What Do You Do? Occupation Recognition in a Photo via Social Context

Ming Shao[1] Liangyue Li[2] Yun Fu[12*]

[1] College of Computer and Information Science, Northeastern University, MA, USA

[2] Department of Electrical and Computer Engineering, Northeastern University, MA, USA

mingshao@ccs.neu.edu, {liangyue, yunfu}@ece.neu.edu

Abstract

In this paper, we investigate the problem of recognizing occupations of multiple people with arbitrary poses in a photo. Previous work utilizing single person's nearly frontal clothing information and fore/background context preliminarily proves that occupation recognition is computationally feasible in computer vision. However, in practice, multiple people with arbitrary poses are common in a photo, and recognizing their occupations is even more challenging. We argue that with appropriately built visual attributes, co-occurrence, and spatial configuration model that is learned through structure SVM, we can recognize multiple people's occupations in a photo simultaneously. To evaluate our method's performance, we conduct extensive experiments on a new well-labeled occupation database with 14 representative occupations and over 7K images. Results on this database validate our method's effectiveness and show that occupation recognition is solvable in a more general case.

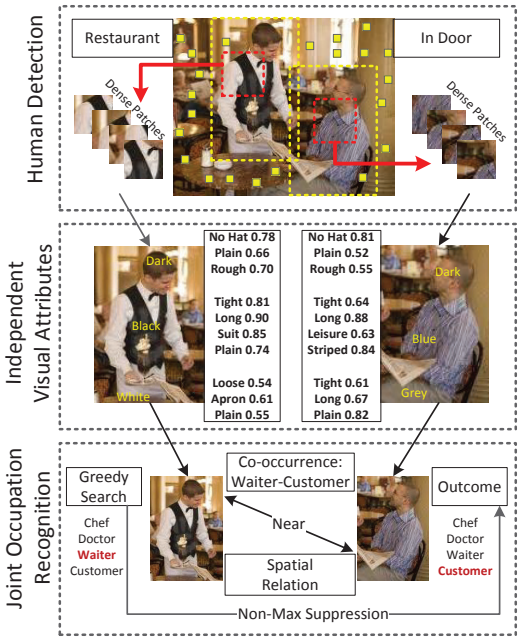

Figure 1. Framework of the proposed method.

1. Introduction

Social characteristics of human, e.g., social status, connections, and roles in a particular situation draw great attention since they are the essence of social life. In the era of social media, more and more social characteristics can be conveyed via digital carriers, e.g., images, videos, and parsed by demographical profiles, e.g., identity [41], gender [3] and age [12]. Recently, an increasing number of works focus on social characteristics under social context: identifying people by shot time of images, fix pattern of co-occurrence, and re-occurrence [22]; recognizing a group of people via social norm and conventional positioning [14]; jointly solving social relation and people identification [31]; recognizing people by linked text in videos or images [2]; exploring kinship via transfer learning and semantics [34].

Nonetheless, computational tools above are now struggling to keep pace [28], in particular with the emergence of many social network websites, e.g., Facebook, Twitter, Google+, and photo sharing websites, e.g., Flickr, Google Picasa, Instagram. Registered users can upload their personal photos, containing themselves or people associated. These photos, as well as any other personal profiles, have already been sufficient to infer users' interests and tastes. Websites could provide better services and useful recommendations if their inferences on users' preferences are precise. Generally, people would like to chat to those with similar occupations or backgrounds. Therefore, understanding people's occupations in customers' photos can significantly

*This research is supported in part by the NSF CNS award 1314484, Office of Naval Research award N00014-12-1-0125 and N00014-12-1-1028, Air Force Office of Scientific Research award FA9550-12-1-0201, and U.S. Army Research Office grant W911NF-13-1-0160.

improve the experience of social connection recommendations and professional services aimed at a particular group.

In this paper, we propose a novel framework towards multiple people's occupations recognition in a photo. The entire framework is shown in Figure 1. Unlike [27] that considers images of single person with "nearly frontal" pose, ours can tackle multiple people with arbitrary poses in an image through spatial constraint and occupations co-occurrence. To this end, (1) dense local clothing patches are extracted to formulate pose invariant low-level feature; (2) a novel visual attribute learning method that adopts discriminative filters learned through training and standard images is proposed; (3) max-margin training is utilized to learn the steady structure model over multiple people in the photo, and a simple greedy forward search is employed to infer on the learned model; (4) to validate our model in practice, we build the largest occupation image database so far, which includes 14 representative occupation categories, and over 7K images. Experimental results on this database demonstrate that the proposed method can deal with occupation recognition problem in a more general case, regardless of pose variation, human interaction, and messy background.

1.1. Related Work

Occupation prediction has been preliminarily discussed and reasonably solved in [27] where a clothing descriptor based framework is proposed. The clothing feature is described via part-based modeling on patches of human body parts, and is semantically represented by informative and noise-tolerant sparse coding [40]. In addition, they use Bag-of-Words model [9] to capture low-level features in both foreground and background, so that the prediction accuracy can be further enhanced. However, there are still problems unsolved that we will address in this paper. First, nearly frontal upper-body cannot be always strictly satisfied in real-world applications. Second, other than low-level features, mid-level features like visual attributes are also helpful. Third, a person's occupation is tightly coupled with others' in the image, by which we can improve overall accuracies of all the people in one photo.

Clothing parsing draws increasing attention recently due to its close relation with people's social identity and commercial value. First, people are inclined to wearing the same cloth in a short time period, which is tightly connected with identities [13, 33]. Second, clothing preference and style are carriers of many social characteristics and demographical information, e.g., gender [3]. Third, content based image retrieval for clothing offers a more flexible way to choose and compare products with high efficiency [21, 36]. Finally, clothing recognition has been extended to video surveillance as a real-world practice [38].

Visual attributes are descriptive words designed by human to capture visually perceptible properties of objects. For example, we use "fluffy" to describe animal's fur and "round" an object's shape. As semantical mid-level features, visual attributes re-organize the complex relations between low-level features and high-level labels, due to its generality over all objects, i.e., color, texture, pattern, shape. They have been widely used to describe objects' properties [8], recognize objects [32], verify faces [17], and parse clothing [21]. Recently, researchers exploit ranking function [24] and augmented parts [26] to better represent attributes.

2. Low-Level Feature Representation

In [27], people use four learned key points on human body to locate clothing patches, i.e., hat, torso, left, and right shoulder. Although their method works effectively under the assumption of nearly frontal upper-body, when heads or bodies are tilt, rotated, their key points learning method may fail. Notably, recent studies on clothing parsing have attempted to tackle the alignment of clothing parts through either auxiliary database [21] or superpixels with Conditional Random Field (CRF) [36]. However, their problems are different from ours in that they either consider single person with simple posing [36] or external database [21].

Differently, we propose to use dense local patches to yield invariant discriminative features. In this paper, local features are generated by the following scheme: (1) obtain local patches through the detector from [39]; (2) bin the features in each patch; (3) concatenate bins into a longer discriminative feature. In-reality, however, clothing parts misalignment is common and mainly due to pose variations [21]. In most cases, the key part of clothing is slightly shifted from the ideal model. This enlightens us to employ the overlapped patches with different shifts to compensate the misalignment. The generated patches, or an image set, reasonably simulate most possible variations due to human poses, therefore potentially become good candidates of local clothing parts. The generation of local dense patches is illustrated in Figure 2. Similar to [3, 27, 21], we use HOG [6], LBP [1], color histogram, and skin fraction to represent local clothing patches, and these local descriptors are implemented in dense-grid fashion.

So far, we only discuss low-level features drawn from human clothing. In fact, background information also proves to be informative in determining occupations [27]. Understandably, some occupations work in specific environments. For example, sports players play games in either venue or stadium, while chef and doctor always work inside a building. According to benchmark test in [35], we employ four representative features, HOG, dense SIFT [19], LBP, and GIST [23] to describe the attached background[1] of

[1]We use the entire image as the input of background feature.

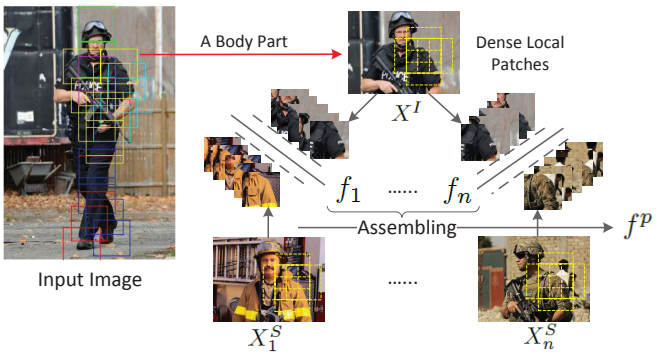

Figure 2. Illustration of dense local patches and assembling of discriminative filters. First, clothing patches of the policeman are detected through the model in [39]. Second, for each clothing patch, we shift its bounding box up, down, right, left by 1/3 of the original size, and yield dense local patches X^I as figure illustrated. Third, we use the dense local patches as positive samples and patches X_1^S from standard images (e.g., firefighter) that do not have this attribute, i.e., blue, as negative samples to train a SVM model whose hyperplane as well as bias will be used as discriminative filter f_1 for visual attributes. Finally, we assemble discriminative filters $f_1, \ldots, f_n$ to form a new visual attribute descriptor f^p.

each image, and their combinations serve as the descriptors for background attributes.

3. Discriminative Filter for Visual Attribute

3.1. Motivation

Traditionally, visual attributes are learned directly through low-level feature [17] or ranks from a wide-margin ranking function [24]. Different from them, we believe the difference between relevant visual attributes is a good descriptor for visual attributes. Intuitively, for some attributes, their uniqueness can be highlighted by conceptually relevant attributes. For example, in our clothing description, pattern "plain" is conceptually defined through "striped" or "spotted", and vice versa. Therefore, the "difference" between "plain", and "striped" or "spotted" can be a potentially ideal descriptor for attribute "plain". Mathematically, this "difference" can be highlighted by decision boundary or discriminative filter in SVM if we consider different visual attributes as data with different labels.

In addition, our formulation of descriptor for visual attributes avoids the vagueness of ranks. For example, if we assign high scores to attribute "plain", then we should assign low scores to both "striped" and "spotted". But relative ranks between "stripped" and "spotted" are difficult to define. In our case, we simply use them together to formulate the discriminative filter.

3.2. Descriptor Formulation

In this paper, image sets with relevant yet different attributes are defined as "`standard images`". For example, clothing patches in other colors are standard images for clothing patches in red. Suppose we have a group of dense local clothing patches $X^I = [x_1, x_2, \ldots, x_m]$ as input for the visual attribute, where x_i is a feature vector for a local patch, and a standard image set $\mathcal{X}^S = [X_1^S, X_2^S, \ldots, X_n^S]$, where each X^S is a group of m local patches from different clothing with relevant yet different visual attributes to the input image. Then we can obtain a group of discriminative filters through:

$$\begin{aligned} w, b = \arg\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i, \\ \text{s.t.} \quad \forall x_i \in X^I, \quad w^T x_i + b \geq 1 - \xi_i, \\ \forall x_i \in X^S, \quad w^T x_i + b \leq -1 + \xi_i, \end{aligned} \tag{1}$$

where ξ_i, C are slack variable and penalty term used in conventional soft margin SVM, respectively. In geometry, linear filter w and corresponding bias b set up a decision boundary for image set X^I and X^S, highlighting the difference between X^I and X^S which could be a better descriptor for visual attribute represented by X^I. To make it concrete, suppose we have two relevant shape visual attributes, i.e., square and round. Low-level feature HOG is able to capture the difference and the learned w and b weight more on HOG feature to differentiate square from round. In practice, we append 1 to the end of each vector x to yield a discriminative filter $f = [w; b]$ without bias term.

3.3. Assembling

In last example, if more comparisons are made (e.g., we compare square with triangle as well.), we may obtain better descriptors for visual attributes. We therefore propose to use several standard images to yield several comparison results. Namely, $f_1, f_2, \ldots, f_n$ are filters generated by training samples pairs $[X^I, X_1^S], [X^I, X_2^S], \ldots, [X^I, X_n^S]$. To enhance the performance, we resort to the following assembling approach:

$$f^p_{(i)} = \frac{1}{n}|F_{(i,:)}|_\alpha, \tag{2}$$

where $F = [f_1, f_2, \ldots, f_n]$, $|\cdot|_\alpha$ is the vector α norm, and $F_{(i,:)}$ denotes the i-th row of filter matrix F. In this paper, we consider using max-pooling (by setting α to ∞), which has been extensively discussed and employed in image classification [25, 37] due to its local translation invariance and biological plausibility. In addition, for filters differentiating input from all standard images, max-pooling along one dimension will return the most significant response, and hence can appropriately describe the traits of the input. The assembling process is shown in Figure 2.

For each attribute $a_i (a_i \in A)$, we will generate a corresponding descriptor $f^p_{a_i}$. To ease the following joint occupation recognition scheme, we predict the probability of each attribute through SVM [4] and its probabilistic output [5], and these visual attributes are now scaled to $[0, 1]$. Finally, we can represent each person with a probabilistic attribute vector $[P(a_i), P(a_2), \ldots, P(a_{|A|})]^T$ of length $|A|$.

4. Joint Occupation Recognition

4.1. Prior Knowledge

We state our learning model dealing with multiple people's occupations recognition in a photo, with arbitrary poses and interactions. Joint recognition of multiple people [22, 14, 31] or objects [18, 15, 7] in an image has been broadly discussed, and the most important inter-class prior are co-occurrence, and spatial context. Specifically, as to occupation recognition problem, the co-occurrence is known as: people with the same or relevant occupations seem to appear in the same photo with high probability. For example, people with relevant occupations could be "teacher—student" and "waiter—customer". Spatial context indicates the structure of images under some social assumption. For example, waiter is standing beside a sitting customer, and a group of people are standing in a line. We integrate both of them into the following model.

4.2. Model Description

We propose a score maximum model to find out the most possible occupations for the people in it. Suppose there are K people in a single photo denoted by $Z = \{z_i | i = 1, 2, \ldots, K\}$, where z_i is an attribute vector with each element indicating the probability of existence of an attribute. We also assume that we have C of classes occupations, and $y_i \in \{1, 2, \ldots, C\}$ denotes the occupation label of the i-th people. Therefore, occupation recognition in an image is equivalent to maximizing the following score function:

$$J(Z, Y) = \sum_{i,j} w^T_{y_i, y_j} l_{ij} + \sum_i w^T_{y_i} z_i, \tag{3}$$

where $Y = \{y_i | i = 1, 2, \ldots, K\}$ is a label vector, and l_{ij} indicates the spatial context feature that quantifies relative location of the i and j-th people into several bins, i.e., above, below, overlapping, next-to, near, and far. Similar definition has been used in [7] to define the relative location of different objects. For each individual, we use a class template w_{y_i} to weight the attribute vector z_i, while for pairwise relation of people i and j, we encode their spatial configuration in w_{y_i, y_j}. Note w_{y_i, y_i} is valid in the situation of people with the same occupation, and is always set to relatively large value in practice.

The score function in Eq. (3) is similar to the model proposed in [11] where it considers matching a pictorial structure to an image through energy minimization. However, both of them prove to be NP-Hard if the underlying structure is arbitrary. Different from the pictorial structure scheme in [10] solved by dynamical programming (DP) under the assumption of underlying tree structure, our model is in discriminative fashion, and people's relations embedded in the photo are not necessarily in tree structure. We instead resort to a non-max suppression (NMS) like greedy search algorithm for inference. In practice, this algorithm works comparably to the exact inference, but in a more efficient manner [7].

Input: $w_{y_i}, w_{y_i, y_j}, z_i, i, j = 1, 2, \ldots, K$
Output: Occupation labels Y for each individual
Initialization: $Y = \emptyset, J = 0, \triangle y_i = w^T_{y_i} z_i$
while *there are unlabeled people in the photo* **do**
  $\triangle y_i = J(Z, Y \cup y_i) - J(Z, Y)$;
  $(i^*, y^*_i) = \arg\max_{i, y_i} \triangle y_i$;
  $Y = Y \cup y^*_i$;
end

Algorithm 1: Inference through greedy search.

4.3. Inference

The proposed greedy search in Algorithm 1 is analogous to the non-max suppression (NMS) proposed in [20], however, each local part of the object in original model is replaced by different people in the occupation problem. Our algorithm can be briefly stated as: first, in candidate pool, find an individual satisfying $i^*, y^*_i = \arg\max_{i, y_i} w^T_{y_i} z_i$, and set $J(Y, Z) = w^T_{y^*_i} z_{i^*}$; second, find the most "compatible" occupation for another individual j, which most enhances the $J(Y, Z)$ by considering both $w^T_{y^*_j} z_{j^*}$ and $w^T_{y^*_i, y^*_j} l_{ij}$ at the same time; repeat this process until all the individuals are added and assigned appropriated occupation labels. We summarize these steps in Algorithm 1. Compared with heuristic algorithms, the greedy search strategy is potentially exponential, especially when the number of total subjects are large. Fortunately, two factors prevent us from exhaustive search. First, the locations of people are fixed, and features are in low-dimension. Second, the number of occupations and people in a photo are not large.

4.4. Learning

We consider optimizing w_{y_i} and w_{y_i, y_j} in a max-margin learning procedure. The output space in our problem incorporates multiple labels and their structure, rather than a single binary label. Therefore, we re-formulate w_{y_i} and w_{y_i, y_j} by w_b and w_a, favoring the multiple labels and output structure, respectively:

$$J(Z, Y) = \sum_{i,j} w^T_a \psi(y_i, y_j, l_{ij}) + \sum_i w^T_b \phi(z_i, y_i). \tag{4}$$

Input: $(Z_1, Y_1), \ldots, (Z_N, Y_N), C, \epsilon$
Output: w, ξ
Initialization: $\mathcal{H} = \emptyset$
repeat
  $(w, \xi) \leftarrow$ solve problem (8) based on current $\mathcal{H}$;
  for $n = 1$ **to** N **do**
    $Y_n^* \leftarrow \arg\max_{Y_n^* \in \mathcal{Y}} \{\triangle(Y_n, Y^*) + w^T \Psi(Z_n, Y^*)\}$;
  end
  $\mathcal{H} \leftarrow \mathcal{H} \cup \{(Y_1^*, \ldots, Y_N^*)\}$;
until $\frac{1}{N}\sum_n^N \triangle(Y_n, Y_n^*) - \frac{1}{N} w^T \sum_n^N [\Psi(Z_n, Y_n) - \Psi(Z_n, Y_n^*)] \leq \xi + \epsilon$;

Algorithm 2: 1-slack formulation for structure SVM.

Recall that we have 6 spatial relations, $|A|$ dimensional feature, and C categories of occupations. Then the dimensionality of w_a and w_b is $6C^2$ and $C \times |A|$, respectively. Analogously, both $\psi(\cdot)$ and $\phi(\cdot)$ are sparse vectors whose elements are allocated by (y_i, y_j) and y_i, respectively. Since we predict labels and their structure together, we integrate weight vectors into one, having the following formulation:

$$J(Z, Y) = w^T \Psi(Z, Y), \tag{5}$$

where $w = [w_a; w_b]$, $\Psi(Z, Y) = [\sum_{ij} \psi(\cdot); \sum_i \phi(\cdot)]$. Next, we will show how to train a max-margin model that given training data $Z_n, n = 1, 2, \ldots, N$, the predicted label $\bar{Y}_n^*$ for Z_n is approaching the true label Y_n, i.e., $\bar{Y}_n^* \approx Y_n$. This essentially is a loss minimization problem plus a regularized term:

$$\begin{aligned} &\arg \min_{w, \xi_n > 0} \frac{1}{2} w^T w + \frac{C}{N} \sum_n^N \xi_n, \\ \text{s.t.} \quad &\forall \bar{Y}_n \quad w^T \triangle\Psi(Z_n, Y_n, \bar{Y}_n) \geq \triangle(Y_n, \bar{Y}_n) - \xi_n, \end{aligned} \tag{6}$$

where $\bar{Y}_n$ is the hypothesis of the true label Y_n, $\triangle\Psi(Z_n, Y_n, \bar{Y}_n) = \Psi(Z_n, Y_n) - \Psi(Z_n, \bar{Y}_n)$, $\triangle(Y_n, \bar{Y}_n)$ is the loss function that quantifies the loss associated with the hypothesis $\bar{Y}_n$, ξ_n is the slack variable in the n-th constraint, and C is the penalty term. More specifically, the loss function here sums over all the single label loss functions indicated by $\triangle(y_i, \bar{y}_i)$, namely,

$$\triangle(Y, \bar{Y}) = \sum_i \triangle(y_i, \bar{y}_i), \text{where} \quad \triangle(y_i, \bar{y}_i) = 1_{y_i \neq \bar{y}_i}, \tag{7}$$

Problem (6) is essentially a structure SVM [29] favoring the constraint term that involves structure output based loss function. This is identified as margin-rescaling in [30].

The key step in solution of problem (6) is to find the most significant violated constraint, namely, to find the most violated hypothesis $\bar{Y}_n$. Intuitively, if the most violated hypothesis satisfies the constraints in problem (6), then all other hypothesis should be valid. However, the runtime for this n-slack formulation in problem (6) is still polynomial with cutting plane method. To accelerate, we refer to 1-slack formulation in [16], which employs a single slack variable ξ, rather than a group of ξ_i for each constraint. We then rewrite the problem (6) in:

$$\begin{aligned} &\arg \min_{w, \xi > 0} \frac{1}{2} w^T w + C\xi, \\ \text{s.t.} \quad &\forall \bar{Y}_n \quad \frac{1}{N} w^T \sum_n^N [\Psi(Z_n, Y_n) - \Psi(Z_n, \bar{Y}_n)] \\ &\geq \frac{1}{N} \sum_n^N \triangle(Y_n, \bar{Y}_n) - \xi. \end{aligned} \tag{8}$$

Differently, since 1-slack formulation shares one unique ξ among all constraints, it adds only one the most violated hypothesis in each iteration. This consequently makes linear runtime possible. To solve this problem, a working set $\mathcal{H}$ is constructed to store the hypothesis and violated constraints. In each iteration, we compute w over the current $\mathcal{H}$, find the most violated constraint based on current w, and add it to the working set. The iteration will not terminate until no constraint can be found that is violated by more than the desired precision ϵ. The solution of problem (8) is summarized in Algorithm 2

Figure 3. Illustrations of the collected occupation database. There are 14 occupations and over 7K images in total.

5. Database

To the best of our knowledge, the occupation database[2] collected in this paper is so far the largest image database for occupation recognition research in computer vision community. There are over 7K images of 14 different occupations, and each category contains at least 500 images. These images are downloaded from the Internet using image

[2]The database will be public available soon.

search engines, e.g., Google Image, Flickr, and social network website, e.g., Facebook. We conduct the query process by trying the names of occupations, their synonyms, and relevant concepts.

We select 14 representative occupation categories (shown in Figure 3) from over 200 well-defined occupations in Wikipedia based on following fundamental criteria. First, we choose the occupation category that is visually informative from human perspective. Hats and uniforms are main characteristics of these categories. Second, we remove some visually informative ones but hard to be recognized by either face detector or human detector, e.g., astronaut. Third, we remove photos with dense crowd and severe overlap among people. Though our framework can deal with multiple people, solving the aforementioned problem is already beyond the scope of this paper. As to the attributes annotation, three specialists are involved to label each attribute over all images. Their responsibility includes: (1) select or remove images not qualified based on the previous criteria; (2) assign attributes to images via majority voting over three of them. Examples of our database for each occupation category is shown in Figure 3.

6. Experimental Results

In this section, we conduct two groups of experiments to test the proposed framework. First, we demonstrate the effectiveness of the proposed descriptor for visual attributes. Second, we compare our joint learning framework with the state-of-the-art method in [27]. Note that we extract low-level features for visual attributes only from corresponding body parts, e.g., hat attributes from head area, upper body attributes from torso and arms. Similar to Figure 2, we use 1/4, 1/3, and 1/2 as offsets in experiments to yield dense local patches. For attributes experiment, we use 250 images from each occupation category with one person in each photo for training and test, while for the joint learning framework, we use another 250 images with more than one person in each photo from each category for training and test.

6.1. Evaluation on Visual Attributes

We illustrate the visual attributes discussed in this paper in Figure 4. The negative samples of each attribute are selected from mutually exclusive samples. For instance, the negative samples for hat attribute "Rimless" are samples from other four attributes, namely, "Uniform", "Helmet", "Cap", "No Hat". Please be notified that for upper or lower body clothing attributes, not all attributes are mutually exclusive, but part of them. For instance, in Figure 4, upper body, "Tight" and "Loose" cannot exist at the same time, but they are compatible with "Long", "Short", and "Vest".

We use one-to-rest binary classification strategy to test each attribute, and 5-fold cross validation is implemented to conduct the test. 25 images are randomly selected from

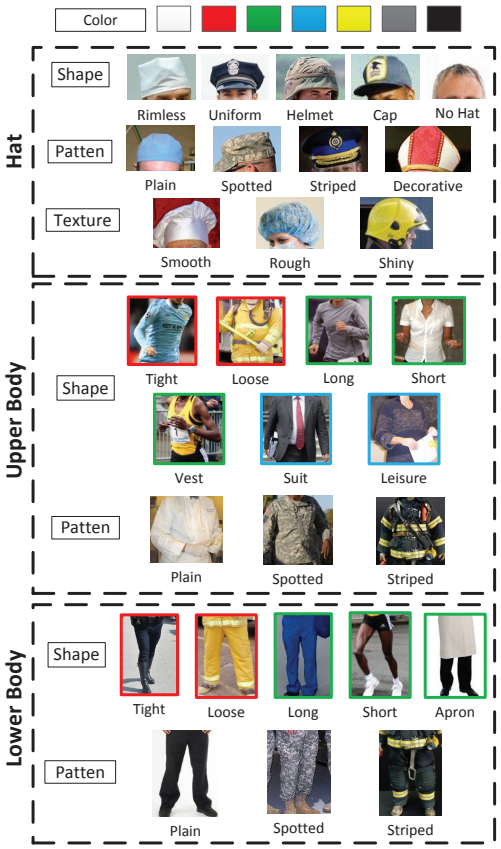

Figure 4. Illustration of the visual attributes used in our framework. Note that attributes in one group are mutually exclusive, e.g., five attributes in shape group of hat. However, for shape group in upper body and lower body, only attributes with the same color border are mutually exclusive.

each attribute as the standard images. The probabilistic outcomes from SVM are used to compute precision and recall. We finally obtain the average precisions for each attribute shown in Figure 5, where "SVM-Visual Attributes" means we directly use low-level feature, while "DF-Visual Attributes" means we use the proposed discriminative filter (DF) as the descriptor. In addition, we also illustrate the impact of numbers of standard images in Figure 6.

In Figure 5, we observe that the proposed visual attributes based on discriminative filters perform better than the attributes based on low-level features plus SVM. While for attributes like color that can be easily differentiated through low-level feature, the improvement is acceptable, the improvement on other attributes, e.g., "Spotted" in hat, is significant. Indeed, we find that soldiers' recognition accuracy is impressive in the later section. Although most attributes achieve acceptable performances, each single attribute cannot directly determine the occupation category. In next experiment, we demonstrate that their combinations offer discriminative features to construct classifiers for both single and multiple people occupation recognition. Final-

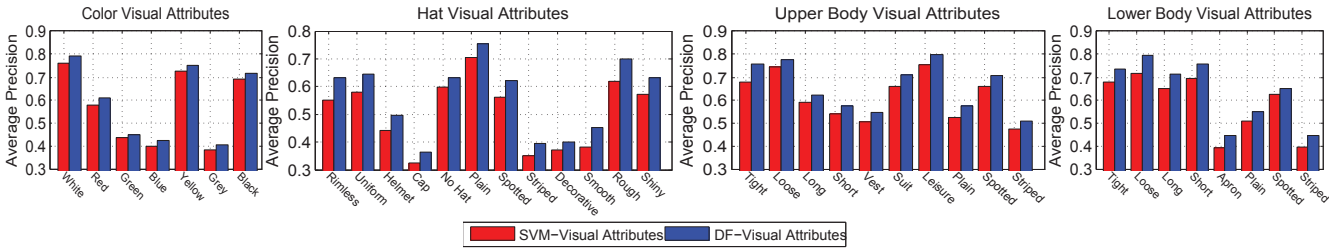

Figure 5. Average precision of different groups of visual attributes. Color group is the average performance over hat, upper and lower body. "SVM-Visual Attributes" means we directly use low-level feature and SVM to predict attributes.

Table 1. Experimental results of average precision (%). The **average performance** of these methods are: Background (15.4%), Method in [27] (35.0%), Single (38.1%), Joint (41.1%).

	chef	clergy	construction labor	customer	doctor
Background	10.3	10.8	11.4	7.4	9.6
Method in [27]	40.8	34.2	42.8	19.2	44.9
Single(Ours)	41.2	35.7	44.9	22.4	46.8
Joint (Ours)	43.8	36.9	47.8	24.8	49.7
	firefighter	lawyer	mailman	marathoner	police
Background	8.3	31.7	19.7	12.8	9.1
Method in [27]	31.3	59.1	21.8	48.2	18.4
Single(Ours)	32.7	58.9	24.7	53.2	18.7
Joint(Ours)	38.1	59.4	27.9	58.7	24.1
	soccer player	soldier	student	teacher	waiter
Background	28.8	31.5	14.8	7.8	17.6
Method in [27]	48.2	60.1	21.5	13.6	20.6
Single(Ours)	59.3	70.4	23.6	14.8	24.4
Joint(Ours)	60.5	75.1	25.1	15.7	28.9

ly, we discover that the numbers of standard images matter in Figure 6. Results show that more standard images yield better results when this number is not very large (5-25).

6.2. Evaluation of Joint Learning Framework

In our joint learning framework, we consider the scores from an individual and its compatibility with others in the same image. In Table 1, we compared four methods whose results are generated by 5-fold cross validation. Note "Single" means single person's occupation recognition which is similar to the method in [27], but not identical, since we use "dense local patches" + "DF-visual attributes" instead of low-level clothing features to deal with pose variations. For "Background", we use features combination mentioned in Section 3, i.e., HOG, SIFT, LBP, GIST. Then we train a one-to-rest binary SVM for each occupation category and use background features in test images as inputs. Note that background is also a visual attribute element in the attribute vector used in this paper. In addition, we add a hidden occupation "customer" in Table 1, indicating people who have interactions with occupations such as "waiter", or "doctor".

From Table 1 We can see that the proposed framework works better than the state-of-the-art method. It becomes significant when people of this occupation tend to show arbitrary poses, e.g., sports player, soldier. Under this situation, clothing patch based method is not stable and many irrelevant factors will be filled into the clothing feature. However, our "dense local patches" + "DF-visual attributes" based method still works well. On the other hand, method in [27] performs well when people of this occupation always show the nearly frontal pose, e.g., lawyer. In addition, the accuracy is enhanced by the interactive occupations, e.g., waiters and customers. We also find a significant improvement in occupations tending to show a group of people, e.g., soldier, marathoner. This proves that our social context based joint learning framework is effective. Finally, we find that the background feature is valuable for some occupations, e.g., lawyer, soccer player, soldier.

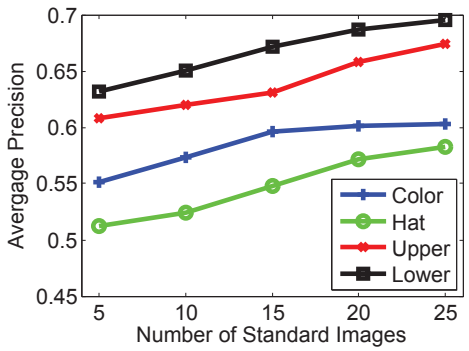

Figure 6. Impact of the numbers of standard images. We use different numbers of standard images to learn DF-visual attributes descriptors, and the average performance over color, hat, upper body, and lower-body is shown indexed by different curves.

In general, we can see that occupation recognition in an image is still challenging, due to the lack of unique attributes, e.g., students, teachers, or large variations of clothing style, e.g., mailman. Consequently, these categories are easily misclassified into other ones. To show the multi-class classification details over 14 categories, we also compute c the confusion matrix by the recognition results from multi-class SVM, and list it in Figure 7. The number in i-th row and j-th column indicates the false alarm rate to i-th class when recognizing j-th class. From this confusion matrix, some typical mistakes made by the classifier are revealed. For example, a construction labor is easily misclassified as a firefighter while teachers are randomly classified as other

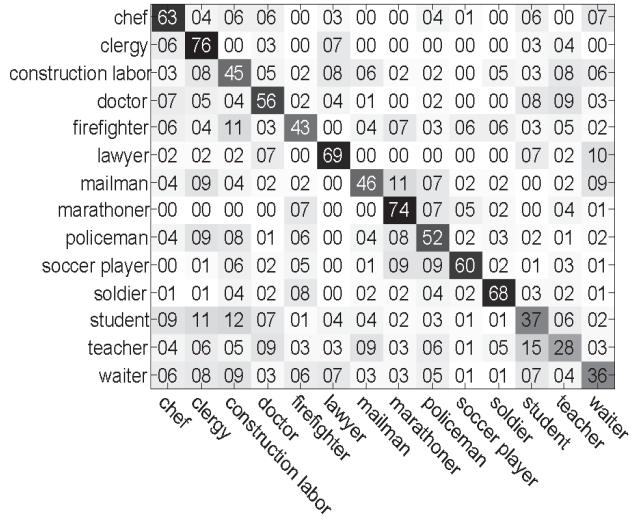

Figure 7. Confusion matrix of occupation recognition results.

occupations since their dressing style is not unique.

7. Conclusions

We proposed a novel framework towards multiple people's occupations recognition in a photo. First, we apply dense local patches scheme to detected human body parts, therefore yielding robust low-level feature representation. Second, visual attributes are learned through assembling discriminative filters to bridge the semantic gap between low-level features and high-level labels — occupation categories. In addition, social context is added to formulate a score maximum model, which is trained through a structure SVM. Extensive experiments on the collected database show that our method works better than the state-of-the-art, especially when there are interactive occupations or a group of people in a photo.

References

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *ECCV*, pages 469–481, 2004. 2

[2] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. G. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *CVPR*, volume 2, pages 848–854, 2004. 1

[3] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, pages 1543–1550, 2011. 1, 2

[4] C. Burges. A tutorial on support vector machines for pattern recognition. *DMKD*, 2(2):121–167, 1998. 4

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm. 2001. 4

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005. 2

[7] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. In *ICCV*, pages 229–236, 2009. 4

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009. 2

[9] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, pages 524–531, 2005. 2

[10] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 4

[11] M. Fischler and R. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computer*, 100(1):67–92, 1973. 4

[12] Y. Fu, G. Guo, and T. Huang. Age synthesis and estimation via faces: A survey. *IEEE TPAMI*, 32(11):1955–1976, 2010. 1

[13] A. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, pages 1–8, 2008. 2

[14] A. C. Gallagher and T. Chen. Understanding images of groups of people. In *CVPR*, pages 256–263. IEEE, 2009. 1, 4

[15] C. Galleguillos, A. Rabinovich, and S. Belongie. Object categorization using co-occurrence, location and appearance. In *CVPR*, pages 1–8, 2008. 4

[16] T. Joachims, T. Finley, and C. Yu. Cutting-plane training of structural svms. *Machine Learning*, 77(1):27–59, 2009. 5

[17] N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Attribute and simile classifiers for face verification. In *ICCV*, pages 365–372, 2009. 2, 3

[18] S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. In *ICCV*, volume 2, pages 1284–1291, 2005. 4

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, pages 2169–2178, 2006. 2

[20] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 17–32, 2004. 4

[21] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *CVPR*, pages 3330–3337, 2012. 2

[22] M. Naaman, R. Yeh, H. Garcia-Molina, and A. Paepcke. Leveraging context to resolve identity in photo albums. In *ACM/IEEE-CS joint conference on Digital libraries*, pages 178–187, 2005. 1, 4

[23] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 2

[24] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510. IEEE, 2011. 2, 3

[25] T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *CVPR*, volume 2, pages 994–1000, 2005. 3

[26] V. Sharmanska, N. Quadrianto, and C. Lampert. Augmented attribute representations. In *ECCV*, pages 242–255, 2012. 2

[27] Z. Song, M. Wang, X. Hua, and S. Yan. Predicting occupation via human clothing and contexts. In *ICCV*, pages 1084–1091, 2011. 2, 6, 7

[28] Z. Stone, T. Zickler, and T. Darrell. Toward large-scale face recognition using social network context. *Proceedings of the IEEE*, 98(8):1408–1415, 2010. 1

[29] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *ICML*. ACM, 2004. 5

[30] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6:1453–1484, 2005. 5

[31] G. Wang, A. Gallagher, J. Luo, and D. Forsyth. Seeing people in social context: Recognizing people and social relationships. In *ECCV*, pages 169–182. Springer, 2010. 1, 4

[32] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *ECCV*, pages 155–168. Springer, 2010. 2

[33] M. Weber, M. Bauml, and R. Stiefelhagen. Part-based clothing segmentation for person retrieval. In *AVSS*, pages 361–366, 2011. 2

[34] S. Xia, M. Shao, J. Luo, and Y. Fu. Understanding kin relationships in a photo. *TMM*, 14(4):1046–1056, 2012. 1

[35] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010. 2

[36] K. Yamaguchi, M. Kiapour, L. Ortiz, and T. Berg. Parsing clothing in fashion photographs. In *CVPR*, pages 3570–3577, 2012. 2

[37] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, pages 1794–1801, 2009. 3

[38] M. Yang and K. Yu. Real-time clothing recognition in surveillance videos. In *ICIP*, pages 2937–2940. IEEE, 2011. 2

[39] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, pages 1385–1392, 2011. 2, 3

[40] X. Yuan and S. Yan. Visual classification with multi-task joint sparse representation. In *CVPR*, pages 3493–3500, 2010. 2

[41] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *Acm Computing Surveys (CSUR)*, 35(4):399–458, 2003. 1