Predicting occupation via human clothing and contexts

Conference Paper *in* Proceedings / IEEE International Conference on Computer Vision. IEEE International Conference on Computer Vision · November 2011

DOI: 10.1109/ICCV.2011.6126355 · Source: DBLP

Zheng Song
National University of Singapore
17 PUBLICATIONS **974** CITATIONS
SEE PROFILE

Meng Wang
University of South China
318 PUBLICATIONS **9,433** CITATIONS
SEE PROFILE

Xian-Sheng Hua
Microsoft
335 PUBLICATIONS **9,394** CITATIONS
SEE PROFILE

Predicting Occupation via Human Clothing and Contexts

Zheng Song[1], Meng Wang[2], Xian-sheng Hua[3], Shuicheng Yan[1]

[1] Department of Electrical and Computer Engineering, [2] School of Computing, National University of Singapore

[3] Microsoft Research Asia

{zheng.s, eleyans}@nus.edu.sg, dcswang@nus.edu.sg, xshua@microsoft.com

Abstract

Predicting human occupations in photos has great application potentials in intelligent services and systems. However, using traditional classification methods cannot reliably distinguish different occupations due to the complex relations between occupations and the low-level image features. In this paper, we investigate the human occupation prediction problem by modeling the appearances of human clothing as well as surrounding context. The human clothing, regarding its complex details and variant appearances, is described via part-based modeling on the automatically aligned patches of human body parts. The image patches are represented with semantic-level patterns such as clothes and haircut styles using methods based on sparse coding towards informative and noise-tolerant capacities. This description of human clothing is proved to be more effective than traditional methods. Different kinds of surrounding context are also investigated as a complementarity of human clothing features in the cases that the background information is available. Experiments are conducted on a well labeled image database that contains more than 5,000 *images from* 20 *representative occupation categories. The preliminary study shows the human occupation is reasonably predictable using the proposed clothing features and possible context.*

1. Introduction

This paper studies the problem of categorizing human via their occupations using the visual information of images. The recognition of human categories like gender and age [10] [17] has already attracted great interest in computer vision research community due to its wide application scenarios and commercial potentials. Specifically, on many social network web sites with photo sharing (e.g., Facebook and Google Picasa), the shared photos contain rich information about the uploaders and the persons they associated with. Many intelligent services, such as friend, group, news or product recommendation, can be better provided if detailed attributes of these persons can be automatically recognized from the photos. While extensive efforts have been dedicated to human gender and age analysis, in this work we further explore the human occupation information which can contribute greatly in analyzing the users' profile as well as their social circle and activities. After that, intelligent recommendation can be provided according to the corresponding person occupation and its related information.

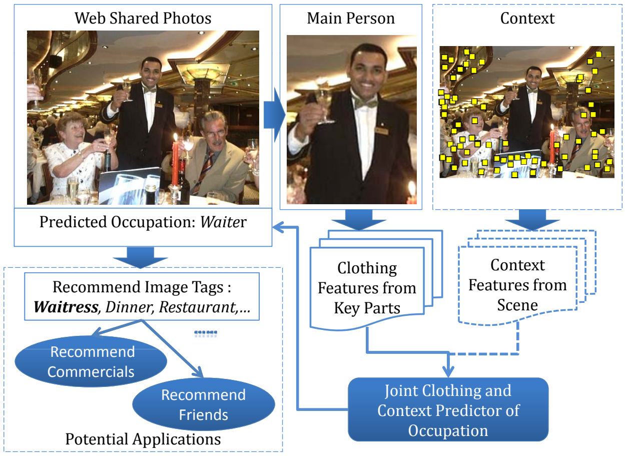

Web Shared Photos
Main Person
Context
Predicted Occupation: *Waiter*
Clothing Features from Key Parts
Context Features from Scene
Recommend Image Tags :
Waitress, *Dinner, Restaurant,...*
Recommend Commercials
Recommend Friends
Potential Applications
Joint Clothing and Context Predictor of Occupation

Figure 1. A schematic illustration of the human occupation prediction framework and its potential applications. The context features part is drawn in dash lines denoting that they are only available in certain occasions.

In this paper, we propose to recognize human occupation via two properties of human-related images, *human clothing* and *scene contexts*. The scheme is shown in Figure 1.

Some efforts have been devoted to preliminary analysis of human clothing such as clothes segmentation [6] and clothes animation [20]. Chen *et al*. [5] proposed an And-Or graph representation of clothes compositions, which models the clothes composition by a graph of clothes parts and provides some preliminary results on the recognition of clothes styles. Gallagher et al. [11] conducted recognition of person identity within family photos via low-level features extracted from both faces and human bodies which can reliably identify the same clothing of person.

However, previous approaches did not dive into the recognition of high-level semantic categories of human such as occupations. The main drawback of traditional recognition/classification methods for human occupation prediction is their insufficiency in semantic modeling. Category models learned from pixel-level features can only express some basic features such as shape, color or texture, and thus can only recognize limited kinds of patterns. On the other hand, the concept of human occupation is a mixture of different visual semantics with complex grammar. Therefore, simple modeling of human occupation category would not be the optimal choice. To further obtain the abstract occupation information, we propose to refine features of human-related images based on part based human alignment as well as the modeling of semantically meaningful patterns.

Many studies have shown that part-based model [9] [3] is able to efficiently improve the prediction reliability owning to its better alignment as well as excellent robustness to occlusion. Semantically meaningful recognition also attracts much attention for complex categorization problems. Kumar *et al*. [12] propose a semantic-level description to faces by modeling hair color, eye clothing, etc., which outperforms most previous approaches using low-level features. The experiments well validate the advantages of semantic-level modeling.

Inspired by these studies, we proposed a part-based representation of human clothing to model possible semantically meaningful patterns of human dressing parts. Note that we only concern human upper body (including hat/hair and faces) due to the fact that the information in the upper body area shows better discriminative ability since lower body areas are frequently occluded and hence non-informative.

Moreover, it is observed that contexts contain direct indicative information for some occupations. Hence in this work, we also investigate the context features with respect to the concerned human in image. The context features within images are proved to be closely related to the image theme by the recent study of Zheng *et al*. [23] and Divvala *et al*. [8]. In addition, context features have been well validated for many tasks, e.g. action recognition in [22].

The main contributions of this paper can be summarized as follows:

1. To our best knowledge, this is the first attempt to predict the occupations of humans in images.

2. We propose a novel approach to robustly model the human clothing. The empirical results show that this model is useful for the prediction of human categories.

3. We collect and annotate a large image dataset as benchmark for human occupation prediction which will be released to encourage further research in this direction.

2. Occupation Prediction Framework

2.1. Part-based Modeling of Human Appearance

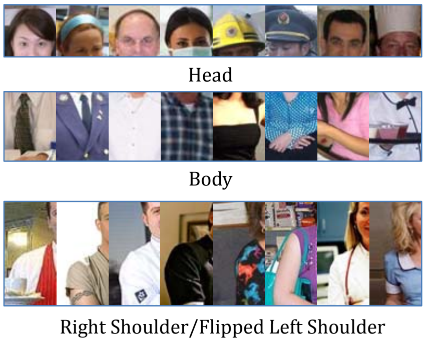


Figure 2. Exemplar image patches of human key parts extracted from the occupation dataset.

We propose to use part-based appearance model for human upper body, and as illustrated in Figure 2 we concern on four key parts of human upper body located at:

- **Head**: This part reveals the human haircut or hat dressing. Haircut is a key attribute for human gender recognition, and hat style is also an important cue for many occupations, such as police and chef.

- **Central Upper Body**: This part mainly contains human upper body clothing including collar, button, color of clothes, etc. Some common dressing styles such as shirts and overcoats can be reliably modeled via this part.

- **Left and Right Shoulders**: These two parts correspond to the arm and shoulder parts of human dressing. These two parts also convey much information of human clothing such as style of sleeves.

Firstly, We introduce an efficient automatic human detection and key point localization method. The state-of-the-art face [19] and human detection [3] approaches are implemented to accurately locate human while handling the large variation in head and upper body poses.

Corresponding to these four pre-defined human part patches, we define four key points of human upper body located respectively at the top of head, the connection point of head and neck, and central point of left and right shoulders. We manually label the key points on our proposed occupation dataset and with the well-labeled ground truth we learn a regression model to infer the four key points using the coordinates of the detected face and body bounding boxes. Consequently, automatic localization of the key points can be achieved after performing face and human detection.

Thereafter centered to these four key points, four image patches are extracted after proper rotation and scaling and low-level image descriptors are extracted from the rectified image patches as representations to the aforementioned human body parts. We adopt five widely-used image low-level features as the basic description of human body parts, i.e., the Histogram of Oriented Gradient (HOG) [7], Local Binary Patterns (LBP) [18], color histogram over the CIELAB color space, and histogram of color/texture gradient [16]. We implement the dense-grid description of these features for the image patches following the state-of-the-art method in [13] since it best compromises feature robustness and discriminative ability for images in natural environment. The sizes of dense grids are set by the image patch size for each body part.

2.2. Sparse Representation Based Human Occupation Prediction

In this subsection, we elaborate on the comprehensive and noise-tolerant descriptors in semantic level for human occupations. Given the extracted low-level image descriptors, traditional classification methods typically use feature concatenation or the Bag-of-Words framework to learn discriminative models [11] [13]. However, our proposed occupation classification problem is for complex categories and hence cannot fully rely on previous framework due to: 1) low-level image features are only sensitive to pure appearance discrimination while discarding the semantically meaningful characteristics for a category; and 2) occupation categories might be divided hierarchically while certain categories could share similar patterns of features which cannot be learned from traditional discriminative learning. Therefore, it demands modeling via intermediate level patterns for occupation classification.

Therefore, we propose a robust and efficient method to adaptively learn sample-based patterns as semantic level representation, where the pattern refers to the low-level descriptors of the human body parts in this problem. Note that there are five kinds of low-level features for each of the four human part and thus there are 20 kinds of low-level descriptors for human clothing.

We require to learn specific patterns which are descriptors from a few training samples which can reliably reconstruct most other samples. This requirement will well guarantee the representative capacity of the learned patterns and hence to be meaningful. Recent research on reconstruction via sparse coding [14] [21] can provide controllable reconstruction basis size while being robust to gross noises. Based on these studies, we implement the multi-task extension of sparse coding method. A rough scheme is illustrated in Figure 3.

For a given image descriptor from the aforementioned 20 descriptors, we formulate the so-called multi-task sparse coding [21] to learn the intermediate level pattern dictionary. Denote $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_J]$ and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_K]$ as the training set and validation set respectively, where $\mathbf{x}_i$ and $\mathbf{v}_j$ are the feature vectors extracted from the part instance in the i-th training sample and the j-th validation sample respectively. We consider the simultaneous reconstruction of all validation $\mathbf{V}$ using $\mathbf{X}$ as the so-called multi-task and a few intermediate level patterns are learned within $\mathbf{X}$ which demands the sparsity.

More specifically, each of the tasks refers to the reconstruction of one validation datum, namely,

$$\mathbf{v_k} = \mathbf{X}\boldsymbol{\beta}_k, k = 1, 2, ..., K, \tag{1}$$

while $[\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, ..., \boldsymbol{\beta}_K] = \mathbf{B}$ is the reconstruction coefficient matrix.

For all these K reconstructions, only a few common training samples in $\mathbf{X}$ are expected to be selected, which implies that many rows in $\mathbf{B}$ are zeros. Therefore the mathematical formulation of multi-task sparse coding can be written as,

$$\hat{\mathbf{B}} = \arg\min_{\mathbf{B}} \frac{1}{2} \sum_{k=1}^{K} \|v_k - \mathbf{X}\boldsymbol{\beta}_k\|^2 + \lambda \sum_{j=1}^{J} \|\mathbf{B}_j\|_2, \tag{2}$$

where $\mathbf{B}_j$ denotes the jth row vector of the matrix $\mathbf{B}$. The multi-task sparsity is then guaranteed by the $\ell_{2,1}$-norm constraint of $\sum_{j=1}^{J} \|\mathbf{B}_j\|_2$ [21]. The parameter λ is empirically set to be $1e-3$ in this work and the sparseness of $\mathbf{B}$ is thus around $0.1 - 0.2$, *i.e.* about 10% to 20% entries of $\mathbf{B}$ are non-zeros.

Eqn. (2) can be solved by an iterative updating rule as shown in [21]:

$$\mathbf{B}_j = (1 - \frac{\lambda}{\|\mathbf{S_j}\|})_+ \mathbf{S_j}, \tag{3}$$

where $\mathbf{S_j} = [s_{j1}, s_{j2}, ...s_{jK}]$ and $s_{jk} = \mathbf{x}_j^T(\mathbf{v}_k - \mathbf{X}\boldsymbol{\beta}_k^{-j})$. $\boldsymbol{\beta}_k^{-j}$ denotes $\boldsymbol{\beta}_k$ with the jth element to be zero and the symbol $(c)_+$ means $\max(c, 0)$. Note that this iterative procedure requires all column feature vectors in $\mathbf{X}$ and $\mathbf{V}$ to be normalized.

After obtaining the sparse coefficient matrix $\mathbf{B}$, we can obtain $\mathbf{X}_b = [\mathbf{x}_{b_1}, \mathbf{x}_{b_2}, \mathbf{x}_{b_s}]$, where $\{b_1, b_2, ...b_s\}$ correspond to the indices of the non-zeros rows of B as the intermediate level patterns. We can then proceed to represent all samples (including training, validation and test) by the learned $\mathbf{X}_b$ using various approaches. In this paper, to be consistent with the learning procedure, we adapt traditional Lasso-based sparse coding [1] for the representation, namely,

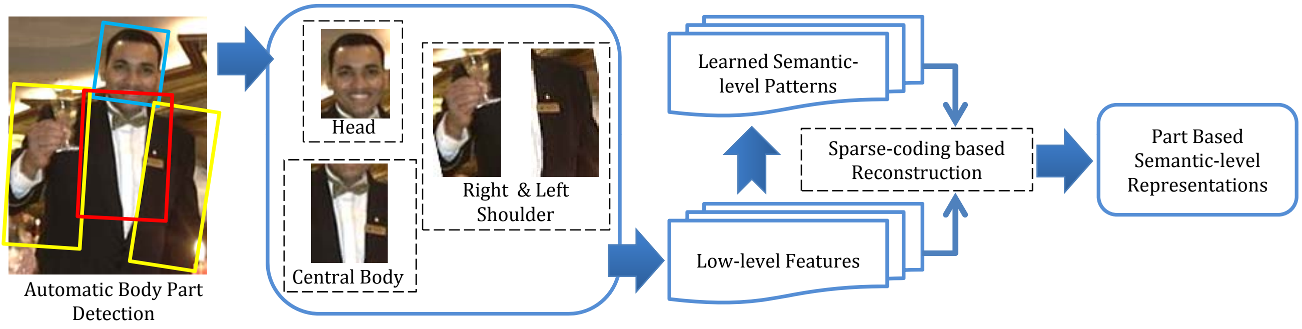

Automatic Body Part Detection
Head
Right & Left Shoulder
Central Body
Learned Semantic-level Patterns
Sparse-coding based Reconstruction
Low-level Features
Part Based Semantic-level Representations

Figure 3. Illustration of the semantic level feature extraction for occupation predition

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{x} - \mathbf{X}_b\boldsymbol{\beta}\|^2 + \lambda'\|\boldsymbol{\beta}\|_1, \qquad (4)$$

where $\boldsymbol{\beta}$ is the reconstruction coefficients, and $\mathbf{x}$ is a new part feature vector to be reconstructed. The parameter λ here is empirically set to be 0.02 to obtain similar the sparseness of $\boldsymbol{\beta}$ with the learned B.

The reconstruction process is repeated for all the 20 aforementioned human clothing descriptors and the proposed occupation descriptor is represented as the concatenation of the reconstruction coefficient $\boldsymbol{\beta}$'s from these image descriptors.

2.3. Foreground Enhanced Context Features

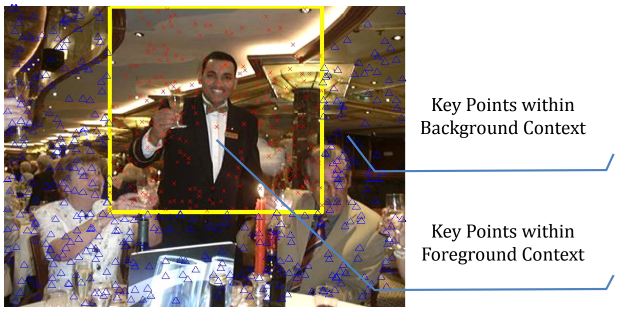

Key Points within Background Context
Key Points within Foreground Context

Figure 4. An illustration of the context feature extraction process. The yellow rectangle denote the division of the defined foreground and background context areas. The blue triangles and red crosses denote the SIFT key points belonging to background and foreground respectively. For better viewing, please see original color pdf file.

Apart from the human clothing information, context information is also critical for recognizing certain occupations, especially for naturally captured photos. The usefulness of context information for occupation prediction can be intuitively demonstrated via the following cases:

1. Objects that have certain interactions with human can indicate occupation information, such as an instrument for a player or a car for a driver.
2. The background of an image may indicate the working place of the human and this information thus helps to predict his/her occupation, such as a construction field for a worker.

Consequently, there are mainly two layers of context information: foreground context and background context. The foreground context is associated with the human area, *e.g.* object(s) the human is holding, and the background context contains the environmental information. Since background context of images is often cropped for human-related images, we propose to model the context information with enhancement to foreground.

We use sparse SIFT [15] features and Bag-of-Words (BoW) modeling to represent the context layers of an image as illustrated in Figure 4. Firstly, the image is divided into the foreground and background layers with a rectangle box. We locate the box by adding a margin to the detected human area. Then the SIFT key points are divided into these two layers according to their locations and two codebooks of visual words are trained using SIFT descriptors from these two layers respectively. Then the BoW features of the images are extracted by projecting the SIFT features onto the learned two codebooks with higher weights for foreground visual words.

Note that, in order to keep the completeness of information, we do not exclude the information of human clothing in the construction of context features. Although duplicate representations with the human clothing features are introduced, it is a reasonable approach to sufficiently model context objects with uncertain positions. Some useful information that is considered as noises in human clothing features may be modeled in the foreground context.

It can be considered that BoW framework is similar to our proposed semantic-level human clothing feature with the difference that BoW seeks representative descriptors in low-level features and hence be faster yet inaccurate. Considering that context features of image background are rather complex but only used for assistance to human clothing features, using BoW features is a good trade-off.

Note that besides the above mentioned features, some facial features, e.g. gender (most farmers are men) and ethnic group, are also useful for occupation prediction. In this work, we do not use these features because 1) the faces may be in non-frontal views although the human bodies are constrained to be near-frontal in this work as later mentioned, and 2) gender and ethnic estimation under non-frontal face setting is still an unsolved problem.

3. Dataset Construction

For this dataset, we aim to collect diverse images with representative occupations that can well verify the effectiveness of algorithms for occupation prediction via human dressing and context features. We identified twenty occupation categories from over 200 existing occupations listed in the Wikipedia. Our criterion is that the occupation categories should be informative enough for recognition from photos . Exemplary images of these 20 occupations categorization along with their intuitively useful feature types are shown in Figure 5.

We collected the images with text queries related to these 20 occupations from two popular image search engines, Google and Bing. Then three volunteers participated in the annotation process to 1) delete the noisy or duplicate samples, and then 2) for those images containing near-frontal human upper bodies, label the four key points at head, neck, and shoulders. The final dataset contains 5589 images in total.

4. Experiments

4.1. Human Clothing Descriptors Comparison

In this subsection, we compare our proposed human clothing features based on part-based semantic-level modeling with traditional descriptions proposed in [11] which implement similar low-level features from aligned upper body and face areas. We hence implement [11] by directly extracting the low-level features described in Section 2.1. Since the feature dimension from the body area is quite large, we reduce the feature dimension to 2000 with Principal Component Analysis (PCA) for efficient modeling. To rule out potential errors caused by automatic human and face detection, we conduct the experiments directly based on the manually labeled key points for both methods.

The proposed human clothing features are constructed following the scheme shown in Figure 3. Finally a sparse descriptor with dimension 2741 can be obtained via the concatenating reconstruction coefficients. To fairly compare the features, we also perform PCA dimensionality reduction to reduce the feature dimension to 2000.

We split the occupation dataset into train, validation and test sets (25% train, 25% validation and 50% test) for the construction of our proposed clothes features. For features of [11], the train and validation set are both used for training. The split is fixed to facilitate bag-of-words related approaches. Otherwise each time the split changes, the dictionary of bag-of-words need to be retrained.

Table 1. Comparison of prediction accuracy on occupation dataset from our proposed human clothing descriptors with low-level descriptor proposed in [11].

	Prediction Accuracy (%)
Descriptors in [11] (Manual)	43.89
Our Descriptors (Manual)	**57.14**
Our Descriptors (Automatic)	**52.01**

The occupation prediction is performed using the "one-vs-one" based multi-class Support Vector Machine (SVM) [4] for the 20 occupation classes. Linear kernel is used and the margin cost C is searched for best performance. The detailed comparison results in Table 1 shows that our proposed human clothing descriptors significantly outperform those proposed in [11].

4.2. Evaluation of Human Key Point Detection

Table 2. Error in x and y coordinates while predicting the positions for human upper body key points. The error is measures by pixel after normalizing the human upper body image to size of 128×128 pixels.

	Head	Neck	Left shoulder	Right shoulder
x	5.7601	5.4308	6.8293	6.8531
y	3.8387	3.7353	8.4754	8.5629

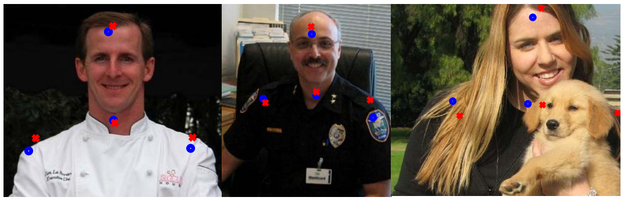


Figure 6. Comparison of manually labeled (blue circles) and automatically detected (red crosses) positions for the four human body key points. The shown images are cropped from the original detection results for better illustration. For better viewing, please see original color pdf file.

We evaluate the automatic human key point detection procedure from two perspectives: the detection accuracy and the effectiveness for occupation prediction. Firstly, we use linear regression model with regularization parameter $\lambda = 1$ to train the proposed human key point detection model on the train and validation set and evaluate the effectiveness on the test set. The average error of key point prediction is shown in Table 2. For a 128×128 human upper body image, the predict errors in x and y dimensions do not exceed 10 pixels. We also show some example predictions

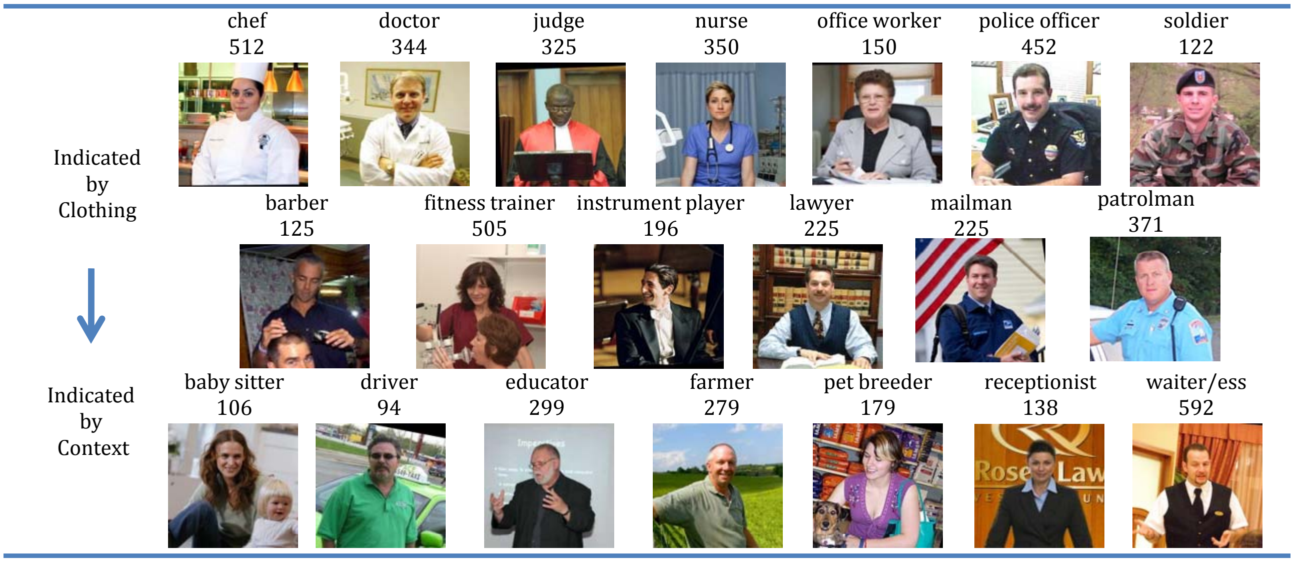


Figure 5. Some exemplary images from the collected human occupation dataset. The occupation categories are grouped by their indicative visual clue. The numbers of data of each occupation category are also listed below the occupation name.

of the four key points in Figure 6. These experiments show that the proposed human key point prediction procedure is reasonably accurate.

We also conduct the experiments to evaluate how the automatic prediction of human key points affect the occupation prediction performance. Firstly, we perform human key point detection on the test set. Samples which fail in the detection part is directly considered as incorrectly classified in occupation recognition. Actually, the human detection is near perfectly accurate on the clear frontal human of our test set.

We then conduct the experiments by extracting the proposed human clothing features using both detected and ground-truth key points on the test set and compare their prediction accuracies based on the previously learnt multi-class SVM model. The prediction accuracies of these two methods are shown in Table 1, from which we can conclude that higher accuracy can be achieved by using manually labeled key points, but for the automatic detection based human clothing descriptors, a performance gain of 8.1% can still be achieved compared with the descriptors proposed in [11].

Since the fully automatic scheme is more practical for real applications, we thus conduct further experiments based on the automatically detected key points.

4.3. Evaluate Human Clothing & Context Features

In this subsection we investigate the effectiveness of context information for the prediction of each occupation category. When predicting an occupation category, all images not belonging to this category are regarded as negative samples. Thus the prediction is implemented by a binary SVM and prediction confidence of each occupation can be obtained for the binary classification model. By varying threshold of prediction confidence, the occupation prediction performance can be measured in Average-Precision (AP) from the Precision-Recall curve. The detailed results are shown in Table 3.

From the performance of individual features, we can observe that for most occupations, the proposed human clothing features significantly outperform the BoW context features, which proves that human clothing features are more informative than the BoW context features which are founded on low-level patterns. Especially, for occupations that are considered easy to distinguish by clothing (chef, police, etc), over 50% AP measurements could be achieved by using only clothing features.

Since context features individually perform poorly for occupation prediction, we use them as the complementary information for the clothing features and the performance of the combined features is also demonstrated in Table 3. In general, the performances of most occupations are improved after combining context features with clothing features, but performance gains from the context features differ greatly from each other for these 20 occupations. We then summarize the main observations as follows:

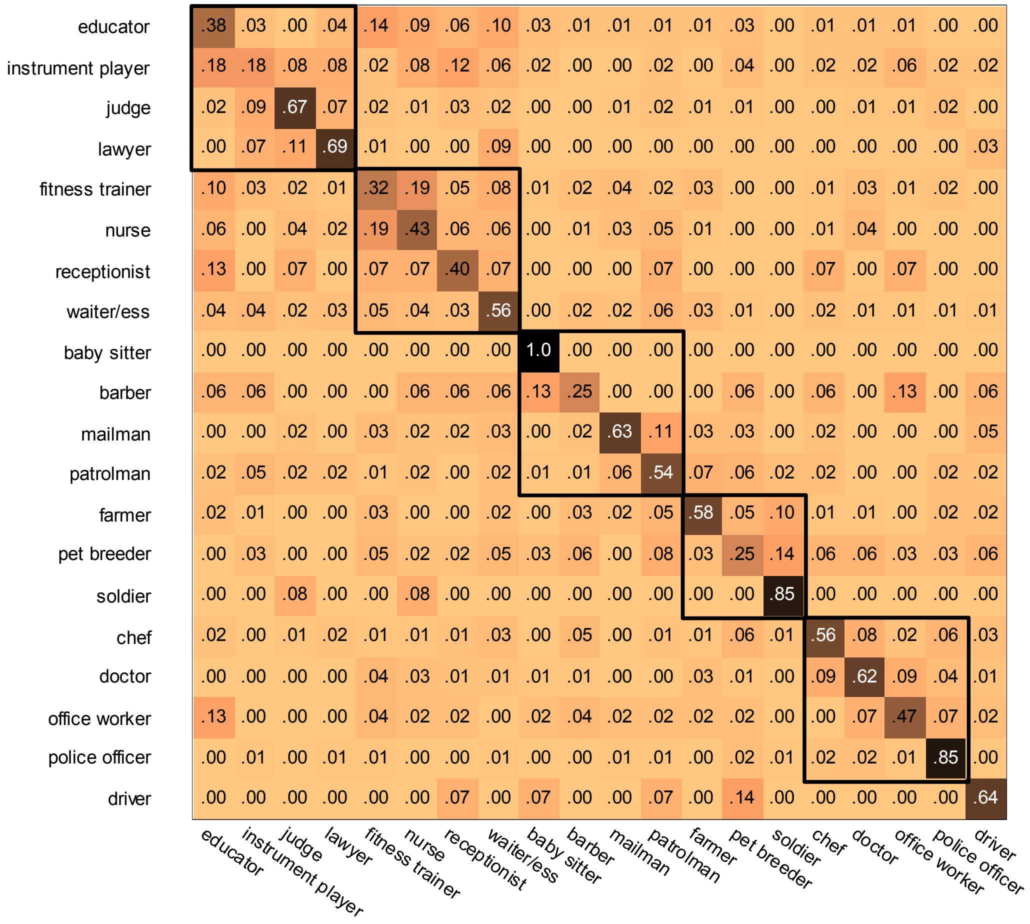
1. For occupations that have obvious clues for prediction in image background or foreground, *e.g.*, “driver”, “officer”, “educator”, “waiter”, etc., combining background and foreground context features would improve the APs by about 5% than human clothing features only. The improvement shows that the proposed two context features both contain their respective information complementary to the human clothing infor-

Table 3. The average precision(%) in predicting 20 occupations via the combination of clothing and context features under one-vs-all setting. We respectively list the performances of 1) context features only, 2) human clothing features only, 3) the combined features, and 4) the improvement of 3) over 2).

					office	police	
	chef	doctor	judge	nurse	worker	officer	soldier
Context Features	33.6	19.3	24.9	24.1	7.5	45.3	9.2
Human Clothing Features	77.4	56.8	67.6	49.0	28.9	88.5	49.3
Combined Features	79.7	58.8	68.3	50.2	32.4	89.7	50.4
Relative Improvement	2.3	2.1	0.7	1.1	3.5	1.2	1.1

		fitness	instrument				pet
	barber	trainer	player	lawyer	mailman	patrolman	breeder
Context Features	3.6	15.6	17.7	34.8	15.1	12.4	11.1
Human Clothing Features	12.7	27.8	15.5	60.2	53.8	52.2	15.9
Combined Features	12.7	30.1	20.3	66.2	55.8	52.9	16.3
Relative Improvement	0.0	2.3	4.7	6.0	2.0	0.7	0.5

	baby						mean
	sitter	driver	educator	farmer	receptionist	waiter/ess	AP
Context Features	1.6	8.4	11.8	30.7	9.0	45.0	19.0
Human Clothing Features	9.4	26.5	20.5	50.9	14.9	66.5	42.2
Combined Features	9.6	31.7	27.2	58.7	18.8	73.4	45.2
Relative Improvement	0.1	5.2	6.8	7.7	3.8	6.9	3.0



	educator	instrument player	judge	lawyer	fitness trainer	nurse	receptionist	waiter/ess	baby sitter	barber	mailman	patrolman	farmer	pet breeder	soldier	chef	doctor	office worker	police officer	driver
educator	.38	.03	.00	.04	.14	.09	.06	.10	.03	.01	.01	.01	.01	.03	.00	.01	.01	.01	.00	.00
instrument player	.18	.18	.08	.08	.02	.08	.12	.06	.02	.00	.00	.02	.00	.04	.00	.02	.02	.06	.02	.02
judge	.02	.09	.67	.07	.02	.01	.03	.02	.00	.00	.01	.02	.01	.01	.00	.00	.01	.01	.02	.00
lawyer	.00	.07	.11	.69	.01	.00	.00	.09	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.03
fitness trainer	.10	.03	.02	.01	.32	.19	.05	.08	.01	.02	.04	.02	.03	.00	.00	.01	.03	.01	.02	.00
nurse	.06	.00	.04	.02	.19	.43	.06	.06	.00	.01	.03	.05	.01	.00	.00	.01	.04	.00	.00	.00
receptionist	.13	.00	.07	.00	.07	.07	.40	.07	.00	.00	.00	.07	.00	.00	.00	.07	.00	.07	.00	.00
waiter/ess	.04	.04	.02	.03	.05	.04	.03	.56	.00	.02	.02	.06	.03	.01	.00	.02	.01	.01	.01	.01
baby sitter	.00	.00	.00	.00	.00	.00	.00	.00	1.0	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00
barber	.06	.06	.00	.00	.00	.06	.06	.06	.13	.25	.00	.00	.00	.06	.00	.06	.00	.13	.00	.06
mailman	.00	.00	.02	.00	.03	.02	.02	.03	.00	.02	.63	.11	.03	.03	.00	.02	.00	.00	.00	.05
patrolman	.02	.05	.02	.02	.01	.02	.00	.02	.01	.01	.06	.54	.07	.06	.02	.02	.00	.00	.02	.02
farmer	.02	.01	.00	.00	.03	.00	.00	.02	.00	.03	.02	.05	.58	.05	.10	.01	.01	.00	.02	.02
pet breeder	.00	.03	.00	.00	.05	.02	.02	.05	.03	.06	.00	.08	.03	.25	.14	.06	.06	.03	.03	.06
soldier	.00	.00	.08	.00	.00	.08	.00	.00	.00	.00	.00	.00	.00	.00	.85	.00	.00	.00	.00	.00
chef	.02	.00	.01	.02	.01	.01	.01	.03	.00	.05	.00	.01	.01	.06	.01	.56	.08	.02	.06	.03
doctor	.00	.00	.00	.00	.04	.03	.01	.01	.01	.01	.00	.00	.03	.01	.00	.09	.62	.09	.04	.01
office worker	.13	.00	.00	.00	.04	.02	.02	.00	.02	.04	.02	.02	.02	.02	.00	.00	.07	.47	.07	.02
police officer	.00	.01	.00	.01	.01	.00	.00	.01	.00	.00	.01	.01	.00	.02	.01	.02	.02	.01	.85	.00
driver	.00	.00	.00	.00	.00	.00	.07	.00	.07	.00	.00	.07	.00	.14	.00	.00	.00	.00	.00	.64

Figure 7. The confusion matrix of the 20-occupation prediction problem. The number of i-th row and j-th column denotes the false alarm rate to i-th class while predicting j-th class.

mation.

2. For the occupations easy to distinguish by human clothing (chef, police, etc.), adding context information would not improve the APs much, which also validate that the human clothing features are generally informative and reliable.

4.4. Confusion Matrix of Occupations

In this subsection, we further study to what extent these 20 occupations can be separated from each other. Hence we implement again the "one-vs-one" based multi-class SVM to the currently best performed features, namely the combination of human clothing features and context features. We show the confusion matrix of the multi-class classification of human occupation in Figure 7.

In this confusion matrix, we group the occupations by Spectral Clustering [2] using the normalized confusion matrix as affinity matrix, and hence occupations which tend to be misclassified will be clustered to the same group. We frame these groups in Figure 7. The demonstrated groups of occupations respectively have certain intuitive characteristics in common, *e.g.*, "chef" and "doctor" (in similar white suit), "barber" and "baby sitter" (both interacting with another human), "judge" and "lawyer" (in similar scene).

5. Conclusions and Future Work

In this paper, we investigated for the first time the human occupation prediction problem. The robust intermediate-level representations were proposed to model human clothing information based on the semantic-level descriptions derived from multi-task sparse coding method. The context features were also modeled to assist human occupation prediction. Encouraging experiment results were obtained by combining these two types of features while the proposed human clothing features show reliable performance regardless of context information. The collected large human occupation dataset provides a valuable benchmark for researchers to further study this new topic. Moreover, the proposed human clothing features are general, and we are planning to further explore the possibility to 1) utilize these features for clothing style classification, and 2) recommend proper clothes to customers by modeling the correlations between facial features and human clothing features.

Acknowledgment

This work is supported by Singapore Ministry of Education under research Grant MOE2010-T2-1-087.

References

[1] Sparse lab. http://sparselab.stanford.edu/. 3

[2] S. Becker and Z. Ghahramani. On spectral clustering: analysis and an algorithm. In *NIPS*, 2002. 8

[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009. 2

[4] C. Chang and C. Lin. LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/ cjlin/libsvm. 5

[5] H. Chen, Z. Xu, Z. Liu, and S. Zhu. Composite templates for cloth modeling and sketching. In *CVPR*, 2006. 1

[6] C. Chien and L. Wang. Color texture segmentation for clothing based on finite prolate spheroidal sequences. *Asian Journal of Health and Information Sciences*, 2007. 1

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 3

[8] S. Divvala, D. Hoiem, J. Hays, A. Efros, and M. Hebert. An empirical study of context in object detection. In *CVPR*, 2009. 2

[9] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008. 2

[10] A. Gallagher and T. Chen. Estimating age, gender, and identity using first name priors. In *CVPR*, 2008. 1

[11] A. Gallagher and T. Chen. Clothing cosegmentation for person recognition. In *CVPR*, 2008. 1, 3, 5, 6

[12] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, 2009. 2

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features, spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 3

[14] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for the multi-task lasso with applications to neural semantic basis discovery. In *ICML*, 2009. 3

[15] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 4

[16] D. Martin, C. Fowlkes, and J. Malik. Learning to detect natural image boundaries using local brightness, color, and texture cues. *TPAMI*, 2004. 3

[17] B. Ni, Z. Song, and S. Yan. Web image mining towards universal age estimator. In *ACM Multimedia*, 2009. 1

[18] T. Ojala, M. Pietikainen, and D. Harwood. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition*, 1995. 3

[19] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57:137–154, 2004. 2

[20] R. White, K. Crane, and D. Forsyth. Capturing and animating occluded cloth. In *ACM SIGGRAPH*, 2006. 1

[21] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In *CVPR*, 2010. 3

[22] B. Yao and L. Fei-Fei. Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities. in *CVPR*, 2010. 2

[23] W. Zheng, S. Gong, and T. Xiang. Quantifying contextual information for object detection. In *ICCV*, 2009. 2