Full body image feature representations for gender profiling

Matthew Collins, Jianguo Zhang, Paul Miller, Hongbin Wang
The Institute of Electronics, Communications and Information Technology (ECIT)
Queens University Belfast
{mcollins17,jianguo.zhang,p.miller,h.wang}@qub.ac.uk

Abstract

In this paper we focus on building robust image representations for gender classification from full human bodies. We first investigate a number of state-of-the-art image representations with regard to their suitability for gender profiling from static body images. Features include Histogram of Gradients (HOG), spatial pyramid HOG and spatial pyramid bag of words etc. *These representations are learnt and combined based on a kernel support vector machine (SVM) classifier. We compare a number of different SVM kernels for this task but conclude that the simple linear kernel appears to give the best overall performance. Our study shows that individual adoption of these representations for gender classification is not as promising as might be expected, given their good performance in the tasks of pedestrian detection on INRIA datasets, and object categorisation on Caltech 101 and Caltech 256 datasets. Our best results, 80% classification accuracy, were achieved from a combination of spatial shape information, captured by HOG, and colour information captured by HSV histogram based features. Additionally, to the best of our knowledge, currently there is no publicly available dataset for full body gender recognition. Hence, we further introduce a novel body gender dataset covering a large diversity of human body appearance.*

1. Introduction

Recently, there has been much work in the area of behaviour analysis for video surveillance. However, an equally important issue, that has received relatively little attention thus far, is the ability to profile people in video data based on age and gender. Such profiling would allow future intelligent CCTV systems which could determine the intrinsic threat posed by certain individuals, or groups of individuals, to others. In this work we will only investigate gender classification without consideration of age. Specifically, we investigate gender classification in static images of full body pedestrians. To date, much of the work in computer vision with regard to gender classification has focused on facial gender recognition. Moghaddam and Yang [19] investigated the use of non-linear support vector machines (SVM) for gender classification with low resolution "thumbnail" faces. The SVM performance (3.4% error) was shown to be superior to traditional classifiers such as linear, quadratic, and Fisher linear discriminate, and more modern techniques such as radial basis function classifiers and large-ensemble RBF networks. Buchala *et al*. [3], investigated principal component analysis (PCA) for face classification with regard to gender, ethnicity, age and identity. They found that gender, ethnicity and age could be encoded in a relatively few number of PCA components, and that these were predominantly to be found amongst the first few. With respect to gender, they found information related to complexion, length of nose, the presence or absence of hair on the forehead, eyebrow thickness, and the presence or absence of a smile useful. Mäkinen and Raisamo [18] presented a systematic study on gender classification with automatically detected and aligned faces. They experimented with 120 combinations of automatic face detection, face alignment and gender classification. They found that the best gender classification rates of around 86% were achieved with an SVM.

Whilst the face contains many discriminating features, in a real world surveillance scenario facial based gender recognition is not really a viable option. Unlike building access controls, a CCTV camera is typically not capturing a close up image of a cooperative subject looking head on into the camera. Generally the camera field of view will be of a much wider area, and as such, if it can be seen at all, a face will be of a much lower resolution than traditional face based classifiers require. Also, it is likely that the face could be partially or entirely occluded if the subject in question is seen from the side or back. Within the area of psychology, studies have shown that reducing resolution and increasing noise leads to reduction in facial gender classification by human subjects [7]. It is clear that a different approach is required to identify gender in this type of scenario. As such it is proposed to investigate image representations for gen-

der classification based on full body images. We examine features such as body shape, clothing colour and other contextual clues in order to do this classification. Full-body based gender classification is a much harder problem than the face based approach. For example females and males may dress in the same colour clothing or not. Long hair can be an indication that a subject is likely to be female but it is not a guarantee. People of both genders can choose clothing or hair of very similar styles. Also full body images tend to contain more background clutter than cropped facial images, so there are further complications in extracting relevant information. Hence, it is still unknown how much information we can get from the human body appearance for gender identification. However, to a large extent, human perceivers can discriminate genders based on the appearance of a person, as well as the way he or she walks.

Making use of temporal information available in video footage, there has been some work done into recognising gender by human gait analysis from silhouettes [15]. Human body shape can be categorized using three components; endomorphy (soft and roundness), mesomorphy (hardness and muscularity) and ectomorphic (linearity and skinniness). These were originally defined by Sheldon *et al*. in 1940 [23] and were qualitative. However, quantitative values for these characteristics can be obtained using the Heath-Carter anthropometric method of somatotyping [6]. In a recent study, Munoz-Cachon *et al*. [20] found that females displayed higher rates of endomorphy or relative body fat, whereas mesomorphy tended to be higher among males and ectomorphy was similar in both sexes. These characteristics may be obscured by clothing but in recent research by Balan and Black [1] it has been shown that it is possible to infer a parametric 3D human body shape from images of clothed people and use it for gender classification. However, they require images of the person obtained with four differently viewing cameras. They capture four different silhouettes from each camera and search through the pose parameters such that for each set of values they generate four silhouettes from a learned model of 3D anthropomorphic data and these are then matched to those obtained from the cameras. Gender classification rates of 94% accuracy were obtained.

For the purposes of full-body gender classification, it would be useful to include automatic person detection, in order to segment out the figure before presenting the result to the gender classifier. This would enable an automated system to be developed which could work in real time surveillance scenarios. Viola *et al*. [27] presented an efficient moving pedestrian detector for video sequences. Their detector was trained using AdaBoost and combined both motion and appearance information to detect a walking person. For still images, Dalal and Triggs [9] present a method of using Histograms of Oriented Gradients which shows experimentally to significantly outperform existing feature sets for human detection. This work showed near perfect results on the publicly available MIT pedestrian database. However, to the best of our knowledge, there has been no previous work done looking at the problem of gender recognition from static full body images - with one exception [5]. It is intended to use the findings of this work as a benchmark for this study and to attempt to achieve comparable or superior results. Cao *et al*. [5] represented full body images in a fixed view (frontal or back) as a collection of patch features to model different body parts and provide a set of clues for gender recognition. They built an ensemble learning algorithm to combine the clues and attempt to recognise gender. Their best results achieved an accuracy of approx 75% for both fixed and mixed view.

2. Approach

Working with a database of still images, described in section 3.1, our approach to the full body gender recognition problem is to represent each image using local features. Unlike in the case of Cao *et al*., our features are computed across the whole image and not computed individually for parts of the images over a grid based part division. The images we use are already quite closely cropped to the figures so the influence of background information is negligible and we have found that our features based on the image as a whole have produced comparable results to Cao. Motivation for using features of this type came from recent work in the field of object categorization by Bosch *et al*. [2] where objects were recognized as part of a global class, and then further classified into specific sub categories. It was felt that this was analogous to the gender recognition problem, where both male and female are of the global class human/pedestrian.

The Pyramid Histogram of Gradients (PHOG) and Pyramid Histogram of Visual Words (PHOW) representations described in this work are state of the art representations. In their experiments on the Caltech 101 dataset, they found that different features achieved different classification performances depending on the specific task. For example when distinguishing between cars and airplanes, the shape based descriptor PHOG was more useful, however the appearance based descriptor PHOW was more useful to distinguish between horses and zebras. For some other object categories it was found to be most useful to use a combination of the two. Their best results, computed as the mean recognition rate per class of the Caltech 101 dataset, so that easier classes are not favoured, are reported as $69.0\% \pm 0.6$ for PHOG and $68.1\% \pm 0.6$ for PHOW [2].

A number of different representations and combinations of representations have been examined here and are described below with the PHOG and PHOW features serving as a starting point. In each case, the result is that for each

image a feature vector is produced. These feature vectors are then presented to an SVM classifier which extracts the discriminative information from the vectors and outputs a prediction score classifying the image as positive (male) or negative (female).

2.1. Representations

The following section gives an overview of the image representations used in our experiments.

2.1.1 Appearance Based Features

Pyramid Histogram of Words (PHOW)

This is a descriptor used to capture object appearance. The image is described using a histogram of visual words [2] drawn from a vocabulary of underlying local SIFT features [16]. SIFT descriptors are computed at points on a regular grid with spacing M pixels, here $M = 10$. At each grid point the descriptors are computed over circular support patches with radii $r = 4$, 8, 12 and 16 pixels. The SIFT descriptors are computed across each of the HSV channels to incorporate colour cues which gives a 128 x 3 D-SIFT descriptor for each point. K-Means clustering is then performed over a sample of training images per category selected at random to build a vocabulary per class. We chose to produce a vocabulary of 150 words for each set of training images. The male and female vocabularies were then combined into a 300 word global vocabulary before computing the feature vectors. The vocabulary is then used to produce a Bag of Visual Words representation for each image, which is in turn used to generate a Pyramid Histogram of Words feature vector for each image using the scheme proposed by Lazebnik *et al*. [14] which is based on spatial pyramid matching. Whilst these features are useful in describing the appearance of the image, a downside is the requirement to first build up vocabularies using a sample of the training images from each class and the need to first compute the underlying SIFT features for each image. However, in a live system, the vocabularies would likely be pre-computed, which would reduce the computational overhead.

2.1.2 Shape Features

Pyramid Histogram of Orientation Gradients (PHOG)

The PHOG descriptor [2] is used to describe object shape. Its motivation for use here is that it is felt that human observers look at body shape when classifying a person as male or female *e.g*. body contours, long/short hair etc. The underlying feature of this is an edge map, computed using Canny's method [4]. The feature vector describes the image by local shape and spatial layout of the shape. Local shape is captured by the distribution over edge orientations within a region, and spatial layout by tiling the image into regions at multiple resolutions. The descriptor consists of a histogram of orientation gradients over each image sub region at each resolution level - a Pyramid of Histograms of Orientation Gradients (PHOG).

Histogram of Canny Oriented Gradients (CHOG)

Bosch *et al*. refer to the Histogram of Oriented Gradients (HOG) of Dalal & Triggs [9] as inspiration for their PHOG descriptor. The HOG descriptor was originally proposed for use as a method of pedestrian detection in static images and showed near perfect performance for this task on the publicly available MIT Pedestrian Database [21] used by Cao *et al*. for their paper.

Looking at the way in which the PHOG descriptor was calculated, it was felt that it did not capture enough spatial information for the gender classification task at hand and was more suited to easier object recognition tasks where the difference in shape of two object categories is more obvious. At each level of the pyramid, each region of the pyramid was divided into a 4x4 grid. So for example at the third level the image is divided into 16 cells, where the size of those cells is determined by the original image size. The vectors from each pyramid level were then concatenated together to produce one long vector at the end.

Dalal & Triggs' INRIA HOG detector divides the image by a gridding process into a number of small connected cells of fixed size (6x6 pixels was found to be optimal for human detection) and for each cell a histogram of edge orientations is computed. Cells are then grouped by overlapping blocks of 2x2 cells so that each cell contributes more than once to the final descriptor and each block is normalised. The resulting feature vector is the concatenated histograms from each block [9].

It was felt it would be better to incorporate the more rigid geometric constraints of the original HOG detector of Dalal & Triggs. The original PHOG process of producing the Canny edge map and dividing it up into a number of bins remains the same, but here rather than computing the histograms over numerous pyramid levels and increasing numbers of regions, a single level is looked at and the gridding process, described in the HOG detector above, is applied to produce the feature vector for each image.

The main difference between this feature vector and the original HOG detector is the way in which the underlying gradient is computed. Dalal & Triggs found that the detectors performance was sensitive to the way in which the gradients were computed; they tested a number of methods but found that the simplest was the best for their purposes. They apply no smoothing to the image and use a 1-D [-1,0,1] derivative mask. Here however, we use an edge map computed using Canny's method, as in the PHOG features above. We found that for the purposes of gender classi-

fication, the soft connecting edges, which the Canny edge map determines whether or not to include through its use of high and low thresholds [4], had a positive effect on performance. This was verified by re-computing the features, using an edge map which used the low value threshold for both high and low, thereby including all connecting edges indiscriminately. This resulted in a drop in classification performance of as much as 4–8% depending on the test pass.

PixelHOG (PiHOG)

Whilst the CHOG feature type proved quite effective for gender classification, it was felt that perhaps its use of a Canny Edge map to compute the underlying gradients may discard too much information from the image as the resulting feature vector was quite sparse. For this reason it was decided to also investigate a more dense HOG based feature type more akin to the original INRIA detector to see if it could better capture the spatial information in the image. This is produced in the same way as the CHOG feature except that in place of computing the edge map using Canny's method, instead a custom edge map is used corresponding to all pixels where the gradient is above a threshold value. The result was an edgemap which included almost all pixels in the image.

We found a reasonable improvement in classification accuracy over the CHOG in using these denser feature vectors, but at a somewhat significant additional computational cost in training the SVM. Much of this computational overhead however relates to generation, and I/O of the text based files that SVM^{light} uses and in a real time system faster methods of input could be used to interact with the classifier.

2.1.3 Colour Features

Local HSV Colour Histogram (LHSV)

It was noted that when manually classifying images as male or female, that colour played an important role for the human performing the task. The modern convention in fashion seems to be that girls and women tend to prefer brighter colours and boys and men more comfortable with muted colours [12] and this pattern is also evident in the dataset images. Whilst this trend would not be robust enough of a discriminating factor alone it was felt that a feature which captured colour information in some way would be useful if combined with shape features.

This feature is essentially a Hue Oriented Histogram [25]. The image is divided into the three HSV channels. The H channel is divided up into a number of bins and the corresponding S value at each pixel is used at the voting weight for the histogram. The value in each histogram bin is the sum of the pixel values in the S Channel whose corresponding H Channel pixel value belongs to the current bin. The same HOG gridding process as used in the previous feature is then applied to the image and a histogram is computed for each cell of 6x6 pixels. And each overlapping block of 2x2 cells is the normalised vector after concatenating each of these four histograms. The final overall feature vector is the concatenation of each of the block vectors.

2.2. Classification

To demonstrate robustness of our features, five-fold cross validation was used. The overall accuracy is deemed to be the mean accuracy of the five test passes.

An SVM was used for classification as it is the underlying classification method of the INRIA HOG pedestrian detector and due to the fact that in Moghaddam and Yang's work on facial gender recognition, the SVM was shown to be robust with respect to scale and degree of detail. We used Joachims SVM^{light} implementation of the SVM classifier in our experiments. [13] which has both the Linear and RBF kernels as built in options and allows for user defined kernels also. As well as the basic Linear SVM we also examined the RBF, χ^2, and Intersection kernels with Linear proving to be the most accurate by a small degree.

- The Radial basis function (RBF) kernel defined as:

$$k(x_i, x_j) = \Sigma_{i=1}^{n}(\exp(-\gamma||x_i - x_j||^2)) \quad (1)$$

- The χ^2 Distance kernel [28]:

$$k(x_i, x_j) = \Sigma_{i=1}^{n} \exp\left(-\gamma\left(\frac{(x_i - x_j)^2}{(x_i + x_j)}\right)\right) \quad (2)$$

- The Intersection kernel [17]:

$$k(x_i, x_j) = \Sigma_{i=1}^{n} \exp(\gamma.min(x_i, x_j)) \quad (3)$$

In each of these kernels, optimisation of the gamma parameter was estimated using the method of Zhang *et al*. of using the inverse of the mean value from a distance matrix of the feature vectors thus reducing the computational cost of doing further cross validation to calculate the optimal value [28].

3. Experimental Results

3.1. Datasets

Due to the fact that the problem of full body gender classification is a new area of research, a publicly available database of human body pictures with gender labels does not currently exist. Cao *et al*. [5] manually labelled images from the MIT pedestrian database [21] for their study. This database consists of only front and back views of people, and a limited range of poses. They determined that there were 600 suitable images which they classified as male and 288 female images.

3.1.1 MIT Dataset

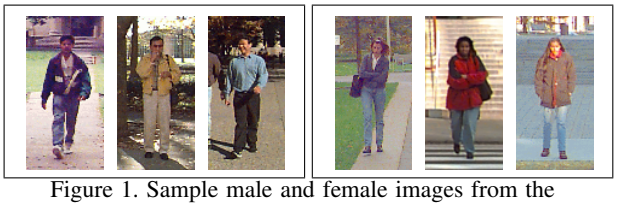

Figure 1. Sample male and female images from the MIT Dataset. (128x64 pixels)

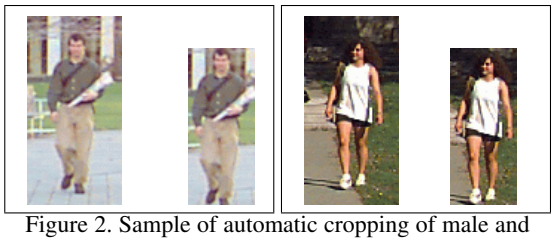

Figure 2. Sample of automatic cropping of male and female images on MIT Dataset. (106x45 pixels)

In order to compare results of our classification process to those of Cao *et al*. we have also manually categorised the MIT Pedestrian Dataset. The actual divisions of this dataset which were used in their work have not yet been released, however we got a similar overall number of images for male and female as they did. The only variable is perhaps which of the original MIT images we each chose to leave out of the set due to being unsure of the genders during manual labelling. The images in the MIT dataset are 128x64 pixels with the figure in the centre. We then further cropped the images down to 106x45 pixels giving a tight crop of the figure. (See Figure 2) This was done automatically based on the INRIA HOG detector which uses a 128x64 pixel search window but includes a margin around the returned person on all four sides. We determined that running the detector on the MIT images always returned the full 128x64 image with the person centred, so we could automatically remove the boundary pixels. In a real time system the detector would be run to locate a person and the margin would be automatically removed as a preprocessing step before generating feature vectors. The HOG detector uses this extra background information to provide context for person detection, however we found it had an adverse effect on our HSV colour based features for gender classification when compared to results from our other more tightly cropped VIPeR image dataset. A number of experiments were run using both the cropped and uncropped versions of the MIT images. For the purposes of our experiments, we decided to focus on frontal view images only. However the process is equally suitable for back and side poses also. Ideally a classifier robust to pose changes could be trained but for now, we focus on front poses for simplicity.

After manual classification of the MIT dataset, looking at front pose images, we had 305 male and only 123 female images. In our preliminary experiments we found that the unbalanced division of the data classes had a negative effect on classification. The classifier appeared to have a bias towards the majority class, classifying almost all images as male and producing misleading results. For example for the combined HOG and LHSV feature type which proved to give our most accurate results on balanced datasets, using this unbalanced class division for classification resulted in an overall mean accuracy across the 5 cross-fold divisions of 71.97%. However looking at this more closely it had classified all but one or two female images at most as male for each pass of the cross validation, resulting in a mean positive accuracy of 100% and a negative accuracy of only 2.47%. For this reason, when looking at results for a binary classification problem such as this, it is important to look at accuracies on an individual class level as well as the overall result. Whilst it is true that this over fitting can be accounted for by the use of a weighted SVM or other techniques, this is beyond the scope of this paper as we are focused on the image representations themselves. For this reason it was decided to balance the MIT dataset by randomly selecting 123 male images to match the number of female images available and discard the rest. It has been shown that results on quite unbalanced data tend to less meaningful [24].

3.1.2 VIPeR Dataset

Figure 3. Sample male and female images from VIPeR Dataset.

It was also felt that after balancing the MIT dataset, perhaps the number of images available would not be sufficient to rigorously evaluate our representations. Also given that the database was initially designed for pedestrian recognition, not gender classification and all the frontal images are in the same standardised pose, it was felt they might not test the feature types robustly enough.

For this reason we have also gathered images from other sources and manually categorised them by gender and pose. A further 1249 images have been categorised from another publicly available dataset. This dataset, VIPeR, was constructed for use with Viewpoint Invariant Pedestrian Recog-

nition [11], but is equally suitable for our work after manual categorisation. This set contains front, back and side profile images. The VIPeR dataset is more evenly split between the genders, so no special balancing considerations had to be made. There are 292 male front view images and 291 female. The images are 128x48 pixels and are a tighter crop to the figure, leaving less background than those in the MIT set so no preprocessing to remove margin pixels was required. Also, within the set of frontal images the poses are more varied than those of the MIT set. Within both datasets each image is of a unique subject thus ensuring that there is no overlap of subjects between training and test sets.

3.2. Comparison

Comparison of SVM Kernels:

In all, four kernels were investigated for the classification task; Linear, RBF, χ^2 and Intersection. The simple Linear SVM which just takes the dot product between each pair of feature vectors seemed to produce the most accurate results (see Figure 4 for example). It has been shown previously that if the dimensionality of feature vectors is high, then the performance of the Linear SVM is generally quite similar to that of the RBF [8] and this is confirmed by our results, however for completeness we also investigated the other kernel functions.

The INRIA HOG detector binaries available can be retrained for any detection problem. We investigated retraining the detector using male images as the positive example, and female images as the negative to see how a detector specifically suited to the human form would perform when differentiating between genders in comparison to our other features. However, we found it would not discriminate between the classes to a satisfactory level. This is perhaps in part to do with the relatively small datasets involved. The original detector used 1239 positive training examples and 12180 negative examples of person-free images. Our present available gender datasets cannot come close to matching these numbers.

Feature	Male Accuracy	Female Accuracy	Overall Accuracy
PHOW	53.67±11.78	60.90±11.50	57.25± 6.81
PHOG	43.83±12.22	59.30± 8.75	51.59± 5.44
CHOG	69.20±10.55	70.73±10.36	69.92± 4.36
LHSV	67.50± 5.61	49.60± 4.41	58.52± 2.48
CHOG + LHSV	72.47± 7.84	69.97±12.71	71.15± 4.79
PiHOG	71.50± 8.75	72.37±10.76	71.90± 7.22
PiHOG + LHSV	72.37± 8.86	72.37±16.45	72.28± 8.07

Table 1. Average classification accuracy (%) of different features by linear SVM on uncropped MIT images

Comparison of Features:

Across all 3 sets of results, a dramatic improvement can be seen between the PHOG and CHOG results. Incorporating the tighter geometric constraints of the INRIA HOG Detector designed for human figures into the PHOG process clearly captures much more relevant spatial information about the image suitable to the task of full body gender recognition. Further improvement is also noted when moving to the PiHOG representations.

Comparing the results in Tables 1 and 2 we can see that the CHOG features actually performed better on the Uncropped version of the MIT images (69.92%) than the cropped versions (65.11%). This is likely due to the same

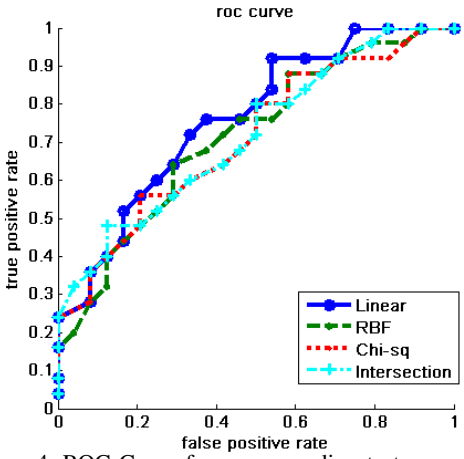

Figure 4. ROC Curve for corresponding test passes on CHOG+LHSV features from MIT dataset over all four kernels indicating similar performance but with Linear proving slightly better overall.

Feature	Male Accuracy	Female Accuracy	Overall Accuracy
PHOW	60.03±10.34	53.50± 9.45	56.84± 6.81
PHOG	47.10± 6.90	56.77±13.94	51.96± 7.83
CHOG	62.70±11.17	67.57± 9.90	65.11± 8.34
LHSV	65.13± 9.30	56.97±12.00	60.98± 5.58
CHOG + LHSV	67.63± 9.76	70.03±17.27	68.71± 7.57
PiHOG	72.37± 9.92	76.50±10.25	74.42± 7.28
PiHOG + LHSV	72.40± 6.35	79.73±16.19	76.00± 8.13

Table 2. Average classification accuracy (%) of different features by linear SVM on cropped MIT images

Feature	Male Accuracy	Female Accuracy	Overall Accuracy
PHOW	69.16± 5.92	58.07± 5.00	63.63± 2.58
PHOG	64.00± 8.14	42.96± 8.58	53.52± 4.22
CHOG	69.17± 5.49	73.90± 7.07	71.53± 5.80
LHSV	73.99± 3.98	66.76± 8.12	70.37± 2.92
CHOG + LHSV	77.08± 3.55	78.35± 4.65	77.70± 0.83
PiHOG	77.74± 2.45	71.83± 8.28	74.79± 3.73
PiHOG + LHSV	79.47± 4.87	81.80± 7.72	80.62± 2.58

Table 3. Average classification accuracy (%) of different features by linear SVM on uncropped VIPeR dataset images

effect observed in the original INRIA HOG detector where the margin of extra pixels around the figure located was used to provide extra contextual information which the detector found useful. It is clear however, that for other features which use the entire image, the extra background information introduces noise and lowers the overall classification accuracy. This can be seen in both the LHSV features and the PiHOG. The Canny edge map would discard much of the boundary information except for some connecting soft edges, whereas in the case of PiHOG the entire boundary would be included accounting for the introduction of noise in this case.

As can be seen from the VIPeR dataset results in Table 3, the highest accuracy of 80.62% is reported when combining the PiHOG feature vectors with the LHSV.

We found that the best method for combining feature types was that presented by Zhang *et al*. in [10] of taking each of the 2 scores outputted by the SVM for both feature types and putting them together into new feature vectors for each image, before presenting these new feature vectors to another SVM for classification. The results confirm the speculation, that the shape based features of the HOG would complement the colour information of the LHSV thus providing more discriminative information for the classifier. The same trend is observed in combing the LHSV features with the CHOG which provides 5% improvement for the VIPeR dataset than CHOG alone. A similar improvement over CHOG/PiHOG alone can be seen in the both cropped and uncropped MIT dataset when LHSV features are also taken into account, although to a slightly lesser degree. The relatively low accuracy score for LHSV alone in the MIT set of only 60.98% when compared to the PiHOG score of 74.42% accounts for this lower increase. The trend of females wearing brighter clothes holds across both datasets, but it is more evident in the VIPeR set and this is reflected in the results. In the case of the VIPeR set both individual scores are quite high to begin with and closer together and in this way they complement each other particularly well. We also tried combining corresponding feature vectors into one long vector and presenting this to the SVM, but found the results were slightly less accurate than our other combination method. In some particular cases in the MIT sets in fact, combining the significantly lower accuracy LHSV features with the HOG based features actually dragged the overall accuracy down a little.

Based on the fact that the CHOG features performed better on the Uncropped MIT images, an experiment was run to check how combining the Uncropped CHOG features with the cropped LHSV features would perform. However there is negligible difference between this and the case where both features are computed from the uncropped images. As in the VIPeR dataset, cropped PiHOG combined with cropped LHSV features still proves to be the best overall feature choice.

The appearance based PHOW features achieved accuracies in the range of 56-64%. This is better than random guess, but they are out performed by the LHSV features. Some experiments where also done in combining these PHOW features with the LHSV and HOG based features, but no improvement in overall performance was observed.

Figure 5. Examples of misclassified images from VIPeR dataset
(a) Misclassified by PiHOG alone (shape)
(b) Misclassified by LHSV alone (colour)
(c) Misclassified even when both LHSV and PiHOG features were used together.

Figure 5 shows example images which were misclassified in the VIPeR dataset. Figure 5 (a) shows images which were misclassified by PiHOG alone (shape) but correctly classified when colour information from LHSV was also taken into account. Note the bright colours of the female images and dark male images. Similarly 5 (b) shows images misclassified by LHSV colour information alone, but adding shape information took body shape into account and corrected the error. Finally, 5 (c) gives a sample of misclassified images even when both LHSV and PiHOG features were used together. This is unsurprising. For example the first image is female, but this may not be immediately obvious even to a human observer and could easily be mistaken for an overweight male. Similarly the long hair on the male in the 4th image may account for its misclassification as the majority of male training images would have had short hair.

4. Conclusion

This paper presents an experimental investigation into the problem of gender recognition from full body static images with a view to emulating the rapid progress of the field of generic object recognition in that of specific gender profiling. It examines a number of feature types for extracting discriminative information from the images suitable for classification. By combining both shape and colour information we achieved the best accuracy of 80.62% demonstrating that multiple cues should be taken into consideration when classifying full body images and no one feature type is sufficient to capture all relevant information alone.

Future work will examine more sophisticated methods of combining features such as multiple-kernel learning [26, 22], to see if this can improve the classifier score. We will also investigate other feature types to try and capture a wider range of information from the images than shape and colour cues. It would also be good to examine some alternative colour representations than the current LHSV implementation we are using to see if any improvement can be gained.

5. Acknowledgement

The authors acknowledge funding from the Department for Employment and Learning Northern Ireland (DEL), the Queen's University Belfast Research Support Package D8203EEC, and from EPSRC grant EP/E028640/1 ISIS.

References

[1] A. O. Balan and M. J. Black. The naked truth: Estimating body shape under clothing. In *ECCV '08*, pages 15–29, Berlin, Heidelberg, 2008. Springer-Verlag.

[2] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *CIVR '07*, pages 401–408, New York, NY, USA, 2007. ACM.

[3] S. Buchala, N. Davey, T. M. Gale, and R. J. Frank. Principal component analysis of gender, ethnicity, age, and identity of face images. In *IEEE ICMI*, 2005.

[4] J. Canny. A computational approach to edge detection. *PAMI, IEEE Transactions on*, 8(6):679–698, 1986.

[5] L. Cao, M. Dikmen, Y. Fu, and T. S. Huang. Gender recognition from body. In *MM '08: Proceeding of the 16th ACM international conference on Multimedia*, pages 725–728, New York, NY, USA, 2008. ACM.

[6] J. E. L. Carter and B. H. Heath. *Somatotyping-development and applications*. Cambridge University Press, Cambridge [England]; New York, 1990.

[7] A. Cellerino, D. Borghetti, and F. Sartucci. Sex differences in face gender recognition in humans. *Brain research bulletin*, 63(6):443–449, 7/15 2004.

[8] O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram-based image classification. *Neural Networks, IEEE Transactions on*, 10(5):1055–1064, 1999.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. *CVPR 2005. IEEE Computer Society Conference on*, 1:886–893 vol. 1, 2005.

[10] M. Everingham, A. Zisserman, C. Williams, and L. V. Gool. The pascal voc 2006 results. *technical report*, 2006.

[11] D. Gray, S. Brennan, and H. Tao. Evaluating appearance models for recognition, reacquisition, and tracking. *PETS. IEEE International Workshop on*, 2007.

[12] http://histclo.com/Gender/color.html.

[13] T. Joachims. *Making large-scale support vector machine learning practical*. Advances in kernel methods: support vector learning. MIT Press, Cambridge, MA, USA, 1999.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *CVPR, 2006 IEEE Computer Society Conference on*, 2:2169–2178, 2006.

[15] X. Li. Gait components and their application to gender recognition. *IEEE SMC. Part C, Applications and reviews*, 38(2):145, 2008.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[17] S. Maji, A. C. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[18] E. Makinen and R. Raisamo. Evaluation of gender classification methods with automatically detected and aligned faces. *PAMI, IEEE Transactions on*, 30(3):541–547, 2008.

[19] B. Moghaddam and M.-H. Yang. Learning gender with support faces. *IEEE TPMI*, 24(5):707–711, 2002.

[20] M. Muoz-Cachn, I. Salces, M. Arroyo, L. Ansotegui, A. Rocandio, and E. Rebato. Body shape in relation to socioeconomic status in young adults from the basque country. *Collegium antropologicum*, 31(4):963–8, 2007.

[21] M. Oren, C. Papageorgiou, P. Sinha, E. Osuna, and T. Poggio. Pedestrian detection using wavelet templates. *CVPR*, pages 193–199, 1997.

[22] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. Simplemkl. *JMLR*, 9:2491–2521, November 2008.

[23] W. H. Sheldon, S. S. Stevens, and W. B. Tucker. *The varieties of human physique : an introduction to constitutional psychology*. Harper, New York, 1940.

[24] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *JMLR*, 7:1531–1565, 2006.

[25] K. van de Sande, T. Gevers, and C. Snoek. Evaluation of color descriptors for object and scene recognition. *CVPR 2008. IEEE Conference on*, pages 1–8, 2008.

[26] M. Varma and D. Ray. Learning the discriminative power-invariance trade-off. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 2007.

[27] P. Viola, M. J. Jones, and D. Snow. Detecting pedestrians using patterns of motion and appearance. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 734–741 vol.2, 2003.

[28] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, Vol. 73, No. 2:213–238, 2007.