

# Predicting Occupation from Images by Combining Face and Body Context Information

WEI-TA CHU, National Chung Cheng University  
CHIH-HAO CHIU, National Chung Cheng University

Facial images embed age, gender, and other rich information that are implicitly related to occupation. In this work, we advocate that occupation prediction from a single facial image is a doable computer vision problem. We extract multilevel hand-crafted features associated with locality-constrained linear coding, and convolutional neural network features, as image occupation descriptors. To avoid the curse of dimensionality and overfitting, a boost strategy called multi-channel SVM is used to integrate features from face and body. Intra-class and inter-class visual variations are jointly considered in the boosting framework to further improve performance. In the evaluation, we verify effectiveness of predicting occupation from face, and demonstrate promising performance obtained by combining face and body information. More importantly, our work further integrates deep features into the multi-channel SVM framework, and shows significantly better performance over the state of the art.

CCS Concepts: • **Computing methodologies** → *Image representations; Object recognition;*

Additional Key Words and Phrases: Occupation prediction, discriminant multi-channel SVM, adaptive weighting, body context, spatial pyramid, locality-constrained linear coding, convolutional neural network

## ACM Reference Format:

Wei-Ta Chu and Chih-Hao Chiu, 2016. Predicting Occupation from Images by Combining Face and Body Context Information. *ACM Trans. Multimedia Comput. Commun. Appl.* 0, 0, Article (October 2016), 20 pages.  
DOI: 0000001.0000001

## 1. INTRODUCTION

Predicting occupation from images emerges as an important computer vision problem because of its great potential in intelligent services and systems [Song et al. 2011]. For example, recommendation systems can more effectively and dynamically suggest news, products, or friends, to users if their occupations are known. Deeper advertising services can be developed on social platforms or expertise networks if occupations of users are considered.

Currently, related studies mainly focus on predicting occupations based on human clothing [Song et al. 2011], scene context [Song et al. 2011], and social context [Shao et al. 2013]. What people dress, where people work, and how people interact, are all important clues for occupation prediction. In this work, we advocate an alternative that may also aid occupation prediction: *facial image*. Song et al. had mentioned this in their work, but they did not finely discover facial information after their pioneering work [Song et al. 2011]. Although *predicting occupation only from faces* seems making little sense at first, we will demonstrate that it is really doable and can be a complementary approach to advance current clothing-based and context-based methods.

Studies based on face information have been widely proposed from various perspectives. For photo collection management, context-aware person identification via face was proposed to facilitate browsing and searching [O'Hare and Smeaton 2009]. With the proliferation of web-scale image collection and video sharing platforms, person identification has been extended to celebrity detection and naming [Zhang et al. 2012][Xiong et al. 2014][Pang and Ngo 2015]. As important foundation for many face-related applications, estimating facial attributes is one of the main research

---

Author's addresses: W.-T. Chu and C.-H. Chiu, Computer Science Department, National Chung Cheng University  
Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
© 2016 ACM. 1551-6857/2016/02-ART0 \$15.00  
DOI: 0000001.0000001

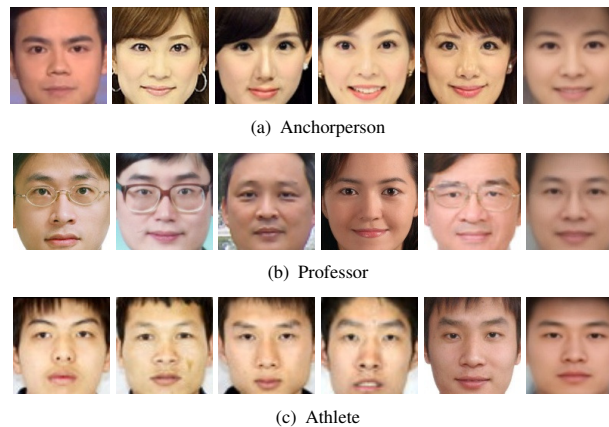


Fig. 1: Face samples of anchorperson, professor, and athlete, and their average faces (the last image of each row).

branches. For example, Chen et al. [Chen et al. 2013c] estimated facial attributes like gender and race based on community-contributed photos with feature selection schemes, and these attributes were demonstrated to be beneficial to face image retrieval [Chen et al. 2013a]. Rich attributes that cannot be explicitly stated by text can also be used to develop interesting applications, such as kin relationships in a photo [Xia et al. 2012], hirability prediction [Nguyen et al. 2014], and spontaneous smile detection [Dibeklioglu et al. 2015].

In this work, we advocate another face-related study: occupation prediction. The relationship between face and occupation can be built based on explicit facial attributes like gender and age as well as implicit characteristics described by visual features. In sociology, age patterns in different occupations have been long studied [Kaufman and Spilerman 1982][Smith 1973], and the gender employment patterns reported in [Gabriel and Schmitz 2007] clearly reveal the occupational differences between men and women. These studies show that visual attributes are related to the occupation distribution, and the reported facts enable occupation prediction from facial images.

Figure 1 shows sample face images of *anchorperson*, *professor*, and *athlete*, respectively. The average faces of each occupation are shown at the rightmost of each row. We can easily observe that anchorpersons tend to be female, professors tend to be elder, and athletes tend to be younger. This phenomenon may be caused by working environment or career culture. More facial features, such as skin color, haircut, and glasses wearing, can also be found by more analysis. We conjecture that an occupation can be viewed as a joint distribution over a set of face attributes as well as body attributes, and thus a computational model can be built to predict occupation from images.

Figure 2 shows the framework of the proposed occupation prediction system, which consists of four components: data preprocessing, feature extraction, discriminant multi-channel SVM, and prediction fusion. In the data preprocessing component, we divide each upper body image into the face part and the body part. We then extract features from three perspectives: low-level features, high-level attributes, and deep learning features, respectively. We will compare the hand-crafted features (the former two) and the learnt features (the last) in the evaluation.

To integrate features extracted from different parts, the multi-channel SVM framework [Chen et al. 2013b]<sup>1</sup> based on the boosting strategy is adopted to train the SVM classifier for occupation prediction. Different from the multi-channel SVM (MC-SVM) framework where a large number of

<sup>1</sup>In [Chen et al. 2013b], their framework is called a multi-feature SVM, which means that features are extracted from multiple regions, and features from each region are viewed representing a channel. To avoid confusion, we rename the framework as multi-channel SVM in this paper, in order to more clearly indicate that features are extracted from multiple parts, rather than multiple features showing different visual perspectives.

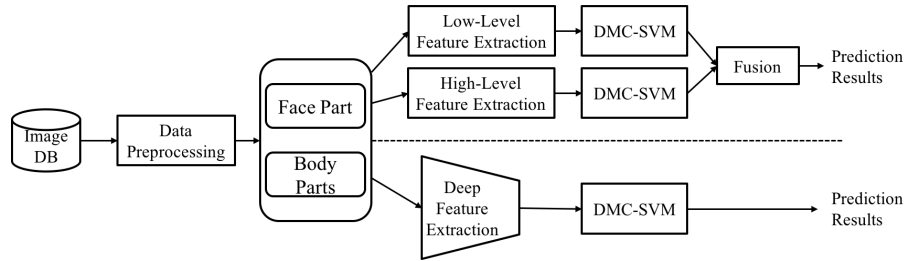


Fig. 2: Framework of the proposed occupation prediction system.

1-vs-1 SVM classifiers were trained, we construct a multiclass SVM in the boosting framework, and further consider the characteristics of intra-class variation and inter-class variation to constitute the proposed discriminant multi-channel SVM (DMC-SVM). From the perspective of hand-crafted features, the fusion component is to integrate prediction results obtained based on low-level features and high-level attributes. From the perspective of deep learning features, the multiclass SVM directly outputs prediction results.

Our contributions are threefold. First, although the head region has been considered in occupation prediction [Song et al. 2011], we more finely investigate the clues conveyed by the face region, such as age and gender. We verify effectiveness of this facial information, and construct a computational model to integrate multi-channel features from the facial part and the body part, with the consideration of inter-/intra-class characteristics. Second, to our best knowledge, our work is the first one to employ deep learning features in occupation prediction. By integrating such features into the multi-channel framework, we verify the effectiveness of this approach. Third, we collect two image datasets as the foundation for future research. One of which consists of frontal faces with occupation information, and another consists of upper body with occupation information.

Note that we are not claiming that occupation can be predicted solely based on facial information. The most important idea is that, by further considering facial information, occupation can be more accurately predicted. In [Chen et al. 2013b], first names can be roughly estimated solely based on face images, also because of sociological trends. Motivated by this work, we think that occupation may be influenced by more sociological and psychological factors, and thus integrate multiple clues in a computational model.

The rest of this paper is organized as follows. Section 2 provides literature survey about the relationship between occupation and facial information, and the state of the arts. Section 3 provides details of preprocessing, and feature extraction from images. Details of the proposed discriminative multi-channel support vector machine is described in Section 4. Comprehensive evaluation is reported in Section 5, followed by the conclusion given in Section 6.

## 2. RELATED WORKS

Human face is commonly viewed to convey rich information, including age, gender, ethnicity and race, emotional state, honesty and deception, and personal identity. The characteristic structures and expressions perhaps make face the most important anatomical subject of mythology, religion, art, and literature. We advocate that occupation can be more accurately predicted if facial features are considered. To settle the foundation of the proposed idea, we resort to researches of sociology and official statistics in Section 2.1. In Section 2.2, we introduce the state-of-the-art occupation prediction works.

### 2.1. Occupation and Faces

Todorov et al. [Todorov et al. 2005] showed that inferences of candidate's competence based on facial appearance can be quickly made by human, and the perceived competence can be used to predict outcomes of elections better than chance. Antonakis and Dalgas [Antonakis and Dalgas 2009] even reported that the abilities of adults and children to predict election results are indistinguish-

able. In a game involving a simulated trip from Troy to Ithaca, the participated children (aged 5 to 13 years) were asked to chose one of the given two faces (who were actually two candidates of another country's election) to be the captain of their boat. Results of logistic regression showed the probability of predicting election results correctly was 0.71. Rule and Ambady [Rule and Ambady 2008] found that chief executive officers (CEO) from more successful companies versus less successful ones can be distinguished by naive users based solely on CEOs' facial appearance. These interesting psychological studies reveal that faces convey subtle but rich information, and human beings have amazing ability to capture it at a glance.

The relationship between face and occupation can be seen from several statistics and studies. Human's temperament and cognitive abilities are affected by gender differences, and this largely influences how a person selects his/her job. More specifically, competitiveness, risk, interest in children, mechanical ability, and verbal ability are all factors related to job selection [Browne 2006]. The gender employment patterns reported in [Gabriel and Schmitz 2007] show the occupational differences between men and women in US. Gender, therefore, is clearly related to occupation, also shown in Figure 1. Recognizing gender from facial images has been a well-known computer vision task for years. Occupation and face images are thus linked by gender from some implicit perspective.

Age is clearly related to job selection. People who are in the occupation needing higher physical strength, e.g., athlete, are usually younger, and people who are in the occupation needing more life experience or specialized knowledge, e.g., professor, are usually elder. In sociology, age patterns in different occupations have been long studied [Kaufman and Spilerman 1982][Smith 1973]. Kaufman and Spilerman [Kaufman and Spilerman 1982] concluded that "systematic forces of an institutional and a demographic nature operate on occupations and are capable of creating a diversity of age patterns". Occupation and face images are thus implicitly linked by age.

In addition to gender and age, other facial features, such as styles of haircut and hat [Song et al. 2011], skin color, and wearing glasses, may also link with occupation.

## 2.2. Predicting Occupation from Images

Song et al. [Song et al. 2011] proposed the first work to predict occupation from images based on clothing and context information. A part-based appearance model was adopted to detect parts of human upper body, which were then described by low-level features with sparse coding to derive semantic-level representation. Key points on human body and background were extracted as context information. They demonstrated that human clothing features were much more promising than context features, while combining both types of features yielded higher accuracy. In Song's work, the upper body images were used, which consist of the head region and of course the face region. They more focused on "top of head" in order to catch the characteristics of haircut and hat style. In our work, we finely consider information from face, i.e., the region from eyebrows to chin, and investigate how facial information influences performance of occupation prediction.

Shao et al. [Shao et al. 2013] focused on recognizing occupations of multiple people with arbitrary poses in an image. In addition to visual appearance, co-occurrence and spatial configuration were jointly modeled by a structure support vector machine. This work pushed occupation recognition to a more general case by considering multiple people with pose variation and various interaction.

In our previous work [Chu and Chiu 2014], we proposed to predict occupation solely from facial images, and verified that this should be an interesting and effective way to conduct occupation prediction. In this work, we extend the discriminant multi-channel SVM framework to integrate face and body information, employ deep learning features as image descriptor, and demonstrate that performance better than the state of the art can be obtained.

Relevant to occupation recognition, social role/group discovery emerged recently [Ramanathan et al. 2013][Kwak et al. 2013]. By analyzing human interaction or how a person interacts with the environment, social roles can be recognized, and such results can be used to facilitate image/video understanding such as multimodal event detection [Ramanathan et al. 2013]. Kwak et al. [Kwak

et al. 2013] studied the relationship between individual’s social identity and social groups, and investigate group descriptors to facilitate a novel vision task, i.e., social categorization.

### 3. FEATURE EXTRACTION

#### 3.1. Preprocessing

As we mentioned in Figure 2, we consider upper body images and divide them into the face part and the body part. The face part can be seen as the *face region*, and the body part is further divided into four *clothing regions*. In [Song et al. 2011], four regions of a human upper body are defined to describe clothing information, i.e., top of head, central upper body, left shoulder and right shoulder. In order to verify that integrating face information with clothing information is helpful to occupation prediction, we focus on the following five regions:

- Face ( $R_1$ ): The face region contains rich high-level semantic features, such as age, gender, wearing glasses, etc. We advocate that different occupations have distinct attribute distributions and thus can be discriminated.
- Top of head ( $R_2$ ): This region usually shows hat or hairstyle. Hairstyle may be different because of the gender or working environment. Hat is also an important attribute to discriminate different occupations, e.g., police officer, firefighter, and chef.
- Central upper body ( $R_3$ ): People of many occupations wear uniforms. This region shows important clothing information like collar and clothing texture.
- Left and right shoulders ( $R_4$  and  $R_5$ ): These two regions correspond to the wrist, arm, and shoulder parts of human dressing, and also convey clothing information.

Figure 3 shows samples of the five regions. To automatically detect these regions, we first calculate relative distances between these regions from a randomly selected subset of our evaluation database. For images in this subset, we detect face and upper body bounding boxes by the method proposed in [Viola and Jones 2004]. The face region is normalized to  $64 \times 64$  pixels, with detected eyes fixed at specific positions. The detected upper body region is also normalized according to the ratio of face normalization. We then manually label central points of the face region and four body regions. We calculate the distance between central points of  $R_1$  and  $R_2$ , the one between  $R_1$  and  $R_3$ , the one between  $R_3$  and  $R_4$ , the one between  $R_3$  and  $R_5$ , and denote them as  $d_1$ ,  $d_2$ ,  $d_3$ , and  $d_4$  shown in Figure 3, respectively. The ratio of  $d_1$ , for example, to the height of the face bounding box is then calculated, denoted as  $r_1$ . Finally, we obtain the mean ratio  $\bar{r}_1$  by averaging all  $r_1$  calculated from the selected subset.

Given a new image, we first detect its face bounding box and locate its central point. According to the distance ratios  $\bar{r}_1$ ,  $\bar{r}_2$ ,  $\bar{r}_3$ , and  $\bar{r}_4$  learnt from the aforementioned process, central points of four body regions are estimated. These central points are used to expand their corresponding body regions. For example, suppose that the location of the central point of  $R_3$  is estimated as  $(x_3, y_3)$ , the coordinates of the upper-left corner  $cl_3$  and the lower-right corner  $cr_3$  of the  $R_3$  region are set as  $cl_3 = (x_3 - 32, y_3 - 50)$  and  $cr_3 = (x_3 + 32, y_3 + 50)$ , respectively. This forms a  $64 \times 100$  pixels region to represent  $R_3$ . Other regions are expanded in the same way, with  $64 \times 120$  pixels for the left and right shoulders regions, and  $32 \times 64$  pixels for the top of head region. Intensity histogram equalization is then applied to all these regions to eliminate the influence of lighting variation.

#### 3.2. Hand-Crafted Feature Extraction

For each region shown in Figure 3, we extract low-level visual features followed by feature coding. Specially for the face region ( $R_1$ ), we further detect high-level attributes related to occupation.

**3.2.1. Low-Level Feature Extraction and Coding.** Low-level features are extracted in a dense sampling manner. We divide each image region into  $16 \times 16$ -pixels *grids* with 2-pixel strides, and extract the features shown in Table I from each grid. Dense Scale Invariant Feature Transform (SIFT) [Liu et al. 2008] descriptors are extracted to represent the face region since it is invariant to image scale and rotation, and is also robust to illumination change. We also extract HSV color moments from

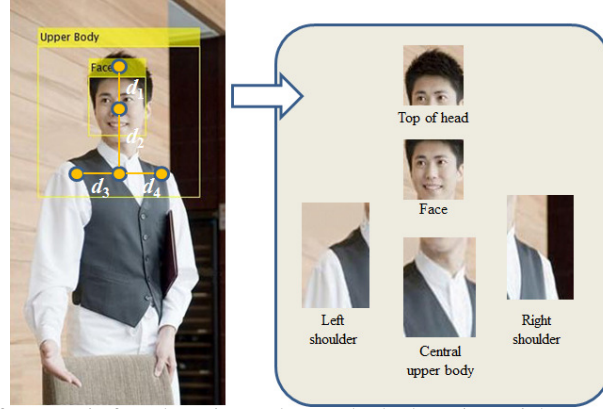


Fig. 3: Left: Results of automatic face detection and upper body detection. Right: Regions of face ( $R_1$ ), top of head ( $R_2$ ), central upper body ( $R_3$ ), left shoulder ( $R_4$ ), and right shoulder ( $R_5$ ).

Table I: The extracted low-level features and their corresponding dimensions.

Regions	Features	Dim.
Face region ( $R_1$ )	Dense SIFT descriptor	128
	HSV color moments	9
Four body regions ( $R_2 - R_5$ )	Histogram of Oriented Gradient (HOG)	128
	Local Binary Patterns (LBP)	196
	Color histogram in the CIELAB space	768
	Histogram of color gradient	768
	Histogram of texture gradient	256

the face region to describe skin color. As suggested in [Song et al. 2011], we extract five features to represent the four body regions, i.e., the Histogram of Oriented Gradient (HOG) [Dalal and Triggs 2005], Local Binary Patterns (LBP) [Ojala et al. 2002], color histogram in the CIELAB color space, and histogram of color/texture gradient [Martin et al. 2004]. These features characterize changes in color, brightness, and texture. They are also commonly used in clothing retrieval [Liu et al. 2012].


We would like to encode low-level features to more effectively describe an image region. The bag-of-visual-words model and sparse coding are widely-used encoding schemes, while the former suffers from the hard quantization problem, and the latter does not guarantee that similar features are transformed into similar codes. Locality-constrained linear coding (LLC) [Wang et al. 2010] was proposed to jointly consider locality and sparsity properties, and was demonstrated to outperform bag of visual words and sparse coding. We thus adopt LLC associated with the spatial pyramid scheme [Lazebnik et al. 2006] to generate an image descriptor.

For a given image region, we first divide it into  $1 \times 1$ ,  $2 \times 2$ , and  $4 \times 4$  spatial sub-regions, respectively. Given an image sub-region, we sample dense grids and extract features such as SIFT from each grid. Let  $X = \{x_1, \dots, x_M\} \in \mathbb{R}^{D \times M}$  be a set of  $D$ -dimensional SIFT descriptors ( $D=128$ ) extracted from  $M$  dense grids. Given an LLC codebook with  $K$  codewords ( $K=1024$ ),  $B = \{b_1, \dots, b_K\} \in \mathbb{R}^{D \times K}$ , the LLC code can be derived by:

$$\min_C \sum_{i=1}^M \|x_i - Bc_i\|^2 + \lambda \|d_i \odot c_i\|^2 \quad (1)$$

$$s.t. \quad \mathbf{1}^T c_i = 1, \forall i$$

where  $C = \{c_1, \dots, c_M\}$  is the set of coefficients indicating how codewords are combined to reconstruct  $X$  with minimum error. The operator  $\odot$  means the element-wise multiplication, and



age	32	44	27
gender	Female	Male	Male
glass	None	None	Normal

Fig. 4: High-level attributes of three sample face images.

$d_i \in \mathbb{R}^K$  is the distance vector with each entry as of the Euclidean distance between  $x_i$  and  $b_j$ ,  $j = 1, \dots, K$ . By solving this optimization problem, each SIFT descriptor  $x_i$  is converted into an LLC code  $c_i \in \mathbb{R}^K$ .

For each sub-region, the LLC codes of a grid are integrated using max pooling. These pooled LLC codes from all 21 sub-regions are then concatenated to describe the image region. With the process mentioned above, an image region is finally described as a vector of 21 (sub-regions)  $\times$  1024 (codebook size  $K$ ) = 21, 504 dimensions.

**3.2.2. High-Level Attributes.** High-level information can be informative in determining occupations. In this work, we adopt the Face++ API [Fan et al. 2014]<sup>2</sup> to estimate age, gender, and the condition of wearing glasses from the face region. Figure 4 shows prediction results of three sample face images. As can be seen, high-level attributes can be effectively estimated, e.g., the rightmost face can be known to be a 27 years old male wearing glasses.

We evaluate performance of high-level attribute estimation in a preliminary experiment. We manually define ground truths of age, gender, and glass for 100 images randomly selected from the first collected dataset (Sec. 5). Overall, accuracy of gender estimation is 96.94%, and accuracy of wearing glasses or not is 98.98%. For age estimation, we group every 10 years as a class, i.e., 0–9 years old, 10–19 years old, and so on, and evaluate whether the estimated age for a given image falls into the class it belongs to. Overall the average accuracy is 22.45%, which is much worse than gender estimation and wearing glasses estimation. But in Sec. 5 we will show that this not-so-accurate age information is still helpful in occupation prediction.

### 3.3. Deep Learning Feature Extraction

We notice that currently features learnt based on convolutional neural network (CNN) largely surpass hand-crafted features in many image classification works [LeCun et al. 1998][Krizhevsky et al. 2012][Razavian et al. 2014]. To evaluate whether CNN features are effective in occupation prediction, we extract CNN features as the descriptor to describe each image region.

We utilize the MatConvNet package [Vedaldi and Lenc 2015] with the pre-trained vgg-f model [Chatfield et al. 2014] to extract CNN features from each image region ( $R_1$  to  $R_5$ ). There are five convolutional layers and three fully-connected layers in the CNN model. We try to take output of the fifth, sixth, and seventh (fully-connected) layers to be CNN features, and found that features from the sixth layer yield the best performance through our preliminary experiments. Therefore, each image region is finally described by a 4096-dimensional vector.

## 4. OCCUPATION PREDICTION

Based on the descriptors extracted from a collection of face regions, for example, we can construct a classifier to predict occupation from face. The classification results obtained from classifiers trained from the five regions (face and four body regions) can then be fused to give the final prediction

<sup>2</sup><http://www.faceplusplus.com/>

result. However, dimensionality of concatenated LLC descriptors is high, and the training process is susceptible to overfitting and the curse of dimensionality. In [Chen et al. 2013b], LLC descriptors extracted from 21 sub-regions are viewed as features extracted from different feature channels. Twenty-one classifiers are separately constructed, and are integrated by the proposed multi-channel SVM (MC-SVM) framework. From this viewpoint, each descriptor has much less dimensionality. Note that “multi-channel” here means integrating features from multiple channels (sub-regions), rather than integrating different features, e.g., SIFT and HOG.

#### 4.1. Multi-Channel SVM

We adopt the MC-SVM framework to integrate features extracted from sub-regions. Algorithm 1 shows the training process. Let us take the face region as an instance in the following explanation. Suppose we have  $N$  face regions from  $N$  training images, and each face region  $x$  is represented by  $T$  feature channels (sub-regions) and is with a class label (occupation)  $y$ . We denote  $\mathbf{x}_{t,i}$  as the  $t$ -th feature channel extracted from the  $i$ -th training face region, where  $t = 1, \dots, T$ , and  $i = 1, \dots, N$ . First, with the equal weights  $D_i$  for all training face regions, we use the first feature channel ( $t = 1$ ) to construct an SVM classifier with the five-fold cross validation scheme (line 4). We thus can calculate the prediction confidence  $f_t(\mathbf{x}_{t,i})$  as well as classification error  $err_t$  in the representation of the ratio of incorrect prediction. The error term  $err_t$  plays an important role to dynamically adjust classifier weight  $\alpha_t$  (line 5) and image weight  $D_i$  (line 6). Intuitively, if the  $i$ -th face region is misclassified by the current SVM (based on the first feature channel), the image weight  $D_i$  gets larger when we train the SVM based on the second feature channel, while the classifier weight  $\alpha_i$  gets smaller. The whole procedure is repeated until all feature channels have been trained. The algorithm finally outputs a set of classifiers  $f_t$ ,  $t = 1, \dots, T$ , specific to the  $t$ -th feature channel and the corresponding classifier weights  $\alpha_t$ . Given a test face region  $q$ , the extracted  $T$  feature vectors  $\mathbf{q}_1, \dots, \mathbf{q}_T$  are fed to the  $T$  classifiers, respectively, and the final prediction confidence is  $\sum_{t=1}^T \alpha_t f_t(\mathbf{q}_t)$ , with the prediction result as  $sign(\sum_{t=1}^T \alpha_t f_t(\mathbf{q}_t))$ .

---

#### ALGORITHM 1: Training Process of Multi-Channel SVM

---

**Input:** Training features  $\mathbf{x}_{t,i}$  and training labels  $y_i \in \{c_1, \dots, c_M\}$ , where  $t = 1, \dots, T$  and  $i = 1, \dots, N$ .

**Output:** SVM classifiers  $f_t$  and classifier weights  $\alpha_t$

- 1 Initialize weights of training images  $D = \{D_1, \dots, D_N\}$ ,  $D_i = 1$ ,  $i = 1, \dots, N$ .
  - 2 **for**  $t=1$  to  $T$  **do**
  - 3     Train SVM  $f_t$  using  $D$ .
  - 4     Using weights  $D$ , perform SVM cross validation to obtain confidence  $f_t(\mathbf{x}_{t,i})$  and prediction  $\hat{y}_{t,i} = sign(f_t(\mathbf{x}_{t,i}))$ . Compute error  $err_t = \frac{\sum_{i=1}^N |\hat{y}_{t,i} \neq y_i|}{N}$ .
  - 5     Compute  $\alpha_t = \log(\frac{1-err_t}{err_t})$ .
  - 6     Set  $D_i = D_i \exp(-\alpha_t y_i f_t(\mathbf{x}_{t,i}))$  and renormalize so that  $\sum_{i=1}^N D_i = N$ .
  - 7 **end**
  - 8 Output classifiers  $f_t$  and classifier weights  $\alpha_t$ .
- 

#### 4.2. Discriminant Multi-Channel SVM

The main idea of Algorithm 1 is embedded at the weight update step shown in line 5. Here we propose an improvement to more deeply adjust classifier weights. Considering that a good classification system should categorize entities of the same class together and discriminate entities of different classes as far as possible, we take the ratio of inter-class variation to intra-class variation into account to update  $\alpha_t$ . The improved algorithm is called *Discriminative Multi-Channel SVM* (DMC-SVM, shown in Algorithm 2), which mainly differs Algorithm 1 in line 5.

Let  $\mathbf{x}^{(i)}$  denote the  $i$ -th face region with the label  $y_i$ , and let  $\mathcal{C}_j = \{\mathbf{x}^{(i)} : y_i = j\}$  denote the set of face regions with the label  $j$ . The inter-class variation  $dr(y_i, t)$  is calculated as



**ALGORITHM 2:** Training Process of Discriminant Multi-Channel SVM.

**Input:** Training features  $\mathbf{x}_{t,i}$  and training labels  $y_i \in \{c_1, \dots, c_M\}$ , where  $t = 1, \dots, T$  and  $i = 1, \dots, N$ .

**Output:** SVM classifiers  $f_t$  and classifier weights  $\alpha_t$

1 Initialize weights of training images  $D = \{D_1, \dots, D_N\}$ ,  $D_i = 1, i = 1, \dots, N$ .

2 **for**  $t=1$  to  $T$  **do**

3     Train SVM  $f_t$  using  $D$ .

4     Using weights  $D$ , perform SVM cross validation to obtain confidence  $f_t(\mathbf{x}_{t,i})$  and prediction

$$\hat{y}_{t,i} = \text{sign}(f_t(\mathbf{x}_{t,i})). \text{ Compute error } err_t = \frac{\sum_{i=1}^N |\hat{y}_{t,i} \neq y_i|}{N}.$$

5     Compute  $\alpha_t = w_1 \log(\frac{1-err_t}{err_t}) + w_2 \frac{dr(y_i, t)}{da(y_i, t)}$ .

6     Set  $D_i = D_i \exp(-\alpha_t y_i f_t(\mathbf{x}_{t,i}))$  and renormalize so that  $\sum_{i=1}^N D_i = N$

7 **end**

8 Output classifiers  $f_t$  and classifier weights  $\alpha_t$ .

$$dr(y_i, t) = \frac{1}{Z_i} \sum_{\substack{p,q \\ \mathbf{x}^{(p)} \in \mathcal{C}_{y_i}, \mathbf{x}^{(q)} \notin \mathcal{C}_{y_i}}}^{|\mathcal{C}_{y_i}|} d(\mathbf{x}_{t,p}, \mathbf{x}_{t,q}), \quad (2)$$

$$\hat{dr}(y_i, t) = \frac{dr(y_i, t)}{dr_{var}(y_i, t)}, \quad (3)$$

where  $Z_i$  is a normalization factor, and  $d(\mathbf{x}_{t,p}, \mathbf{x}_{t,q})$  is the Euclidean distance between face regions  $\mathbf{x}^{(p)}$  and  $\mathbf{x}^{(q)}$ , based on the  $t$ -th feature channel. The value  $dr(y_i, t)$  is thus the average Euclidean distance between face regions belonging to different classes. The value  $\hat{dr}(y_i, t)$  is the variance of the inter-class distance distribution.

On the other hand, the intra-class variation  $da(y_i, t)$  is defined as the average Euclidean distance between face regions within the same class, and is calculated as

$$da(y_i, t) = \frac{1}{Z'_i} \sum_{\substack{p,q \\ \mathbf{x}^{(p)} \in \mathcal{C}_{y_i}, \mathbf{x}^{(q)} \in \mathcal{C}_{y_i}}}^{|\mathcal{C}_{y_i}|} d(\mathbf{x}_{t,p}, \mathbf{x}_{t,q}), \quad (4)$$

$$\hat{da}(y_i, t) = \frac{da(y_i, t)}{da_{var}(y_i, t)}. \quad (5)$$

where  $Z'_i$  is a normalization factor, and the value  $\hat{da}(y_i, t)$  is the variance of the intra-class distance distribution, based on the  $t$ -th feature channel.

At line 5 of Algorithm 2, classification error  $err_t$  and the ratio of inter-class variation to intra-class variation are jointly considered to update the classifier weight  $\alpha_t$ . If for some feature channel the ratio is larger, the corresponding classifier weight also gets larger. In this work, the weights  $w_1$  and  $w_2$  are both set as  $\frac{1}{2}$ .

Figure 5 shows the intra-class distance distributions of athlete's facial images, and the inter-class distance distributions between athlete and policeman, derived from the first four feature channels of the face region. As can be seen, although there is overlap, the intra-class distribution is distinct from the inter-class distribution. The intra- and inter-class distributions between other occupations also present this distinction. Figure 6 shows the intra- and inter-class distance distributions between doctor and anchorperson, and between athlete and professor. These characteristics give clues to update the classifier weight  $\alpha_t$  and yield better performance that will be described later.

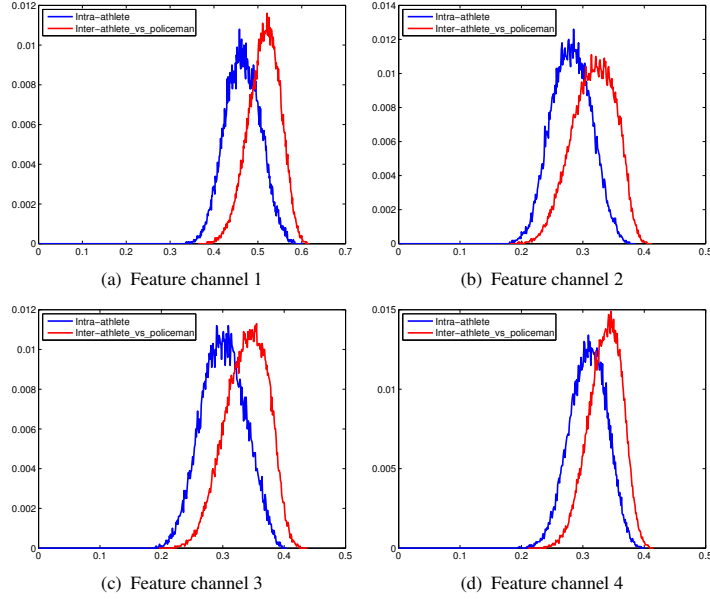


Fig. 5: Samples of inter-class distance distributions between athlete and policeman (red), and intra-class distance distributions of athlete (blue), derived from four different feature channels.

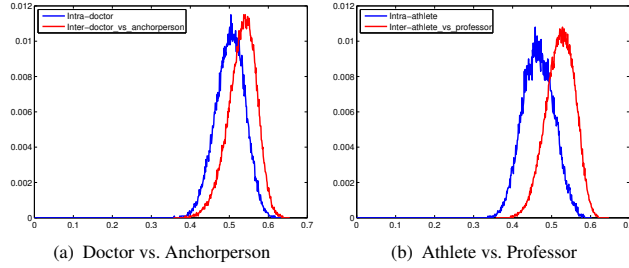


Fig. 6: Samples of intra- and inter-class distance distributions between doctor and anchorperson (left), and between athlete and professor (right).

### 4.3. Prediction Fusion

Note that the aforementioned description of DMC-SVM is trained for only one image region based on one descriptor, e.g., dense SIFT descriptors extracted from the face region. According to Table I, we can also train a DMC-SVM for the face region based on HSV color moments. Let these two DMC-SVMs denote as  $\mathcal{M}_1$  and  $\mathcal{M}_2$ , respectively. Similarly, for each of the body region, we can train a DMC-SVM based on HOG, LBP, color histogram, color gradient, and texture gradient, respectively, and denote them as  $\mathcal{M}_3, \dots, \mathcal{M}_7$  (for the top of head region),  $\mathcal{M}_8, \dots, \mathcal{M}_{12}$  (for the central upper body region),  $\mathcal{M}_{13}, \dots, \mathcal{M}_{17}$  (for the left shoulder region), and  $\mathcal{M}_{18}, \dots, \mathcal{M}_{22}$  (for the right shoulder region). In addition to low-level features, we especially extract high-level attributes including age, gender, and wearing glasses to describe the face region. These three face attributes can also be seen as different feature channels, and we can construct a (high-level) DMC-SVM  $\mathcal{M}_{23}$ .

To combine the classification results obtained from these DMC-SVMs, we adaptively calculate the weight  $\beta_i$  for  $\mathcal{M}_i$  by considering its classification error derived from the training set. The weight

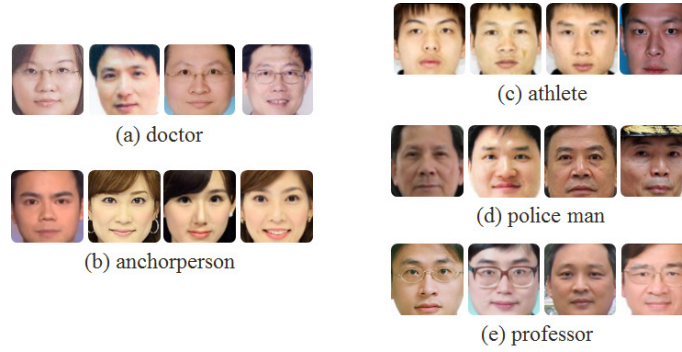


Fig. 7: Sample images of DB1. There are five different occupations and 2,062 images in total.

$\beta_i$  is calculated as  $\beta_i = \log\left(\frac{1-ERR_i}{ERR_i}\right)$ , where  $ERR_i$  is the classification error of the model  $\mathcal{M}_i$ . Given an upper body image consisting of the face and four body regions  $R_1, \dots, R_5$ , the integrated prediction confidence is obtained by

$$\begin{aligned}
 p_k = & \sum_{i=1}^2 \beta_i \mathcal{M}_i(R_1) + \sum_{i=3}^7 \beta_i \mathcal{M}_i(R_2) + \sum_{i=8}^{12} \beta_i \mathcal{M}_i(R_3) \\
 & + \sum_{i=13}^{17} \beta_i \mathcal{M}_i(R_4) + \sum_{i=18}^{22} \beta_i \mathcal{M}_i(R_5) + \beta_{23} \mathcal{M}_{23}(R_1),
 \end{aligned} \tag{6}$$

where  $\mathcal{M}_i(R_j)$  is, for the region  $R_j$ , the prediction confidence of the occupation  $k$  given by the DMC-SVM  $\mathcal{M}_i$ . The given image is claimed to belong to the occupation  $k^*$  if  $k^* = \arg \max_k p_k$ . Similar prediction process is also conducted for prediction results obtained by DMC-SVMs trained based on deep learning features.

## 5. EVALUATION

### 5.1. Databases

Because previous works [Song et al. 2011][Shao et al. 2013] didn't put their datasets in public, we collect two databases by ourselves to verify the proposed method. For the first database (DB1), we focus only on frontal face images that will be used to verify the effectiveness of DMC-SVM solely based on face. Images of DB1 were downloaded from official web sites of some organizations in Asia, such as hospitals, universities, and TV channels. From official web sites, reliability and quality of images in DB1 are guaranteed. We exclude ethnic variations by focusing on Eastern Asian people in DB1. Figure 7 shows sample images, where totally 2,062 images belonging to five different occupations, i.e., doctor, anchorperson, athlete, policeman, and professor, are included. The number of images of each occupation ranges from 300 to 500.

For the second database (DB2), we target at verifying that occupation prediction by combining face and body context information can achieve better performance. We collect images from two popular image search engines, Google Images and Bing Images, by text queries related to occupations. Following the selection criteria mentioned in [Song et al. 2011] and [Shao et al. 2013], twenty-one occupations are selected from over 100 well-defined occupations in Wikipedia. Figure 8 shows sample upper body images of each occupation. DB2 has ethnic diversity and high intra-class variations, making it more challenging than DB1. It contains 5,671 images in total, and the number of images of each occupation ranges from 122 to 553.



Fig. 8: Sample images of DB2. Inside the parentheses are the numbers of images of each occupation. There are twenty-one different occupations and 5,671 images in total.

Table II: Average prediction accuracy of different methods, based on low-level hand-crafted features extracted from DB1.

	Random	SVM	MC-SVM	DMC-SVM
Avg. accuracy	20.00	67.81	71.61	72.89

## 5.2. Occupation Prediction from Face

Based on DB1, for each occupation 240 images were randomly selected for training, and the remaining images were for testing. The random split scheme was adopted for five times to train and test the constructed DMC-SVMs.

The average prediction accuracy based on low-level hand-crafted features is shown in Table II. We compare performance obtained by conventional SVM, MC-SVM, and DMC-SVM. In conventional SVM, classifier weight  $\alpha_t$  and image weight  $D_i$  are set as unity always (without any updating). In MC-SVM, the classifier weight  $\alpha_t$  is updated as described in Algorithm 1, while in DMC-SVM, the classifier weight is updated as described in Algorithm 2. Table II shows that encouraging performance can be obtained when occupation is predicted based solely on face information. Performance superiority of the MC-SVM scheme over the SVM scheme shows effectiveness of classifier weighting and data weighting. With improved weighting and normalization by distance variance, the DMC-SVM prediction model yields the best performance.

Combining low-level features and high-level attributes further improves performance. Table III shows average accuracies obtained by conventional SVM based on single high-level attributes (the top row), average accuracies obtained by considering low-level features and high-level attributes in the DMC-SVM framework (the middle row), and average accuracy obtained based on deep learning features and the DMC-SVM (the bottom row). We see that using a single high-level attribute does not yield satisfactory performance. On the other hand, combining multiple low-level features by DMC-SVM yields much higher accuracy (72.89%) than using high-level attributes only by SVM (e.g., 25.45% based on gender information). Combining all high-level attributes by DMC-SVM largely improves performance over single high-level attributes, i.e., 40.65% vs. 25.45%, and more performance improvement can be achieved if low-level hand-crafted features and high-level attributes are jointly considered (73.18%). The last row of Table III shows that the DMC-SVM framework constructed based on deep learning features yields better performance than low-/high-level hand-crafted features (77.50% vs. 73.18%). This result is consistent with the flourishing deep learning research trend.

Table IV shows the confusion matrix of occupation prediction based on deep learning features. The prediction accuracy of athlete is quite high, probably because of its uniqueness in gender and age (mainly young male). Figure 9 shows age distributions of five different occupations, where we see very distinct distributions from athlete (90% of athletes are from 20 to 30) and anchorperson

Table III: Average prediction accuracies based on different levels of features and deep leaning features, for DB1.

Methods & Features	Average accuracy
Age (SVM)	23.01
Gender (SVM)	25.45
Glasses (SVM)	24.41
Low-level (DMC-SVM)	72.89
High-level (DMC-SVM)	40.65
Low+High (DMC-SVM)	<b>73.18</b>
Deep learning features (DMC-SVM)	<b>77.50</b>

 Table IV: Confusion matrix of occupation prediction based on deep learning features, for DB1. The  $(i, j)$ -th entry indicates the ratio that images of the  $i$ -th occupation are classified as the  $j$ -th occupation.

	Doctor	Anchorperson	Athlete	Policeman	Professor
Doctor	<b>75.36</b>	3.44	0.56	12.56	8.08
Anchorperson	1.47	<b>85.31</b>	1.39	2.70	9.14
Athlete	0.38	0.77	<b>97.69</b>	1.15	0.00
Policeman	18.46	1.23	1.54	<b>65.54</b>	13.23
Professor	10.96	12.48	0.72	12.24	<b>63.60</b>

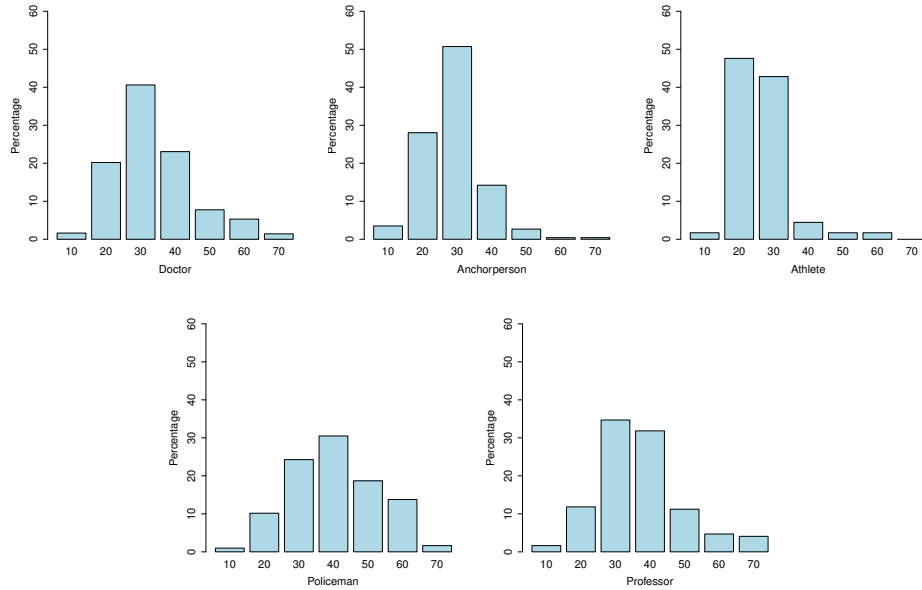


Fig. 9: Age distributions of five occupations in DB1.

(more than half are around 30), and professor and doctor have relatively similar distributions. We can confirm these trends by observing that prediction accuracy is relatively higher for athlete and anchorperson.

Figure 10 shows samples that are correctly classified (left part) and falsely classified (right part). The text corresponding to each image in the right part shows the falsely predicted class and is shown in italic. From this figure we can see that some images are confusing even for human beings, e.g., doctor vs. professor. More facial features would be needed to improve prediction performance.

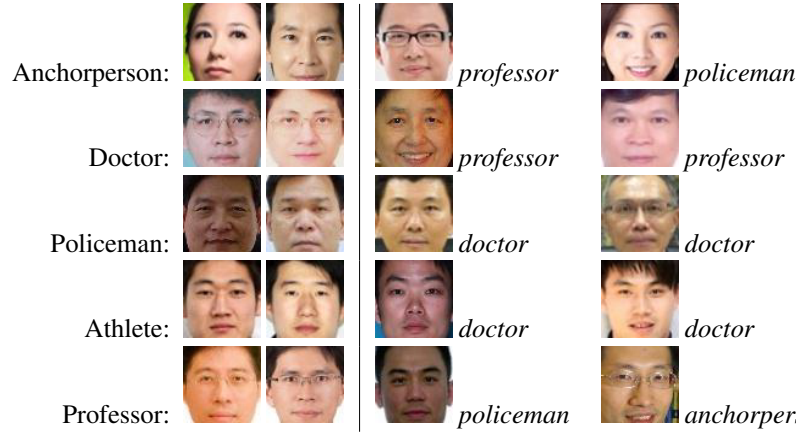


Fig. 10: Sample images that are correctly classified (left part) and falsely classified (right part) in DB1.

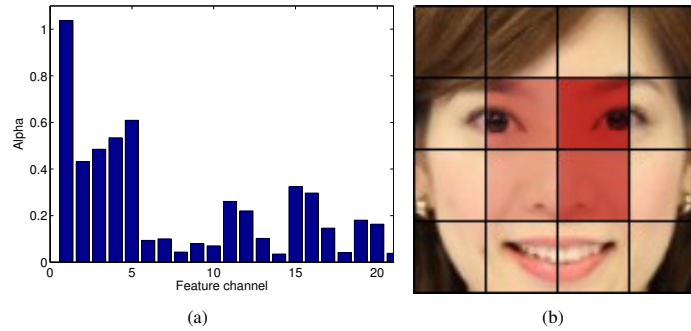


Fig. 11: (a) Weights of classifiers ( $\alpha_t$ ) based on different feature channels. (b) Local patches with higher weights are highlighted.

Classifier weights give some clues of feature effectiveness. Figure 11(a) shows weights of classifiers ( $\alpha_t$ ) learnt for different feature channels by Algorithm 2. According to the spatial pyramid scheme, the first feature is extracted from the whole image, the second to the fifth features are extracted from four semi-global sub-regions, and the sixth to the twenty-first features are extracted from sixteen local sub-regions. From this figure, it can be seen that classifiers trained based on global information are given higher weights ( $\alpha_1$  to  $\alpha_5$ ). Notice that  $\alpha_{11}$ ,  $\alpha_{12}$ ,  $\alpha_{15}$ , and  $\alpha_{16}$  are relatively higher than that for other local feature channels. We especially highlight the local sub-regions corresponding to these larger  $\alpha$ 's in Figure 11(b). This result conforms to our expectation that local sub-regions in the central part of a face are more important in prediction.

### 5.3. Occupation Prediction by Combining Face and Body Context Information

We have verified that face information indeed provides clues for occupation prediction. In this section, we want to verify that performance can be further enhanced by combining face and body context information. From DB2, 100 images are randomly selected from each occupation for training, and the remaining images are for testing. The random split scheme is applied five times, and the average prediction accuracy is reported.

Similar to Table II, Table V show average prediction accuracy of different methods, based on low-level handcrafted features extracted from DB2. This table again verifies the effectiveness of the MC-SVM and the DMC-SVM approaches.

Table V: Average prediction accuracy of different methods, based on low-level hand-crafted features extracted from DB2.

	Random	SVM	MC-SVM	DMC-SVM
Avg. accuracy	4.76	21.43	27.86	28.81

Table VI: Average prediction accuracy by combining face and body context information, based on hand-crafted features extracted from DB2.

Context	Average accuracy
Random	4.76
Face (low-level)	<b>21.74</b>
Face (age)	6.54
Face (gender)	12.23
Face (glass)	6.86
Face (high-level)	12.68
Face (low+high)	<b>24.37</b>
Body (low-level)	<b>29.15</b>
Face+Body	<b>34.36</b>

From Table VI we observe that occupation prediction using low-level features of the face region can be greatly better than random guess. We also see that high-level features of the face region are also helpful. By integrating low-level and high-level features of the face region, over 24% accuracy can be achieved. Comparing the results obtained based on face (low+high) with that obtained based on body context, we see that body context information is more robust. This result is not surprising and explains why existing occupation prediction works start from body context. More importantly, we verify that integrating face and body information yields better performance (34.36% vs. 29.15%).

Table VII shows average accuracy obtained based on deep features extracted from different contexts. Similarly, classification performance based solely on deep features extracted from body is better than that extracted from face. The best performance can be obtained by combining both deep features extracted from face and body. By comparing Table VI with Table VII, we see that significant performance improvement can be obtained (41.10% vs. 34.36%) when deep learning features are used.

Because CNN features are designed to describe an image in a holistic way, it would be interesting to study the obtained performance when CNN features are extracted from the entire image, rather than from the face region or the body region. The bottom half of Table VII shows performance variations based on different settings, i.e., with/without the spatial pyramid (SP) scheme and with SVM/DMC-SVM to do occupation prediction. The result shows that, with spatial pyramid but without DMC-SVM, the obtained performance is just like random. This is due to the curse of dimensionality, i.e., concatenating features from all spatial pyramids forms a very high-dimensional vector. On the other hand, if we describe the entire image, without spatial pyramid, by a single 4096-dimensional vector, and use the conventional SVM to do classification, we obtain 46.07% accuracy. This again shows the superiority of deep features. If we divide the image by the spatial pyramid scheme and adopt the proposed DMC-SVM, we get the highest accuracy 48.8%.

The confusion matrix of prediction accuracy is shown in Figure 12. Perhaps because the dressing style is unique, some occupations have higher prediction accuracy, such as *chef*, *doctor*, and *soldier*. On the contrary, *driver*, *educator*, and *instrument player* have less characteristics to be distinguished from others.

Figure 13 shows some samples that are correctly predicted (left part) and falsely predicted (right part). From this figure we see that samples with typical images of some occupation can be correctly predicted, while not so typical samples may be falsely predicted. We also find that the body context gives important information, which conform to previous works [Song et al. 2011][Shao et al. 2013].

Table VII: Average prediction accuracy by combining face and body context information, based on deep features extracted from DB2.

Features	Channels	Model	Average accuracy
Deep features (face)	SP	DMC-SVM	28.35
Deep features (body)	SP	DMC-SVM	38.70
Deep features (face+body)	SP	DMC-SVM	<b>41.10</b>
Deep features (entire image)	SP	SVM	4.76
Deep features (entire image)	one	SVM	46.07
Deep features (entire image)	SP	DMC-SVM	<b>48.68</b>

baby sitter	19.77	3.49	2.33	2.33	0.00	2.33	0.00	3.49	11.63	3.49	1.16	6.98	4.65	8.14	4.65	2.33	5.81	1.16	10.47	2.33	3.49
barber	5.10	21.94	6.63	8.16	5.10	0.51	2.55	2.04	3.57	9.18	3.06	5.10	4.59	5.10	2.04	2.55	4.08	0.51	1.53	3.06	3.57
chef	0.00	3.70	59.26	7.41	0.00	0.00	0.00	0.00	3.70	3.70	7.41	3.70	3.70	3.70	0.00	0.00	0.00	0.00	0.00	3.70	0.00
doctor	0.44	0.44	3.09	80.57	0.00	0.22	0.66	0.22	1.10	0.88	0.44	0.22	0.88	7.28	0.44	1.55	0.22	0.44	0.22	0.22	0.44
driver	1.90	0.95	3.81	7.62	11.43	2.86	5.71	5.71	5.71	3.81	5.71	10.48	1.90	2.86	1.90	10.48	3.81	6.67	0.00	3.81	2.86
educator	4.14	4.14	2.07	2.76	0.00	17.93	3.45	2.07	4.83	4.83	4.83	16.55	4.14	3.45	1.38	4.14	6.21	2.07	4.14	6.21	0.69
farmer	4.55	0.00	0.00	0.00	0.00	68.18	0.00	0.00	0.00	0.00	0.00	0.00	4.55	0.00	0.00	4.55	4.55	0.00	0.00	9.09	0.00
firefighter	2.00	0.00	0.00	6.00	2.00	0.00	0.00	46.00	4.00	6.00	2.00	2.00	2.00	12.00	2.00	2.00	4.00	0.00	6.00	2.00	0.00
fitness trainer	2.69	0.45	1.79	3.14	2.24	0.45	1.79	5.83	41.70	4.93	4.48	4.48	1.79	3.14	1.35	7.62	1.79	2.69	2.69	2.24	2.69
instrument player	7.50	1.25	3.75	1.25	6.25	0.00	0.00	1.25	11.25	18.75	11.25	6.25	1.25	3.75	2.50	6.25	2.50	2.50	6.25	5.00	1.25
judge	1.45	0.73	1.82	5.45	1.09	1.09	1.09	1.45	2.55	2.18	39.64	17.82	1.82	2.18	1.82	4.73	0.73	3.27	5.45	1.45	2.18
lawyer	0.77	2.31	3.08	3.08	1.54	0.00	0.00	0.00	3.08	1.54	10.77	54.62	0.00	1.54	7.69	0.77	3.08	2.31	0.00	0.00	3.85
mailman	5.41	0.00	0.00	2.70	0.00	2.70	10.81	10.81	2.70	0.00	0.00	5.41	35.14	2.70	0.00	2.70	2.70	5.41	0.00	10.81	0.00
nurse	2.04	2.04	3.85	20.18	0.23	0.23	0.45	0.68	3.17	0.68	0.23	0.68	1.81	56.24	1.36	3.17	0.68	0.00	1.59	0.68	0.00
office worker	1.65	2.47	2.20	12.36	1.65	1.37	1.10	0.27	1.10	1.37	7.42	11.54	0.82	7.14	26.37	3.85	4.12	4.67	4.12	1.37	3.02
patrolman	0.00	0.73	1.46	0.00	1.46	0.00	1.46	0.73	0.00	0.00	1.46	3.65	2.19	0.73	1.46	55.47	2.19	17.52	1.46	5.84	2.19
pet breeder	8.00	2.00	2.00	6.00	0.00	2.00	10.00	2.00	2.00	2.00	4.00	4.00	0.00	2.00	2.00	6.00	38.00	0.00	4.00	2.00	2.00
police officer	1.15	0.00	0.57	4.02	1.15	0.57	2.87	0.57	0.00	3.45	3.45	7.47	0.00	2.30	2.87	17.24	0.57	43.10	0.57	5.75	2.30
receptionist	5.57	1.47	2.05	3.23	0.88	0.59	0.88	4.69	5.87	3.52	4.99	7.92	2.64	5.28	6.45	2.93	3.52	2.64	26.69	3.23	4.99
soldier	0.00	0.00	0.00	0.00	0.00	0.00	3.13	6.25	6.25	0.00	3.13	3.13	0.00	0.00	0.00	9.38	3.13	3.13	0.00	82.50	0.00
waiter	1.97	0.00	4.43	11.82	0.00	0.49	0.99	1.97	5.42	2.46	4.93	2.46	0.99	8.87	0.99	3.45	0.49	1.48	5.91	1.97	38.92

Fig. 12: Confusion matrix of prediction accuracy based on DB2.

To study performance difference between DB1 and DB2, we randomly select five occupation classes from DB2, so that the number of occupations is the same as DB1. We then predict occupation for the selected dataset based on CNN features with DMC-SVMs. The same process runs three times, and the average prediction accuracy is 62.54%. In Table III, the average prediction accuracy based on the same settings for DB1 is 77.50%. This shows that, by removing the factor of the number of occupation classes, DB2 is somewhat harder than DB1. Although more contextual cues are available in images of DB2, higher race variations and pose variations give rise to more challenges.

#### 5.4. Comparing with the State of the Art

Two most important existing works on occupation prediction are predicting via clothing and context [Song et al. 2011] and predicting via social context [Shao et al. 2013]. Shao et al. investigated recognizing occupations of multiple people with arbitrary poses in a photo [Shao et al. 2013]. Visual attributes, co-occurrence, and spatial configuration were jointly considered to build the prediction model. However, in our work we focus on upper-body photos consisting of one single person, and thus the method and dataset in [Shao et al. 2013] cannot be fairly used for performance comparison.

Song et al. [Song et al. 2011] detected head, central body, and left/right shoulders from upper-body photos, and extracted five types of features (as mentioned in the second half of Table I)





Fig. 13: Samples that are correctly classified (left part) and falsely classified (right part) for DB2.

Table VIII: Performance comparison with an existing work, based on data in DB2.

Methods & Features	Average accuracy
Body	29.15
Face+Body	34.36
Deep features (Body)	38.70
Deep features (Face+Body)	<b>41.10</b>
Body (modified from [Song et al. 2011])	32.11

from each region. Low-level features were then encoded by the sparse coding method to constitute intermediate-level representation. Based on this representation, SVM classifiers were constructed to predict occupation for a given photo. We attempt to compare our work with [Song et al. 2011]. However, neither their database nor codes are released. We thus implement their system but replace sparse coding by LLC based on our own DB2 database. Although the feature coding method is different from that mentioned in [Song et al. 2011], the essential idea behind [Song et al. 2011] is maintained. Besides, many studies have shown the superiority of LLC over sparse coding in image classification [Wang et al. 2010].

Table VIII shows performance comparison with the work [Song et al. 2011]. We only show the performances obtained based on the best settings of hand-crafted features and deep features. As can be seen, deep features associated with the DMC-SVM significantly outperform the modified version of [Song et al. 2011] (41.10% vs. 32.11%). The proposed method, therefore, is confirmed to be very effective in the challenging dataset.

Table IX: Average prediction accuracies when one specific high-level attribute is intentionally corrupted, for DB1.

Ratio of change	Gender	Glass	Age
10%	41.72	40.49	41.26
30%	40.06	39.02	39.52
50%	39.33	39.60	38.59

## 6. CONCLUSION

We have presented a work for predicting occupation from images by combining face and body context information. We verify that different occupations present different distributions of visual attributes, and thus predicting occupation is a doable computer vision problem. To describe face and body context information, we extract multilevel features, including hand-crafted features and deep features, from regions at different granularities. To avoid overfitting, features from different spatial sub-regions are combined based on the boosting strategy, and a discriminant multi-channel SVM classifier is constructed to achieve occupation prediction. We report evaluation results from various perspectives, including demonstrating the effectiveness of predicting occupation from face, and showing promising performance obtained by combining face and body information. We further verify the effectiveness of using deep features in the multi-channel SVM framework, which may push forward many deep-learning-based applications.

In this work, we focus on predicting occupation for frontal upper body images. Accurately detecting body parts and extracting robust descriptors from humans in arbitrary poses are still open problems. Besides, the number of occupations being recognized are still limited. It is very challenging to recognize occupations without clear symbols (uniform, hat, specific working environment, or other accessories) if only visual attributes are used. Considering heterogeneous information like social media interaction, sentiment, photo composition, or human behaviors in occupation prediction would be interesting future works.

## ACKNOWLEDGMENTS

The work was partially supported by the Ministry of Science and Technology of Taiwan under the grant MOST103-2221-E-194-027-MY3, MOST104-2221-E-194-014, and MOST105-2628-E-194-001-MY2.

## APPENDIX

To separately verify the effect of correctness of gender, glass, and age information on occupation prediction, we especially focus on the case “High-level (DMC-SVM)” in Table III. If we concatenate all three high-level attributes and employ DMC-SVM, we obtain 40.65% accuracy. Now we randomly change 10%, 30%, and 50% of the estimated (by Face++) gender, glass, and age, respectively, and still concatenate three high-level attributes to predict occupation based on DMC-SVM. Table IX shows average accuracies based on different settings. Note that, for example, the 40.06% accuracy shown in the (2,1)th cell is obtained by that we intentionally change 30% of the estimated gender labels, and concatenate gender (with corruption or without corruption) with non-corrupted glass and age information.

As can be seen from Table IX, when we intentionally change 10% of predicted labels, sometimes performance better than 40.65% can be obtained, i.e., the (1,1)th cell (41.72%) and the (1,3)th cell (41.26%). Note that the Face++ module is not perfect. Therefore, when we intentionally change the predicted labels, sometimes we change its wrong prediction to correct label. On the other hand, if we intentionally change too many of predicted labels, e.g., 50%, performance drops clearly (39.33%; 39.60%; 38.59% vs. 40.65%).

## REFERENCES

John Antonakis and Olaf Dalgas. 2009. Predicting Elections: Child’s Play! *Science* 323, 5918 (2009), 1183.

- Kingsley R. Browne. 2006. Evolved Sex Differences and Occupational Segregation. *Journal of Organizational Behavior* 27 (2006), 143–162.
- K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. In *British Machine Vision Conference*.
- Bor-Chun Chen, Yan-Ying Chen, Yin-Hsi Kuo, and Winston H. Hsu. 2013a. Scalable Face Image Retrieval Using Attribute-Enhanced Sparse Codewords. *IEEE Transactions on Multimedia* 15, 5 (2013), 1163–1173.
- Huizhong Chen, Andrew C. Gallagher, and Bernd Girod. 2013b. What’s in a Name? First Names as Facial Attributes. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 3366–3373.
- Yan-Ying Chen, Winston H. Hsu, and Hong-Yuan Mark Liao. 2013c. Automatic Training Image Acquisition and Effective Feature Selection From Community-Contributed Photos for Facial Attribute Detection. *IEEE Transactions on Multimedia* 15, 6 (2013), 1388–1399.
- Wei-Ta Chu and Chih-Hao Chiu. 2014. Predicting Occupation from Single Facial Images. In *Proceedings of IEEE International Symposium on Multimedia*. 9–12.
- Navneet Dalal and Bill Triggs. 2005. Histograms of Oriented Gradients for Human Detection. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 886–893.
- Hamdi Dibeklioglu, Albert A. Salah, and Theo Gevers. 2015. Recognition of Genuine Smiles. *IEEE Transactions on Multimedia* 17, 3 (2015), 279–294.
- Haoqiang Fan, Mu Yang, Zhimin Cao, Yuning Jiang, and Qi Yin. 2014. Learning Compact Face Representation: Packing a Face into an Int32. In *Proceedings of ACM Multimedia*. 933–936.
- Paul E. Gabriel and Susanne Schmitz. 2007. Gender Differences in Occupational Distributions among Workers. *Monthly Labor Review* 130 (2007), 19–24.
- Robert L. Kaufman and Seymour Spilerman. 1982. The Age Structures of Occupations and Jobs. *Amer. J. Sociology* 87, 4 (1982), 827–851.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of Advances in Neural Information Processing System*.
- Iljung S. Kwak, Ana C. Murillo, Peter N. Belhumeur, David Kriegman, and Serge Belongie. 2013. From Bikers to Surfers: Visual Recognition of Urban Tribes. In *Proceedings of British Machine Vision Conference*.
- Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 2006. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 2169–2178.
- Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient based Learning Applied to Document Recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
- Ce Liu, Jenny Yuen, Antonio Torralba, Josef Sivic, and William T. Freeman. 2008. SIFT Flow: Dense Correspondence across Different Scenes. In *Proceedings of European Conference on Computer Vision*. 28–42.
- Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-Shop: Cross-Scenario Clothing Retrieval via Parts Alignment and Auxiliary Set. In *Proceedings of European Conference on Computer Vision*. 3330–3337.
- David R. Martin, Charless C. Fowlkes, and Jitendra Malik. 2004. Learning to Detect Natural Image Boundaries using Local Brightness, Color, and Texture Cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26, 5 (2004), 530–549.
- Laurent S. Nguyen, Denise Frauendorfer, Marianne S. Mast, and Daniel Gatica-Perez. 2014. Hire me: Computational Inference of Hirability in Employment Interviews Based on Nonverbal Behavior. *IEEE Transactions on Multimedia* 16, 4 (2014), 1018–1031.
- Neil O’Hare and Alan F. Smeaton. 2009. Context-Aware Person Identification in Personal Photo Collections. *IEEE Transactions on Multimedia* 11, 2 (2009), 220–228.
- Timo Ojala, Matti Pietikainen, and Topi Maenpaa. 2002. Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 7 (2002), 971–987.
- Lei Pang and Chong-Wah Ngo. 2015. Unsupervised Celebrity Face Naming in Web Videos. *IEEE Transactions on Multimedia* 17, 6 (2015), 854–866.
- Vignesh Ramanathan, Bangpeng Yao, and Li Fei-Fei. 2013. Social Role Discovery in Human Events. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 2475–2482.
- Ali S. Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *Proceedings of CVPR Workshop on DeepVision*.
- Nicholas O. Rule and Nalini Ambady. 2008. The Face of Success: Inferences From Chief Executive Officers’ Appearance Predict Company Profits. *Psychological Science* 19, 2 (2008), 109–111.

- Ming Shao, Liangyue Li, and Yun Fu. 2013. What Do You Do? Occupation Recognition in a Photo via Social Context. In *Proceedings of International Conference on Computer Vision*. 3631–3638.
- John M. Smith. 1973. Age and Occupation: The Determinants of Male Occupational Age Structures – hypothesis H and Hypothesis A. *Journal of Gerontology* 28, 4 (1973), 484–490.
- Zheng Song, Meng Wang, Xian-Sheng Hua, and Shuicheng Yan. 2011. Predicting Occupation via Human Clothing and Contexts. In *Proceedings of International Conference on Computer Vision*. 1084–1091.
- Alexander Todorov, Anesu N. Mandisodza, Amir Goren, and Crystal C. Hall. 2005. Inferences of Competence from Faces Predict Election Outcomes. *Science* 308, 5728 (2005), 1623–1626.
- Andrea Vedaldi and Karel Lenc. 2015. MatConvNet – Convolutional Neural Networks for MATLAB. In *Proceedings of ACM International Conference on Multimedia*.
- Paul Viola and Michael J. Jones. 2004. Robust Real-Time Face Detection. *International Journal of Computer Vision* 57 (2004), 137–154.
- Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. 2010. Locality Constrained Linear Coding for Image Classification. In *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*. 3360–3367.
- Siyu Xia, Ming Shao, Jiebo Luo, and Yun Fu. 2012. Understanding Kin Relationships in a Photo. *IEEE Transactions on Multimedia* 14, 4 (2012), 1046–1056.
- Chao Xiong, Guangyu Gao, Zhengjun Zha, Shuicheng Yan, Huadong Ma, and Tae-Kyun Kim. 2014. Adaptive Learning for Celebrity Identification With Video Context. *IEEE Transactions on Multimedia* 16, 5 (2014), 1473–1485.
- Xiao Zhang, Lei Zhang, Xin-Jing Wang, and Heung-Yeung Shum. 2012. Finding Celebrities in Billions of Web Images. *IEEE Transactions on Multimedia* 14, 4 (2012), 995–1007.