extracted from the i-LIDS underground scene. They are of low-resolution and subject to significant directional lighting changes.

In this paper, we propose a novel approach to head pose classification in crowded public scenes using low-resolution images captured under challenging viewing conditions. In particular, we avoid the need for explicit segmentation of skin and hair regions from a head image. Spatial positional information is also utilised in our model representation. However, unlike previous techniques using shape explicitly [14], we implicitly encode spatial information using a grid map for more robustness given low-resolution images. Moreover, in order to cope with large degree of variations in the positions of pixels that correspond to skin and hair textures across different poses, and the non-uniform nature of their distributions in a head image (i.e. there is often no clean-cut separation between skin and hair textures at the pixel level), instead of using pixel appearance information directly [11], we propose a novel approach to construct feature vectors using similarity distance maps by indexing each pixel of a head image to the mean appearance templates of head images at different poses using KL divergence. These distance feature maps are then used to train a multi-class Support Vector Machine for pose classification. We demonstrate significant performance advantages of our representation compared to a state-of-the-art model [13] and other established techniques [3, 14] for head pose estimation in crowded public space under challenging viewing conditions captured by the UK Home Office i-LIDS dataset [9].

2 Framework

Pose Specific Mean Appearance Templates To construct a head appearance representation, segmented head images need be background whitened in order to minimise the effect of background pixels surrounding a head [13]. This is especially important for images from crowded public scenes. However, due to uncontrolled lighting causing significant changes in background, such pre-processing can be unstable and error-prone (see examples in Fig 2). To overcome this problem, we propose a different representational scheme as follows.

Figure 2: Poor image quality and uncontrolled lighting cause errors in background subtraction and whitening for skin segmentation as proposed by [13].

A set of head images are collected from the i-LIDS dataset to create a training set. All initial images are manually cropped and normalised in size (note that manual cropping is not needed for testing, see Sec. 3 on Dataset and Head Detection). Similar to [2, 13], this training set is labelled by pose and grouped into eight discrete pose classes $k45^\circ$ where $k = 1, ..., 8$ (see Fig. 3 (a)). However, different from [2, 13], we do not attempt to distinguish between skin and non skin pixels. Hence there is no need for texture labelling or segmentation and we avoid any assumption on hair and skin configuration within a head/face image.

Due to low image resolution, we assume that each head image pixel value in each pose class and RGB colour channel is a random variable that can be represented approximately by a single Gaussian distribution. Consequently, Head images of each pose class and RGB channel can then be represented by a multivariate normal distribution with Gaussian parameters being also images, i.e. the mean image $\mathbf{M}$. This metric contains the mean values of each pixel at their image location. For a given set of training images, we compute per pose class and RGB channel a shape-free mean appearance template, $\mathbf{M}^c$ (see Fig. 3 (b)).

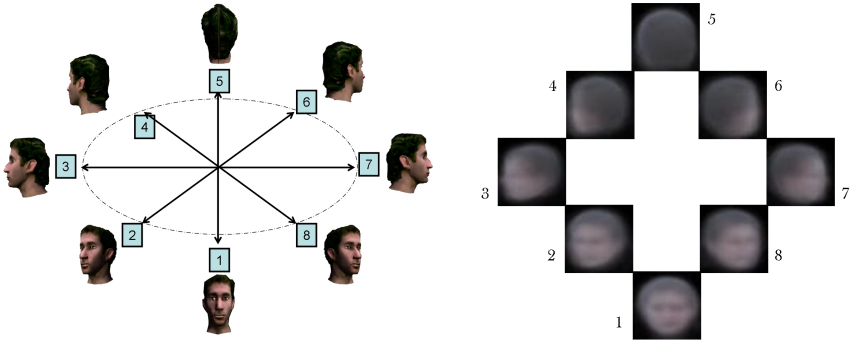


Figure 3: (a) Head pose of 360° in panning angles are quantised into eight discrete pose classes representing pose angles at 0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°. (b) The average (mean) head appearance templates for the eight pose classes.

Feature Descriptor by Similarity Distance For pose classification, it is critical to represent head appearance based on good separation of background, hair and skin/non-skin pixels. However, due to the highly non-uniform spread of those pixels, we avoid hard-labelling and categorisation of pixels and instead, we consider an indirect similarity distance measure based representation. The central idea is to compare each input image pixel to a set of mean appearance templates regardless pose, i.e. across all poses, given that the true pose of the given image is unknown. Therefore, for a given image $\mathbf{N}$, we compute a set of eight

weights $x_{i,j}$ for each pixel $n_{i,j}$. These weights measure a score of similarity between each pose class appearance mean template and the input image at each pixel. More precisely, given a mean template image $\mathbf{M}^c \in \Re^{axb}$ for each pose class and an input image $\mathbf{N} \in \Re^{axb}$, where $c = \{1,...,8\}$, their corresponding pixels are denoted as $m^c_{i,j}$ and $n_{i,j}$. Each pixel from the input image is profiled by exhaustive comparison to the corresponding pixels from each mean appearance template. To that end, we measure the *Kullback Leibler divergence* (*KL*) between the input image $\mathbf{N}$ and each pose class for every pixel of $\mathbf{M}^c$ in each of the three colour channels. Note that the standard *KL* is defined as $D_{KL}(p||q) = p\left(\log \frac{p}{q}\right)$ which measures the divergence between two $p.d.f$. Here for a single input image no distribution can be estimated for the pixel value at each pixel location. The standard *KL* formulation is thus not appropriate. What we measure instead is subtly different from the similarity between two distributions, and is referred to as *KL* coefficients (δ_{KL}) :

$$\delta_{KL}\left(m^c_{i,j}||n^c_{i,j}\right) \quad = \quad \max_{RGB}\left\{m^c_{i,j}\left(\log\frac{m^c_{i,j}}{n_{i,j}}\right)\right\} \tag{1}$$

where $n_{i,j}$ and $m^c_{i,j}$ are pixel intensity values from the same RGB colour channel. Here, we measure the disparity between actual pixel appearance of $\mathbf{N}$ and the expected values from each mean appearance template at each pixel position. Note that similar idea has been exploited for measuring the similarity of two random variables in the domain of prediction theory of classification [1].

Since we aim to keep the topological independence of pixel variation, we construct a similarity distance weighting map as a feature descriptor (2D matrix) containing the maximum divergence coefficients between each pose class and $\mathbf{N}$ at each pixel location:

$$x_{i,j} = \max_{c}\left\{\delta_{KL}(m^c_{i,j}||n^c_{i,j})\right\}, \text{ and } c = \{1,...,8\} \tag{2}$$

Thus, $x_{i,j}$ contains the maximum coefficients from all 8 classes and 3 colour channels at each pixel position. We impose an additional constraint so that $\delta_{KL}(m^c_{i,j}||n^c_{i,j}) = 0$ when $n^c_{i,j} \geq m^c_{i.j}$. This effectively removes those divergent pixels deem to be background. Fig. 4 shows some examples of the extracted feature descriptors for a range of input head images. It is evident that by exhaustively projecting an input image to all eight templates and measuring their similarities by δ_{KL} at each pixel position, the proposed feature descriptor effective separates hair, skin and background at the pixel-level. In particular, the hair region is represented by high values whilst the face region yields low values in the descriptor. The robust head modelling is achieved even though textures' distributions are not modelled explicitly.

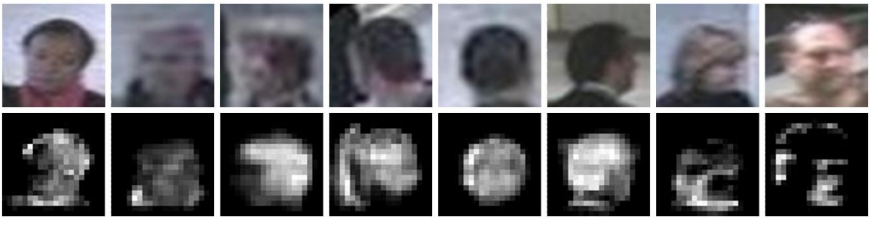


Figure 4: Examples of head image feature descriptors (bottom row) constructed from input images (top row) by selecting the maximum KL divergence δ_{KL} between each input image and multiple pose mean templates for each pixel position.

Head Pose Estimation by Multi-class SVM We shall now describe the use of this descriptor as feature representation for pose classification. Instead of classifying head pose by comparing image appearances with those of appearance template models [3, 14], we compare similarity distance maps between images and models at the pixel-level and shall demonstrate its significant advantage in recognition performance when input images are of low quality and subject to significant lighting variations and possible occlusion (see Sec. 3).

In order to classify any input image by eight discrete head poses, we apply a Multi-class Support Vector Machine (SVM). We build a model where the i-th SVM constructs a hyperplane between the class i-th and the $C-1$ remaining classes. Pose classification is determined by a majority vote among all eight classifiers. More specifically, we adopt a *one-against-rest* SVM strategy [7] using a polynomial kernel with the objective of finding a hyperplane capable of separating one pose class from the rest. Suppose $\mathbf{X}_c$ denotes the training samples, $\mathbf{Y}_c \in \{1,-1\}$ denote the corresponding labels, and α_{c_i} are the *Lagrange* coefficients determining class boundaries, the hyperplane $f(\mathbf{X}_c)$ for classification is:

$$f(\mathbf{X}_c) = \mathbf{W}_c^T \mathbf{X}_c + b_c \geq 1 \tag{3}$$

$$\mathbf{w}_c = \sum_{i=1}^{||c||} \alpha_{c_i} y_{c_i} \mathbf{x}_{c_i} \tag{4}$$

where b_c is the margin bound between the hyperplane and the Support Vectors (SVs), which are all $\mathbf{x}_{c_i}$ vectors with an α_{c_i} greater than zero. A SVM with fewer SVs and lower α_{c_i} values has better classification power with more generalisation capability. We shall demonstrate in our experiments that the proposed similarity distance feature descriptor enables a SVM learning machine to be constructed with much fewer SVs compared to other existing techniques (see Fig. 9).

Considerations on Similarity Metrics All similarity measures should be compared according to their independent coefficients between two pixel values, since additional scaling as integration along the image implies an image comparison and a pixel correlation. Kullback Leibler Divergence or the alternative Information Gain Measure provides several advantages over other metrics for similarity measure. Bhattacharyya coefficients measure the orientation between two distributions, which requires a holistic normalization on the whole image, which correlates the pixels variations. An Euclidean distance on the other hand measures only intensity variation spatially without taking into account the distribution estimate on pixel variations overtime. Assuming that the values at each pixel location across the training set follow a Gaussian distribution, one could aim to compute the join distributions of RGB channels for each image pixel at its image location under all pose variations. One can then employ the Mahalanobis distance for density weighted clustering of each pixel into hair, skin and background. However, it is unrealistic to assume that these models can be sufficiently well-estimated due to the diversity in pixel variations and difficulties in manually labelling pixels into different types of texture/appearance for hair, skin and background. The pseudo-Bhattacharyya coefficients was also employed by [13] to compare mean appearance models and image pixels, $w_{i,j} = \sqrt{\frac{m^c_{i,j}}{n_{i,j}}}$. Although the pseudo-Bhattacharyya measure has similarity to the KL-divergence, the later can cope with large non-linear variations due to its logarithmic function whilst the former is linear.

3 Experiments

Dataset and Head Detection We used the i-LIDS [9] underground scene dataset for all our experiments. The dataset consists of extensive CCTV footages of a busy underground scene captured under challenging lighting and viewing conditions. Our video data are from two underground stations with video frame size of 640×480 recorded at 25 fps (see examples in Fig 1). Typically the head image size varies from 40×60 to 10×20 pixels depending on the distance to camera. They were normalised to a size of 20×20. These scenes were crowded most of the time with many people present at any given time. People were often under some degree of occlusion and exhibited large head pose variations. Appearance variation of people includes beard faces, bold heads, light and dark hair and skin colours, all of which challenge modelling head/face image appearance with any assumption on clear-cut hair, skin and background segmentation.

For training our head pose classifiers, we randomly selected and manually cropped 800 head images with 100 images per pose class. During testing, given an image sequence we apply a background subtraction algorithm to highlight foreground areas in video frames where connected components analysis was performed to give candidate people search windows for head localisation. Head candidates were then obtained using a sliding window pedestrian detection model based on Histogram of Oriented Gradients (HOG) [6]. Two models were trained and hierarchically applied to both the whole and upper body region.

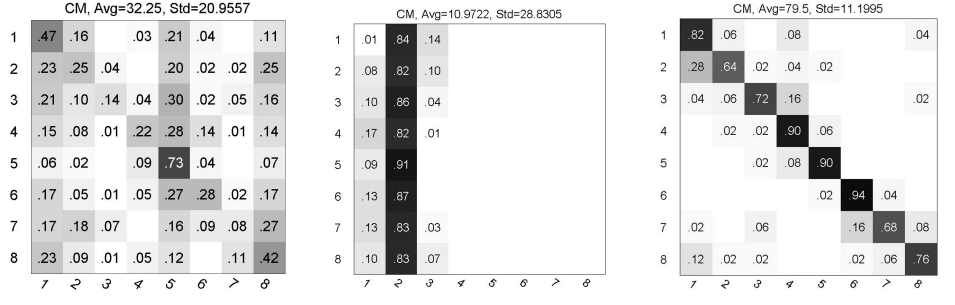
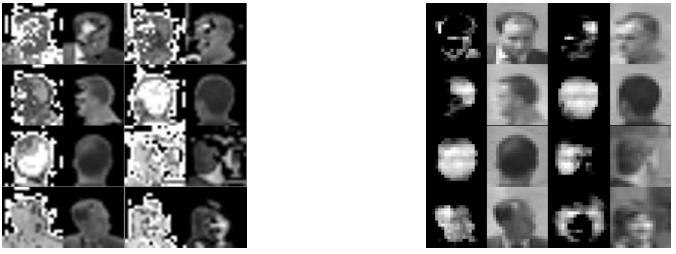


Figure 5: Confusion matrix on head pose classification using i-LIDS dataset. A $10-fold$ cross validation over 100 images per class provide the results given by: (a) Robertson and Reid [13], with average rate 32.25%. (b) Apparance models proposed by Beymer [3] and Sherrah et al. [14], with average rate 10.97%. (c) Our model, with average rate 79.5%.

Performance Comparison on Head Pose Classification To compare the performance of the proposed head pose classification model against other existing schemes by performing 10-$fold$ cross validation experiments. In particular we compared the effectiveness of using the proposed similarity distance feature descriptor with two existing alternative representations proposed in [13] and [3] respectively. For training a generic skin colour model in the representation of [13] , we extracted over 100,000 data points per colour from both the i-LIDS database and other publicly available face datasets. Matching of probe image colour histograms against models was performed using pseudo-Bhattacharyya coefficients. Fig. 5 shows that our model outperforms significantly existing techniques using explicit skin and non-skin histogram based modelling. It was evident that the skin colour from head images



(a) Skin and non-skin descriptors **(b)** Similarity distance feature escriptors

Figure 6: i-LIDS dataset difficult for extracting explicit skin and non-skin based models. (a) Skin and non-skin textures extracted using the model proposed by Robertson and Reid [13]. (b) Similarity distance feature maps extracted by our proposed model.

is severely degenerated using the i-LIDS dataset. Moreover, it is noted that for the representation in [13], poor quality in chromaticity of background, hair and skin textures often led to significant misidentification of pixels by explicit skin and non-skin colour modelling (see Fig. 6). Robertson and Reid [13] reported head pose classification rate of 80% on their original dataset but we could only obtain an average rate of 32% using the i-LIDS dataset (Fig. 5 (a)). This is largely because that many dark pixels in i-LIDS data correspond to the background rather than hair. Similarly many skin and non-skin pixels were easily confused.

Similarity Metrics Comparison We compared six different similarity metrics for constructing our similarity distance feature maps for pose classification on the i-LIDS dataset. Table 3 shows that *KL divergence* outperforms all other five alternative measures by obtaining the best pose classification average rate of 80% for $10-fold$ Cross Validation.

Head Pose	1	2	3	4	5	6	7	8	Mean Rate
Probability	0.36	0.34	0.38	0.41	0.64	0.36	0.40	0.43	42%
Euclidean	0.51	0.37	0.68	0.57	0.89	0.81	0.57	0.53	62%
Bhattacharyya	0.74	0.58	0.63	0.73	0.87	0.75	0.65	0.64	70%
Mahalanobis	0.72	0.68	0.66	0.73	0.81	0.82	0.67	0.74	73%
pseudo Bhattacharyya	0.76	0.61	0.71	0.81	0.87	0.82	0.70	0.74	75%
Kullback Leibler	**0.82**	**0.64**	**0.72**	**0.90**	**0.90**	**0.94**	**0.68**	**0.76**	**80%**

Table 1: Compare different similarity measures for head pose classification.

The Effect of Descriptor Size We evaluated the effect of choosing different head image size therefore the similarity distance feature map/descriptor size on head pose classification rate. This is to evaluate the robustness of our proposed descriptor against size variations. The dataset used in previous experiments were randomly rescaled from their original size of between 10×20 to 40×60 to some new sizes between 5×5 to 40×40 pixels. Fig. 7 shows that our similarity distance feature descriptor remains largely stable above image size of 5×5. If we disregard size 5×5, the rate remains stable at about $76\%\pm 4\%$.

Interpretation of the Learnt Support Vectors In order to give more in depth understanding on what the multi-class SVM has learnt and how different representations may affect the learning of the support vectors, Fig. 8 shows the learnt average positive and negative support vectors (SVs) for the eight different pose classes using our proposed similarity distance feature descriptors. It is evident that hair texture gives less variation at the top of a head. In contrast, skin texture is highly variable, suggesting its less robustness therefore poorer

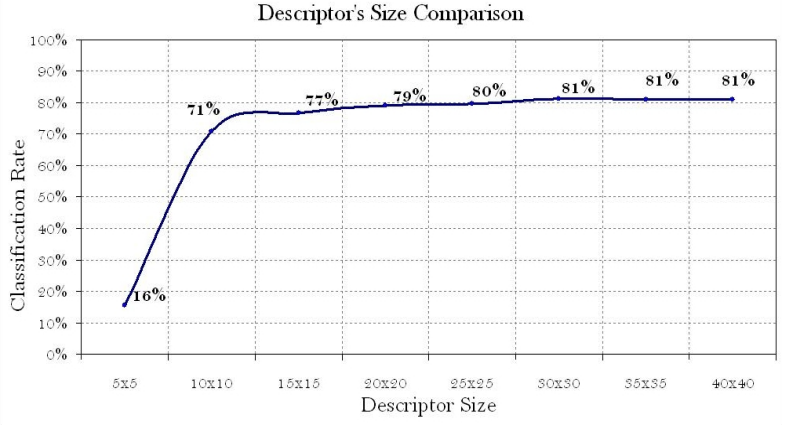


Figure 7: Head pose classification rate against image/feature map size.

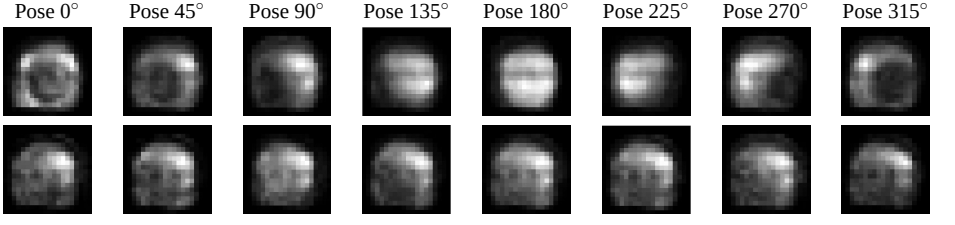
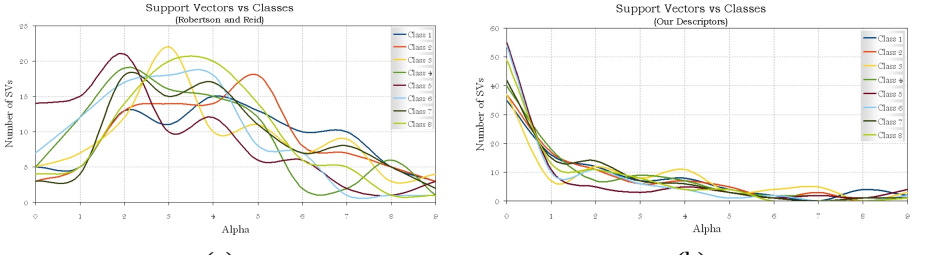


Figure 8: Learnt average support vectors (top row) and average biased feature vectors from the SVMs (bottom row) for the eight different pose classes.

ability in encoding information for head pose classification. Moreover, these SVs seem to be highly separable. For multi-class SVMs using polynomial kernels, fewer support vectors with smaller Lagrange coefficients give better classification performance in general [5]. Fig. 9 (a) shows that using Roberston and Reid descriptor resulted in a large number of SVs with poor separability of the decision boundaries. This is evident from many SVs with high α values. In contrast, our descriptors give much smaller number of highly separable SVs with low α values.



(a) (b)

Figure 9: (a) SVs of the SVM trained with Robertson and Raid descriptors see Fig. 5.(a). Similarly, (b) shows the SVs of the SVM trained with our descriptors see Fig. 5.(b).

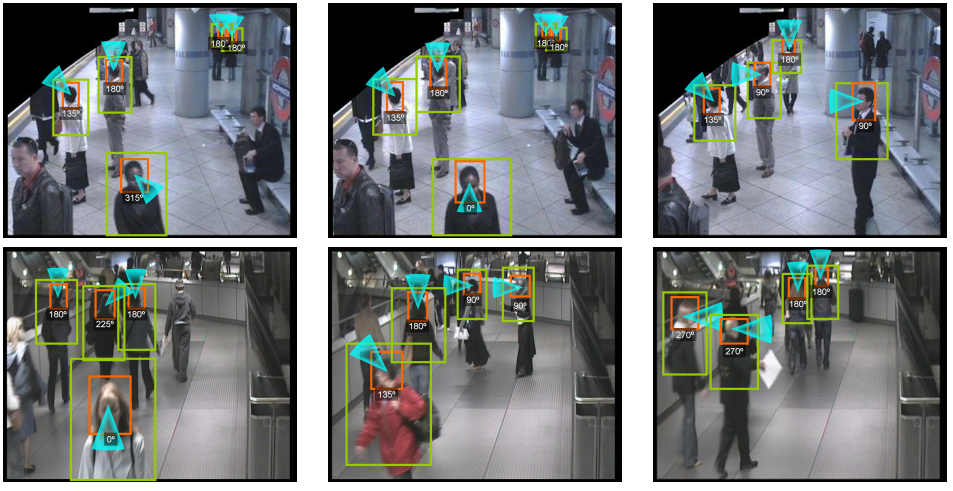


Figure 10: Examples of automated head pose classification of unknown multiple heads in two crowded underground stations. A search window is allocated using connected component after background subtraction and a HoG based pedestrian detection (green boxes). Head bounding boxes (red boxes) were determined using a hierarchical HoG sliding window detector. Head pose of each head bounding box was estimated by classification using a multi-class SVM (blue dial with pose value).

Automated Inference of Multiple Head Poses We evaluated the effectiveness of deploying our model for automated head pose inference on multiple people detected in a crowded public scene captured by the i-LIDS underground dataset. We obtain head candidates in two videos of over 10,000 frames by sliding a HoG based pedestrian detector in each video frame after background subtraction. In addition, we trained an upper body model which is hierarchically performed within a window of a detected pedestrian. Subsequently, the head candidate is located by connected components analysis inside the upper body part. Similarity distance feature maps were extracted from all detected head bounding boxes. Fig. 10 show some examples from extracted i-LIDS video sequences. It is evident that head pose estimation depends on the distance of the camera to each detected person in the scene. The head classification rate is 75% without prior information from previous frames. Misclassification of heads is achieved when head is misdetected. Therefore, the maximum head descriptor is built by sliding and scaling the templates, thus achieving a classification of 73%.

4 Conclusion

In this paper, we proposed a novel approach to head pose classification in crowded public scenes using low-resolution images captured under challenging viewing conditions. Our model is designed to avoid the need for explicit segmentation of skin and hair regions from a head image. More specifically, in order to cope with large degree of variations in the positions of pixels that correspond to skin and hair textures across different poses, and their non-uniform spread within a head image (i.e. there is often no clean-cut separation between skin and hair textures at the pixel level), we formulated a novel representational scheme to construct feature vectors using similarity distance maps. These distance feature maps are

then used to train a multi-class SVM for pose classification. We demonstrate significant performance advantages of our proposed model compared to a state-of-the-art model and another established technique for head pose classification under challenging viewing conditions in crowded public space given by the UK Home Office i-LIDS dataset.

References

[1] A. Ambroladze, E. Parrado-Hernandes, and J. Shawe-Tayor. Tighter pac-bayes bounds. In *NIPS*, 2006.

[2] B. Benfold and I. Reid. Colour invariant head pose classification in low resolution video. In *BMVC*, 2008.

[3] D. Beymer. Face recognition under varying pose. In *CVPR*, pages 756–761, 1994.

[4] L. Brown and Y. Tian. Comparative study of coarse head pose estimation. In *MOTION*, 2002.

[5] C. Burges. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[6] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[7] R. Debnath, N. Takahide, and H. Takahashi. A decision based one-against-one method for multi-class support vector machine. *PAA*, 7(2):164–175, 2004.

[8] S. Gong, S. Mckenna, and J. Collins. An investigation into face pose distributions. In *FG*, 1996.

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

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

[11] J. Ng and S. Gong. Composite support vector machines for detection of faces across views and pose estimation. *IVC*, 20(5):359–368, 2002.

[12] S. Niyogi and W. Freeman. Example-based head tracking. In *FG*, 1996.

[13] N. Robertson and I. Reid. Estimating gaze direction from low-resolution faces in video. In *ECCV*, pages 402–415, 2006.

[14] J. Sherrah, S. Gong, and E. Ong. Face distributions in similarity space under varying head pose. *IVC*, 19(12):807–819, 2001.

[15] H. Shimizu and T. Poggio. Direction estimation of pedestrian from multiple still images. In *IEEE Intelligent Vehicles Symposium*, 2004.

[16] Y. Tian, L. Brown, C. Connell, S. Pankanti, A. Hampapur A. Senior, and R. Bolle. Absolute head pose estimation from overhead wide-angle cameras. In *FG*, 2003.

[17] M. Voit, K. Nickel, and R. Stiefelhagen. Multi-view head pose estimation using neural networks. In *Canadian Conf. on Computer and Robot Vision*, pages 347–352, 2005.

[18] M. Voit, K. Nickel, and R. Stiefelhagen. A bayesian approach for multiview head pose estimation. In *IEEE Conf. Multisensor Fusion for Intelligent Sys.*, pages 31–34, 2006.