Predicting Occupation from Single Facial Images

Wei-Ta Chu
Dept. of Computer Science and Information Engineering
National Chung Cheng University
Chiayi, Taiwan
wtchu@ccu.edu.tw

Chih-Hao Chiu
Dept. of Computer Science and Information Engineering
National Chung Cheng University
Chiayi, Taiwan
kennex2conan14@gmail.com

***Abstract*—Facial images embed age, gender, and other rich information that is implicitly related to occupation. In this work, we advocate that occupation prediction from a single facial image is a doable research direction. We first extract visual features from multiple levels of patches and describe them by locality-constrained linear coding. To avoid the curse of dimensionality and overfitting, a boost strategy called multi-feature SVM is used to integrate features. Intra-class and inter-class visual variations are jointly considered in the boosting framework to further improve performance. In the evaluation, we verify that this is a promising research topic with encouraging performance, and also discuss interesting issues from various perspectives.**

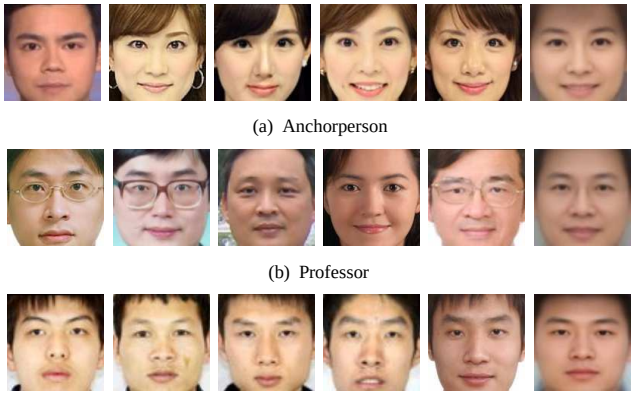
***Keywords*-Occupation prediction; face; discriminant multi-feature SVM; classifier weighting**

I. Introduction

Predicting occupation from images emerges as an important computer vision problem because of its great potential in intelligent services and systems [1]. 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 network web sites or expertise networks if occupations are considered.

Currently, related studies mainly focus on predicting occupations based on human clothing [1], scene context [1], and social context [2]. What people dress, where people work, and how they interact with each other, are all important clues for us to predict occupations. In this work, we advocate an alternative modality that may also reveal important clues on occupation prediction: *facial image*. Although *predicting occupation only from faces* seems making little sense at first, we will demonstrate that it is really a doable direction and can be a complementary approach to advance current clothing-based and context-based methods.

Figure 1 shows sample face images of *anchorperson*, *professor*, and *athlete*, and their average faces. We can easily observe that anchorpersons tend to be female, and athletes tend to be younger. Differences in more facial features, such as skin color, haircut, and glasses wearing, can also be found by further analysis. We conjecture that an occupation can be viewed as a joint distribution over a set of face attributes, and thus a computational model can be built to predict occupation from facial images.



(a) Anchorperson

(b) Professor

(c) Athlete

Figure 1: Face examples of anchorperson, professor, and athlete and their average faces (the last image of each row) from aligned faces. We can clearly see that anchorpersons tend to be female, professors tend to be elder, and athletes tend to be younger.

Our contributions are twofold. First, we propose the first attempt to predict occupation from facial images and verify its effectiveness. A computational model combining texture features and inter-/intra-class characteristics is developed. Second, we collect a face image database consisting of frontal faces with accurate occupation information as the foundation for future research.

II. Related Works

A. Predicting Occupation from Images

Song et al. [1] proposed the first work to predict occupation from images based on human 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 foreground (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.

Shao et al. [2] 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.

Relevant to occupation recognition, social role/group discovery emerged recently [3] [4]. 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 [3]. Kwak et al. [4] 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.

B. Occupation and Faces

In this work, we advocate that occupation can be predicted from facial features only, as an important alternative and complementary cue to existing studies. *What is the intuition behind the idea?* We resort to researches of sociology and official statistics.

First, human mind is sexually dimorphic. Human's temperament and cognitive abilities are affected by gender differences, and this largely influences how a human selects his/her job. More specifically, competitiveness, risk, interest in children, mechanical ability, and verbal ability are all factors relating to how a man or a woman selects his/her job [6]. The gender employment patterns reported in [7] clearly show the occupational differences between men and women in US. Gender, therefore, is clearly related to occupation, also shown in Figure 1. On the other hand, 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.

Second, age is clearly related to job selection. People who are in the occupation needing larger 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, the age patterns in different occupations have been long studied [8] [9]. Kaufman and Spilerman [8] 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 linked by age from some implicit perspective.

In addition to gender and age, other facial features may also link with occupation, such as styles of haircut and hat [1], skin color, and wearing glasses. Based on the foundation of literature mentioned above, we propose to predict occupation from facial images only, as a complementary approach to existing context-based and clothing-based methods.

III. Predicting Occupation

A. Overview

Inspired by the interesting work on first name prediction [5], we follow the main flow proposed in [5] with modification specially designed for occupation prediction. Input facial images are normalized to 64×64 pixels, and are aligned with detected eyes fixed at specific positions. Intensity histogram equalization is conducted for each image to prevent the influence of lighting variation. Dense SIFT (Scale Invariant Feature Transform) descriptors [10] are then extracted by sampling on a dense grid with 2-pixel intervals. Each 128-dimensional SIFT descriptor is then transformed by the Locality-constrained Linear Coding (LLC) [11] to be a 1024-dimensional code. To consider information at multiple scales, the spatial pyramid scheme [12] is used to aggregate LLC codes. Totally 21 pyramid grids are constructed, and LLC codes in each pyramid grid are aggregated. Finally a feature vector of $21 \times 1024 = 21504$ dimensions is extracted for each facial image.

The concatenated LLC codes extracted from the training data consisting of M different occupations are then fed to the LibSVM package [13] to construct a M-class support vector machine classifier. The occupation of a given facial image is predicted by the SVM classifier based on the extracted concatenated LLC code. Note that the concatenated LLC code is very high-dimensional (21504-dim), and the training process described above is susceptible to overfitting and the curse of dimensionality. Therefore, we view a facial image as 21 feature vectors (LLC codes), each vector coming from a pyramid grid and having only 1024 dimensions. In other words, a facial image is represented by vectors from multiple feature channels, which embed information at different granularities with different spatial displacement.

To integrate multiple features, the multi-feature SVM (MF-SVM) framework [5] based on the boosting strategy is adopted to train the SVM classifier. Different from the MF-SVM framework where a large number of 1-vs-1 SVM classifiers were trained, we construct a multiclass SVM in the framework. The characteristic of intra-class variation and inter-class variation is further considered to constitute the proposed discriminant multi-feature SVM (DMF-SVM).

B. Discriminant Multi-Feature SVM

Algorithm 1 shows the training procedure of the DMF-SVM classifier. Suppose we have N training images, and each image $\boldsymbol{x}$ is represented by T feature channels and is with a class label y. We denote $\boldsymbol{x}_{t,i}$ as the tth feature vector extracted from the ith training image, where $t = 1, ..., T$ and $i = 1, ..., N$. First, with the equal weights D_i for all training image, we use the first feature channel ($t = 1$) to construct a multiclass SVM with the five-fold cross validation scheme (line 3). We thus can calculate the prediction confidence $f_t(\boldsymbol{x}_{t,i})$ and an error term based on prediction label $\hat{y}_{t,i}$.

The error term err_t plays an important role in dynamically adjusting classifier weight α_i (line 4) and image weight D_i (line 5). Intuitively, if the ith image is misclassified by the current SVM (based on the first feature channel), its weight D_i will be enlarged when we train the SVM based on the second feature channel. The algorithm finally outputs a set of classifiers f_t specific to the tth feature channel and the corresponding classifier weights α_t. Given a test image $\boldsymbol{q}$, the extracted t feature vectors $\boldsymbol{q}_1, ..., \boldsymbol{q}_T$ are fed to the T classifiers, respectively, and the final classification result is $\sum_{t=1}^{T} \alpha_t f_t(\boldsymbol{q}_t)$.

The update step shown in line 4 is critical and needs more explanation. The first term indicates that, when the classification error err_t of the current SVM for the tth feature channel is larger, the weight α_t of the tth classifier (constructed based on the tth feature channel) gets smaller. By the second term of line 4, we further take the ratio of inter-class variation to intra-class variation into account to update α_t. Let $\boldsymbol{x}^{(i)}$ denote the ith image with the label y_i in the dataset, and let $\mathcal{C}_j = \{\boldsymbol{x}^{(i)} : y_i = j\}$ denote the set of images with the label j. The inter-class variation $dr(y_i, t)$ is calculated as

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

where Z_i is a normalization factor, and $d(\boldsymbol{x}_{t,p}, \boldsymbol{x}_{t,q})$ is the Euclidean distance between images $\boldsymbol{x}^{(p)}$ and $\boldsymbol{x}^{(q)}$. The value $dr(y_i, t)$ is thus the average Euclidean distance, based on the tth feature channel, between images belonging to different classes. On the other hand, the intra-class variation $da(y_i, t)$ is defined as the average Euclidean distance between images within the same class and is calculated as

$$da(y_i, t) = \frac{1}{Z'_i} \sum_{\boldsymbol{x}^{(p)} \in \mathcal{C}_{y_i}, \boldsymbol{x}^{(q)} \in \mathcal{C}_{y_i}}^{|\mathcal{C}_{y_i}|} d(\boldsymbol{x}_{t,p}, \boldsymbol{x}_{t,q}). \qquad (2)$$

Figure 2 shows the intra-class distance distribution of the athlete facial images, and the inter-class distance distribution between athlete and policeman. From this figure we can see that, although there is overlap, the intra-class distribution is distinct from the inter-class distribution. This distinction gives clues to update the classifier weight α_t and yields better performance that will be described later. In this work, the weights w_1 and w_2 to combine two terms are both $\frac{1}{2}$.

IV. Experimental Results

A facial image dataset consisting of 2,062 images belonging to five different occupations, i.e., doctor, anchorperson, athlete, policeman, and professor, was constructed for evaluation. The number of images of each occupation ranges from 300 to 500. Among them, 240 images were randomly selected from each occupation for training, and the remaining images were for testing.

Algorithm 1 Training of Discriminant Multi-Feature SVM

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

Output: SVM classifiers f_t and classifier weights α_t

1: Initialize weights of training images $D_i = 1$
2: **for** t=1 to T **do**
3: Using weights D, perform SVM cross validation to obtain confidence $f_t(\boldsymbol{x}_{t,i})$ and prediction $\hat{y}_{t,i} = sign(f_t(\boldsymbol{x}_{t,i}))$, compute error $err_t = \frac{\sum_{t=1}^{N} |\{\hat{y}|\hat{y}_{t,i} \neq y_i\}|}{N}$
4: Compute $\alpha_t = w_1 \log(\frac{1-err_t}{err_t}) + w_2 \frac{dr(y_i,t)}{da(y_i,t)}$
5: Set $D_i = D_i \exp(-\alpha_t y_i f_t(\boldsymbol{x}_{t,i}))$ and renormalize so that $\sum_{t=1}^{N} D_i = N$
6: Train SVM f_t using D
7: **end for**
8: Output classifiers f_t and classifier weights α_t

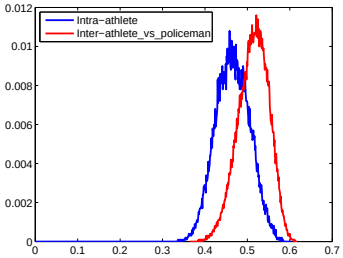


Figure 2: Samples of inter-class distance distribution between athlete and policeman (red) and intra-class distance distribution (blue).

The random split scheme was adopted for five times, and the average classification accuracy is reported in Table I. In the baseline setting, classifier weight α_t and image weight D_i are set as unity always (without any updating). In the MF-SVM setting, the classifier weight α_t is updated as in [5]. The classifier weight is updated as described in Algorithm 1 in the DMF-SVM setting. In the DMF-SVM+ setting, the second term of line 4 of Algorithm 1 is further normalized by variances of distributions, i.e., $\hat{dr}(y_i, t) = dr(y_i, t)/dr_{var}(y_i, t)$ and $\hat{da}(y_i, t) = da(y_i, t)/da_{var}(y_i, t)$, where $dr_{var}(y_i, t)$ and $da_{var}(y_i, t)$ are variances of the inter-class distance distribution and intra-class distance distribution, respectively. The step in line 4 is then modified as $\alpha_t = w_1 \log(\frac{1-err_t}{err_t}) + w_2 \frac{\hat{dr}(y_i,t)}{\hat{da}(y_i,t)}$.

Table I shows that encouraging performance can be obtained for the challenging occupation classification problem. The classification accuracy of all these methods is much better than random guess. Performance superiority of the

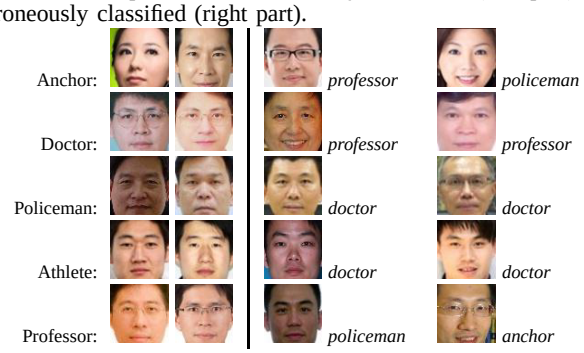
Table I: Average occupation classification accuracy.

	baseline	MF-SVM	DMF-SVM	DMF-SVM+
Avg. accuracy	59.98%	63.78%	63.99%	64.95%

Table II: Confusion matrix of classification accuracy (%).

	Doctor	Anchor	Athlete	Policeman	Professor
Doctor	**56.72**	8.00	5.58	17.30	12.40
Anchor	5.30	**71.18**	2.86	4.90	15.76
Athlete	2.97	0.00	**95.00**	9.50	1.08
Policeman	*21.54*	1.54	3.08	**57.85**	16.00
Professor	*23.60*	9.20	3.60	*19.60*	**44.00**

Table III: Samples that are correctly classified (left part) and erroneously classified (right part).

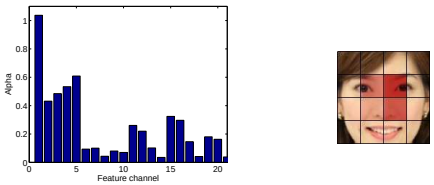


MF-SVM scheme over the baseline scheme shows the effectiveness of classifier weighting and data weighting. The proposed DMF-SVM slightly improves MF-SVM, and with normalization by distance variance, the DMF-SVM+ scheme yields the best performance.

Table II shows the confusion matrix of occupation classification. The classification accuracy of athlete is quite high, probably because of its uniqueness on gender and age (mainly young male). Characteristics of doctor, policeman, and professor are relatively similar and misclassification happens with higher probability. Note that the collected face images are simply portraits with clean background. In such portraits, policemen usually do not wear caps, and doctors do not take their stethoscopes. This task is thus very challenging and worth future study.

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

Figure 3(a) shows weights of classifiers (α_t) learnt for different feature channels by the Algorithm 1. 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 subregions of the image, and the sixth to the twenty-first features are extracted from sixteen local subregions. From this figure, it can be seen that classifiers trained based on global information are given higher weights (α_1 to α_5). Notice that α_{11}, α_{12}, α_{15}, and α_{16} are relatively higher than that for other local feature channels. We especially highlight the local pathes corresponding to these larger α's in Figure 3(b). This result confirms with



(a) (b)

Figure 3: (a) Weights of classifiers (α_t) based on different feature channels. (b) Local patches with higher weights are highlighted.

our expectation that patches in the central region of a face are more important in classification.

V. Conclusion

We have presented predicting occupation from single facial images. Based on facial images, we extract multilevel features from pathes at different levels, followed by LLC to improve feature robustness. To avoid overfitting, features from different patches are combined based on the boosting strategy, and a discriminant multi-feature SVM classifier is constructed to achieve occupation classification. We report encouraging performance in the evaluation, showing that this is a doable computer vision task and would be a promising research topic.

Acknowledgement The work was partially supported by the Ministry of Science and Technology in Taiwan under the grants NSC101-2221-E-194-055-MY2 and MOST103-2221-E-194-027-MY3.

References

[1] Z. Song, M. Wang, X. S. Hua, and S. Yan, *Predicting Occupation via Human Clothing and Contexts*, Proc. of ICCV, 2011.

[2] M. Shao, L. Li, and Y. Fu, *What Do You Do? Occupation Recognition in a Photo via Social Context*, Proc. of ICCV, 2013.

[3] V. Ramanathan, B. Yao, and L. Fei-Fei, *Social Role Discovery in Human Events*, Proc. of CVPR, 2013.

[4] I. S. Kwak, A. C. Murillo, P. N. Belhumeur, D. Kriegman, and S. Belongie, *From Bikers to Surfers: Visual Recognition of Urban Tribes*, Proc. of BMVC, 2013.

[5] H. Chen, A. C. Gallagher, and B. Girod, *What's in a Name? First Names as Facial Attributes*, Proc. of CVPR, 2013.

[6] K. R. Browne, *Evolved Sex Differences and Occupational Segregation*, Journal of Organizational Behavior, vol. 27, pp. 143–162, 2006.

[7] P. E. Gabriel and S. Schmitz, *Gender Differences in Occupational Distributions among Workers*, Monthly Labor Review, vol. 130, pp. 19–24, 2007.

[8] R. L. Kaufman and S. Spilerman, *The Age Structures of Occupations and Jobs*, American Journal of Sociology, vol. 87, no. 4, pp. 827–851, 1982.

[9] J. M. Smith, *Age and Occupation: The Determinants of Male Occupational Age Structures hypothesis H and Hypothesis A*, Journal of Gerontology, vol. 28, no. 4, pp. 484–490, 1973.

[10] D. Lowe. *Distinctive image features from scale-invariant keypoints*, IJCV, vol. 60, no. 2, pp. 91–110, 2004.

[11] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. *Locality constrained linear coding for image classification*, In Proc. CVPR, pp. 3360–3367, 2010.

[12] S. Lazebnik, C. Schmid, and J. Ponce. *Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories*, In Proc. CVPR, pp. 2169–2178, 2006.

[13] C.-C. Chang and C.-J. Lin. *LIBSVM: A library for support vector machines*, ACM Tran. on Intelligent Systems and Technology, 2011.