Describing Clothing by Semantic Attributes

Huizhong Chen[1], Andrew Gallagher[2,3], and Bernd Girod[1]

Department of Electrical Engineering, Stanford University, Stanford, California[1]
Kodak Research Laboratories, Rochester, New York[2]
Cornell University, Ithaca, New York[3]

Abstract. *Describing clothing appearance with semantic attributes is an appealing technique for many important applications. In this paper, we propose a fully automated system that is capable of generating a list of nameable attributes for clothes on human body in unconstrained images. We extract low-level features in a pose-adaptive manner, and combine complementary features for learning attribute classifiers. Mutual dependencies between the attributes are then explored by a Conditional Random Field to further improve the predictions from independent classifiers. We validate the performance of our system on a challenging clothing attribute dataset, and introduce a novel application of dressing style analysis that utilizes the semantic attributes produced by our system.*

1 Introduction

Over recent years, computer vision algorithms that describe objects on the semantic level have attracted research interest. Compared to conventional vision tasks such as object matching and categorization, learning meaningful attributes offers a more detailed description about the objects. One example is the FaceTracer search engine [1], which allows the user to perform face queries with a variety of descriptive facial attributes.

In this paper, we are interested in learning the visual attributes for clothing items. As shown in Fig.1, a set of attributes is generated to describe the visual appearance of clothing on the human body. This technique has a great impact on many emerging applications, such as customer profile analysis for shopping recommendations. With a collection of personal or event photos, it is possible to infer the dressing style of the person or the event by analyzing the attributes of clothes, and subsequently make shopping recommendations. The application of dressing style analysis is shown in Fig.9.

Another important application is context-aware person identification, where many researchers have demonstrated superior performance by incorporating clothing information as a contextual cue that complements facial features [2–4]. Indeed, within a certain time frame (i.e., at a given event), people are unlikely to change their clothing. By accurately describing the clothing, person identification accuracy can be improved over conventional techniques that only rely on faces. In our study, we also found that clothing carry significant information to infer the gender of the wearer. This observation is consistent with the prior work of [5, 6], which exploits human body information to predict gender. Consequently, a better gender classification system can be developed by combining clothing information with traditional face-based gender recognition algorithms.

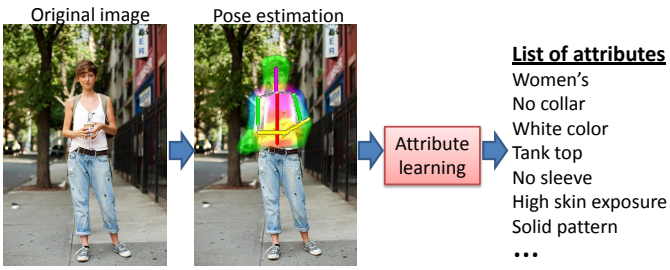

Fig. 1: With the estimated human pose, our attribute learning algorithm generates semantic attributes for the clothing.

We present a fully automatic system that learns semantic attributes for clothing on the human upper body. We take advantage of the recent advances in human pose estimation [7], by adaptively extracting image features from different human body parts. Due to the diversity of the clothing attributes that we wish to learn, a single type of feature is unlikely to perform well on all attributes. Consequently, for each attribute, the prediction is obtained by aggregating the classification results from several complementary features. Last, but not the least, since the clothing attributes are naturally correlated, we also explore the mutual dependencies between the attributes. Essentially, these mutual dependencies between various clothing attributes capture the Rules of Style. For example, neckties are rarely worn with T-shirts. To model the style rules, a conditional random field (CRF) is employed on top of the classification predictions from individual attribute classifiers, and the final list of attributes will be produced as the inference result of the CRF.

To evaluate the descriptive ability of our system, we have created a dataset with annotated images that contain clothed people in unconstrained settings. Learning clothing attributes in unconstrained settings can be extremely difficult due to the wide variety of clothing appearances, such as geometric and photometric distortions, partial occlusions, and different lighting conditions. Even under such challenging conditions, our system demonstrates very good performance. Our major contributions are as follows: 1) We introduce a novel clothing feature extraction method that is adaptive to human pose; 2) We exploit the natural rules of fashion by learning the mutual relationship between different clothing attributes, and show improved performance with this modeling; 3) We propose a new application that predicts the dressing style of a person or an event by analyzing a group of photos, and demonstrate gender classification from clothing that advocate similar findings from other researchers [5, 6].

2 Related Work

The study of clothing appearance has become a popular topic because many techniques in context-aware person identification use clothing as an important contextual cue. Anguelov et. al. [2] constructed a Markov Random Field to incorporate clothing and other contexts with face recognition. They extract clothing features by collecting color and textual information from a rectangular bounding box under the detected face, but this approach suffers from losing clothing information by neglecting the shape and variable pose of the human body. To overcome this problem, Gallagher and Chen [4]

proposed clothing cosegmentation from a set of images and demonstrated improved performance on person identification. However, clothing is described only by low-level features for the purpose of object matching, which differs from our work that learns mid-level semantic attributes for clothes.

Very recently, attribute learning has been widely applied to many computer vision tasks. In [8], Ferrari and Zisserman present a probabilistic generative model to learn visual attributes such as "red", "stripes" and "dots". With the challenge of training models on noisy images returned from Google search, their algorithm has shown very good performance. Farhadi et. al. [9] learn discriminative attributes and effectively categorize objects using the compact attribute representation. Similar work has been performed by Russakovsky and Fei-Fei [10], by utilizing attributes and exploring transfer learning for object classification on large-scale datasets. Also, for the task of object classification, Parikh and Grauman [11] build a set of attributes that is both discriminative and nameable by displaying categorized object images to a human in the loop. In [12], Kumar et. al. propose attribute and simile classifiers that describe face appearances, and have demonstrated very competitive results for the application of face verification. Siddiquie et. al. [13] explore the co-occurrence of attributes for image ranking and retrieval with multi-attribute queries. One of the major challenges of attribute learning is the lack of training data, since the acquisition of labels can be both labor-intensive and time-consuming. To overcome this limitation, Berg et. al. [14] proposed mining both the catalog images and their descriptive text from the Internet and perform language processing to discover attributes. Another interesting application is to incorporate the discovered attributes with language models to generate a sentence for an image, in a similar manner that a human might describe an image [15, 16]. Most work in the attribute learning literatures either assumes a pre-detected bounding box that contains the object of interest, or uses the image as a whole for feature extraction. Our work is different from previous work not simply because we perform attribute learning on clothing, but also due to the fact that we extract features that are adaptive to unconstrained human poses. We show that with the prior knowledge of human poses, features can be collected in a more sensible way to make better attribute predictions.

Recently, Bourdev et. al. [6] proposed a system that describes the appearance of people, using 9 binary attributes such as "is male", "has T-shirt", and "long hair". They used a set of parts from Poselets [17] for extracting low-level features and perform subsequent attribute learning. In comparison, we propose a system that comprehensively describes the clothing appearance, with 23 binary attributes and 3 multi-class attributes. We not only consider high-level attributes such as clothing categories, but also deal with some very detailed attributes like "collar presence", "neckline shape", "striped", "spotted" and "graphics". In addition, we explicitly exploit human poses by selectively changing sampling positions of our model and experimentally demonstrate the benefits of modeling pose. Further, [6] applies an additional layer of SVMs to explore the attribute correlations, whereas in our system, the attribute correlations are modeled with a CRF which benefits from computation efficiency. Besides the system of [6] that learns descriptive attributes for people, some limited work has been done to understand the clothing appearance on the semantic level. Song et. al. [18] proposed an algorithm that predicts job occupation via human clothes and contexts, showing emerging applica-

tions by investigating clothing visual appearance. Yang and Yu [19] proposed a clothing recognition system that identifies clothing categories such as "suit" and "T-shirt". In contrast, our work learns a much broader range of attributes like "collar presence" and "sleeve length". The work that explores similar attributes to ours is [20]. However, their system is built for images taken in a well controlled fitting room environment with constrained human poses in frontal view, whereas our algorithm works for unconstrained images with varying poses. In addition, their attributes predictions use hand-designed features, while we follow a more disciplined learning approach that more easily allows new attributes to be added to the model.

3 Clothing Attributes and Image Data

By surveying multiple online catalogs, we produced a list of common attributes to describe clothing. As shown in Table 1, some of these attributes are binary like "collar presence", while some are multi-class such as "clothing category". Although we tried to include a full set of attributes that comprehensively describes clothing, we dropped a few attributes like "contains logo" due to very small number of positive examples.

Clothing images and labels are required to train attribute classifiers. We have collected images from Sartorialist [1] and Flickr, by applying an upper body detector [21] to select pictures with people. Altogether, we harvested 1856 images that contain clothed people (mostly pedestrians on the streets). Some image samples from our dataset are shown in Fig.8. We then use Amazon Mechanical Turk (AMT) to collect ground truth labels. When labeling a piece of clothing on the human body, the AMT workers were asked to make a choice for each attribute. For instance, the worker may select one answer from "1) no sleeve; 2) short sleeve; 3) long sleeve" for the attribute "sleeve length". Note that for the "gender" attribute, in order to avoid the ambiguity of judging whether this piece of clothing is "men's" or "women's", workers were explicitly told to label the gender of the wearer. To eliminate noisy labels, every image was labeled by 6 workers. Only labels that have 5 or more agreements are accepted as the ground truth. We gathered 283,107 labels from the AMT workers, and the ground truth statistics of our clothing attribute dataset are summarized in Table 1.

4 Learning Clothing Attributes

The flowchart of our system is illustrated in Fig.2. For an input image, human pose estimation is performed to find the locations of the upper torso and arms. We then extract 40 features from the torso and arm regions, and subsequently quantize them. For each attribute, we perform SVM classification using the combined feature computed from the weighted sum of the 40 features. Each attribute classifier outputs a probability score that reflects the confidence of the attribute prediction. Next, a CRF is employed to learn the stylistic relationships between various attributes. By feeding the CRF with the probability scores from the attribute classifiers, the attribute relations are explored, which leads to better predictions than independently using the attribute classifiers.

[1] http://www.thesartorialist.com

Clothing pattern (Positive / Negative)	Solid (1052 / 441), Floral (69 / 1649), Spotted (101 / 1619) Plaid (105 / 1635), Striped (140 / 1534), Graphics (110 / 1668)
Major color (Positive / Negative)	Red (93 / 1651), Yellow (67 / 1677), Green (83 / 1661), Cyan (90 / 1654) Blue (150 / 1594), Purple (77 / 1667), Brown (168 / 1576), White (466 / 1278) Gray (345 / 1399), Black (620 / 1124), $>$ 2 Colors (203 / 1541)
Wearing necktie	Yes 211, No 1528
Collar presence	Yes 895, No 567
Gender	Male 762, Female 1032
Wearing scarf	Yes 234, No 1432
Skin exposure	High 193, Low 1497
Placket presence	Yes 1159, No 624
Sleeve length	No sleeve (188), Short sleeve (323), Long sleeve (1270)
Neckline shape	V-shape (626), Round (465), Others (223)
Clothing category	Shirt (134), Sweater (88), T-shirt (108), Outerwear (220) Suit (232), Tank Top (62), Dress (260)

Table 1: Statistics of the clothing attribute dataset. There are 26 attributes in total, including 23 binary-class attributes (6 for pattern, 11 for color and 6 miscellaneous attributes) and 3 multi-class attributes (sleeve length, neckline shape and clothing category)

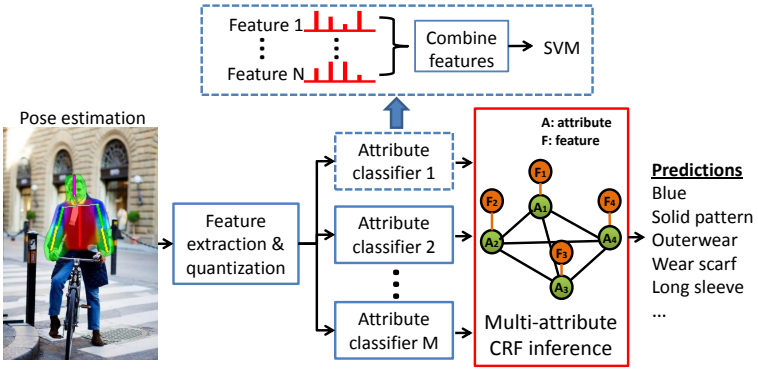

Fig. 2: Flowchart of our system. Several types of features are extracted based on the human pose. These features are combined based on their predictive power to train a classifier for each attribute. A Conditional Random Field that captures the mutual dependencies between the attributes is employed to make attribute predictions. The final output of the system is a list of nameable attributes that describe the clothing appearance.

4.1 Human Pose Estimation

Thanks to the recent progress in human pose estimation [7, 22, 23], the analysis of complex human poses has been made possible. With the knowledge of the physical pose of the person, clothing attributes can be learned more effectively, e.g., features from the arm regions offer a strong clue for sleeve length.

Estimating the full body pose from 2-D images still remains a challenging problem, partly because the lower body is occasionally occluded or otherwise not visible in some images. Consequently, we only consider the clothing items on the upper body. We apply the work in [7] for human pose estimation and briefly review the technique here for the completeness of the paper. Given an input image, the upper body of the person is firstly located by using complementary results of an upper body detector [21] and Viola-Jones face detector [24]. The bounding box of the detected upper body is then enlarged, and the GrabCut algorithm [25] is used to segment the person from the background. Person-specific appearance models for different body parts are estimated within the detected upper body window. Finally, an articulated pose is formed within the segmented person area, by using the person-specific appearance models and generic appearance models.

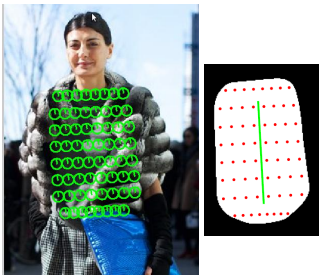

Fig. 3: Extraction of SIFT over the torso region. For better visualization, we display fewer descriptors than the actual number. The left figure shows the location, scale and orientation of the SIFT descriptors in green circles. The right figure depicts the relative positions between the sampling points (red dots), the torso region (white mask) and the torso stick (green bar).

The outputs of the pose estimation algorithm are illustrated in Fig.1, which include a stick-man model and the posterior probability map of the six upper body regions (head, torso, upper arms and lower arms). We threshold the posterior probability map to obtain binary masks for the torso and arms, while ignoring the head region since it is not related to clothing attribute learning.

4.2 Feature Extraction

Due to the large number of attributes that we wish to learn, the system is unlikely to achieve optimal performance with a single type of feature. For example, while texture descriptors are useful for analyzing the clothing patterns such as "striped" and "dotted", they are not suitable for describing clothing colors. Therefore, in our implementation we use 4 types of base features, including SIFT [26], texture descriptors from the Maximum Response Filters [27], color in the LAB space, and skin probabilities from our skin detector.

As mentioned in Section 1, the features are extracted in a pose-adaptive way. The sampling location, scale and orientation of the SIFT descriptors depend on the estimated human body parts and the stick-man model. Fig.3 illustrates the extraction of the SIFT descriptors over the person's torso region. The sampling points are arranged according to the torso stick and the boundary of the torso. The configuration of the torso sampling points is a 2-D array, whose size is given by the number of samples along the stick direction times the number of samples normal to the stick direction. The scale of the SIFT descriptor is determined by the size of the torso mask, while the orientation is simply chosen as the direction of the torso stick. The extraction of SIFT features over the 4 arm regions is done in a similar way, except that the descriptors are only sampled along the arm sticks. In our implementation, we sample SIFT descriptors at 64×32 locations for the torso, and 32 locations along the arm stick for each of the 4 arm segments. The remaining three of our base features, namely textures descriptors, color and skin probabilities, are computed for each pixel in the five body regions.

Once the base features are computed, they are quantized by soft K-means with 5 nearest neighbors. As a general design guideline, features with larger dimensionality should have more quantization centroids. Therefore we allocate 1000 visual words for SIFT, 256 centroids for texture descriptor, and 128 centroids for color and skin probability. We employed a Kd-tree for efficient feature quantization. The quantized features

Feature type	Region	Aggregation
SIFT	Torso	Average pooling
Texture descriptor	Left upper arm	Max pooling
Color	Right upper arm	
Skin probability	Left lower arm	
	Right lower arm	

Table 2: Summary of feature extraction.

are then aggregated by performing average pooling or max pooling over the torso and arm regions. Note that the feature aggregation for torso is done by constructing a 4×4 spatial pyramid [28] over the torso region and then average or max pooled.

Table 2 summarizes the features that we extract. In total, we have 40 different features from 4 feature types, computed over 5 body regions with 2 aggregation methods.

Last, but not the least, we extract one more feature that is specifically designed for learning clothing color attributes, which we call skin-excluded color feature. Obviously, the exposed skin color should not be used to describe clothing colors. Therefore we mask out the skin area (generated by our skin detector) to extract the skin-excluded color feature. As the arms may contain purely skin pixels when the person is wearing no-sleeve clothes, it may not be feasible to aggregate skin-excluded color feature for the arm regions. Hence we perform aggregation over the whole upper body area (torso + 4 arm segments), excluding the skin region. Also, since the maximum color responses are subject to image noise, only average pooling is adopted. So compared to the set of 40 features described above, we only use 1 feature for learning clothing color.

4.3 Attribute Classification

We utilize Support Vector Machines (SVMs) [29] to learn attributes since they demonstrate state-of-the-art performance for classification. For clothing color attributes like "red" and "blue", we simply use the skin-excluded color feature to train one SVM per color attribute. For each of the other clothing attributes such as "wear necktie" and "collar presence", we have 40 different features and corresponding weak classifiers but it is uncertain which feature offers the best classification performance for the attribute that we are currently learning. A naive solution would be to concatenate all features into a single feature vector and feeding to an SVM for attribute classification. However, this approach suffers from three major drawbacks: 1) the model is likely to overfit due to the extremely high dimensional feature vector; 2) due to the high feature dimension, classification can be slow; and most importantly, 3) within the concatenated feature vector, high dimensional features will dominate over low dimensional ones so the classifier performance is similar to the case when only high dimensional features are used. Indeed, in our experiments we found that the concatenated feature vector offers negligible improvements over the SVM that uses only the SIFT features.

Another attribute classification approach is to train a set of 40 SVMs, one for each feature type, and pick the best performing SVM as the attribute classifier. This method certainly works, but we are interested in achieving a even better performance with all the available features. Using the fact that SVM is a kernel method, we form a combined kernel from the 40 feature kernels by weighted summation, where the weights correspond to the classification performance of the features. Intuitively, better feature kernels are assigned heavier weights than weaker feature kernels. The combined kernel

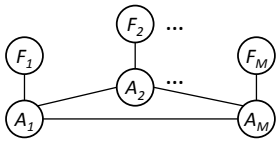

Fig. 4: A CRF model with M attribute nodes pairwise connected. F_i denotes the feature for inferring attribute i.

is then used to train an SVM classifier for the attribute. This method is inspired by the work in [30], where they demonstrated significant scene classification improvement by combining features. Our experimental results in Section 5.1 also show the advantage offered by feature combination.

4.4 Multi-attribute Inference

Due to the functionality and fashion of clothing, it is common to see correlations between clothing attributes. As an example, in our ground truth dataset there is only 1 instance when the person wears a necktie but does not have a collar. It should be noted that the dependencies between the attributes are not symmetric, e.g., while having a necktie strongly suggest collar presence, the presence of a collar does not indicate that the person should wear a necktie.

We explore the mutual dependencies between attributes by applying a CRF with the SVM margins from the previous attribute classification stage. Each attribute functions as a node in the CRF, and the edge connecting every two attribute nodes reflects the joint probability of these two attributes. We build a fully connected CRF with all attribute nodes pairwise connected, as shown in Fig.4.

Let us consider the relation between two attributes, A_1 and A_2. F_1 and F_2 are the features that we use to infer A_1 and A_2 respectively. Our goal is to maximize the conditional probability $P(A_1, A_2|F_1, F_2)$:

$$P(A_1, A_2|F_1, F_2) = \frac{P(F_1, F_2|A_1, A_2)P(A_1, A_2)}{P(F_1, F_2)} \tag{1}$$

$$\propto P(F_1, F_2|A_1, A_2)P(A_1, A_2) \tag{2}$$

$$= P(F_1|A_1)P(F_2|A_2)P(A_1, A_2) \tag{3}$$

$$\propto \frac{P(A_1|F_1)}{P(A_1)} \frac{P(A_2|F_2)}{P(A_2)} P(A_1, A_2) \tag{4}$$

Equation 3 is consistent with our CRF model in Fig.4, assuming that the observed feature F_i is independent of all other features once the attribute A_i is known. From the ground truth of the training data, we can estimate the joint probability $P(A_1, A_2)$ as well as the priors $P(A_1)$ and $P(A_2)$. The conditional probabilities $P(A_1|F_1)$ and $P(A_2|F_2)$ are given by the SVM probability outputs for A_1 and A_2 respectively. We define the unary potential $\Psi(A_i) = -\log\left(\frac{P(A_i|F_i)}{P(A_i)}\right)$ and the edge potential $\Phi(A_i, A_j) = -\log(P(A_i, A_j))$. Following Equation 1, the optimal inference for (A_1, A_2) is achieved by minimizing $\Psi(A_1)+\Psi(A_2)+\Phi(A_1, A_2)$, where the first two terms are the unary potentials associated with the nodes, and the last term is the edge potential that describes the relation between the attributes.

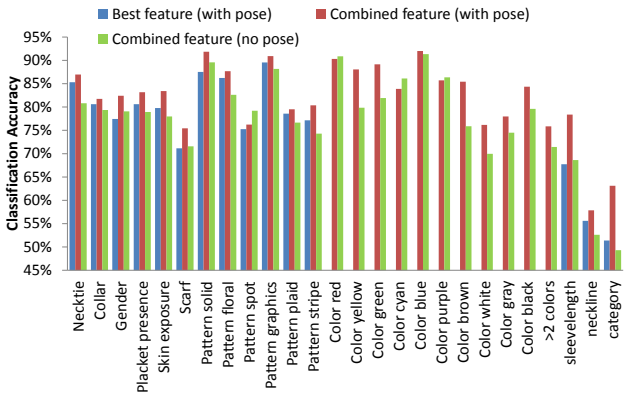

Fig. 5: Comparison of attribute classification under 3 scenarios.

Actual \ Predict	No sleeve	Short sleeve	Long sleeve
No sleeve	86.17%	10.11%	3.72%
Short sleeve	23.40%	64.90%	11.70%
Long sleeve	5.85%	10.11%	84.04%

(a) Sleeve length

Actual \ Predict	V-shape	Round	Other style
V-shape	67.27%	13.45%	19.28%
Round	24.22%	51.56%	24.22%
Other style	24.66%	20.63%	54.71%

(b) Neckline shape

Actual \ Predict	Shirt	Sweater	T-shirt	Outerwear	Suit	Tank top	Dress
Shirt	43.56%	19.35%	1.61%	8.06%	8.06%	9.68%	9.68%
Sweater	14.52%	54.84%	0%	16.13%	6.45%	1.61%	6.45%
T-shirt	4.84%	1.61%	80.65%	3.23%	0%	6.44%	3.23%
Outerwear	12.90%	6.45%	0%	61.30%	8.06%	4.84%	6.45%
Suit	9.68%	8.06%	0%	16.13%	66.13%	0%	0%
Tank top	4.84%	1.61%	0%	1.61%	0%	79.04%	12.90%
Dress	4.84%	11.28%	3.23%	3.23%	3.23%	17.74%	56.45%

(c) Clothing category

Fig. 6: Multi-class confusion matrices. Predictions are made using the combined feature extracted with the pose model.

For a fully connected CRF with a set of nodes S and a set of edges E, the optimal graph configuration is obtained by minimizing the graph potential, given by:

$$\sum_{A_i \in S} \Psi(A_i) + \lambda \sum_{(A_i, A_j) \in E} \Phi(A_i, A_j) \tag{5}$$

where λ assigns relative weights between the unary potentials and the edge potentials. It is typically less than 1 because a fully connected CRF normally contains more edges than nodes. The actual value of λ can be optimized by cross validation. In our implementation, we use belief propagation [31] to minimize the attribute label cost.

5 Experiments

We have performed extensive evaluations of our system on the clothing attribute dataset. In Section 5.1, we show that our pose-adaptive features offer better classification accuracies compared to the features extracted without a human pose model. The results from section 5.2 demonstrates that the CRF improves attribute predictions since it explores relations between the attributes. Finally, we show some potential applications that directly utilize the output of our system.

5.1 Attribute Classification

We use the chi-squared kernel for the SVM attribute classifiers since it outperforms both the linear and the RBF kernels in our experiments. To examine the classification

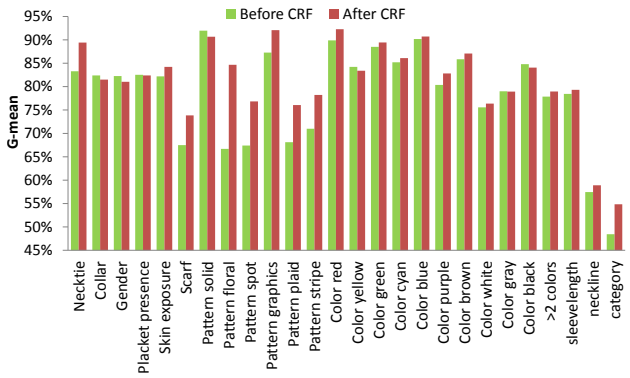

Fig. 7: Comparison of G-means before and after the CRF.

accuracy, we partition the data such that each attribute has equal number of examples in all classes. For example, we balance the data of the "wearing necktie" attribute so that it has the same number of positive and negative examples. We use leave-1-out cross validation to report the classification accuracy for each attribute.

Fig.5 compares the SVM performances with three types of feature inputs: 1) With pose model, using the best feature out of our 40 features; 2) With pose model, combining the 40 features with the method described in Section 4.3; 3) No pose model, same experiment settings as case 2 but with features extracted within a scaled clothing mask [4] under the face. Note that the single best feature accuracies for the the color attributes are not displayed in Fig.5, because color attribute classifiers only use the skin-excluded color feature, which we regard as the "combined" feature for color attributes.

First of all, we observe that the SVM classifiers with combined features always perform better than those using the single best feature. This is because the combined feature utilizes the complementary information of all extracted features. More importantly, we are interested to know whether considering human pose helps the classification of clothing attributes. As can be seen in Fig.5, pose-adaptive features offer improved performance for most attributes, except a slight decrease in performance for 4 attributes in colors and patterns. In particular, for those attributes like "sleeve length" that heavily depend on human poses, we observe a significant boost in performance with pose-adaptive features, compared to the classifier performance of using features without prior knowledge of the human pose.

We also show the confusion matrices of the multi-class attributes in Fig.6, where the attributes are predicted using the combined feature extracted with the pose model. The misclassification patterns of the confusion matrices are plausible. For example, "no sleeve" is misclassified more often as "short sleeve" than "long sleeve", and "tank top" is more often confused with "dress" than with other clothing categories.

5.2 Exploiting Attribute Relations with CRF

We apply a CRF as described in Section 5.2 to perform multi-attribute inference. As shown in Equation 1, the CRF uses the prior probability of each attribute, and the prior

Fig. 8: For each image, the attribute predictions from independent classifiers are listed. Incorrect attributes are highlighted in red. We describe the abnormal attributes generated by independent classifiers for the above images, from left to right: man in dress, suit with high skin exposure, no sleeves but wearing scarf, T-shirt with placket. By exploring attribute dependencies, the CRF corrects some wrong attribute predictions. Attributes that are changed by the CRF are shown in parentheses, and override the independent classifier result to the left.

reflects the proportion of each attribute class in the dataset. Therefore, instead of testing the classifier performance on the balanced data, we evaluate the classification performance on the whole clothing attribute dataset. Since this is an unbalanced classification problem, classification accuracy is no longer a proper evaluation metric. We use the Geometric Mean (G-mean), which is a popular evaluation metric for unbalanced data classification.

$$\text{G-mean} = \left(\prod_{i=1}^{N} R_i \right)^{\frac{1}{N}} \tag{6}$$

where R_i is the retrieval rate for class i, and N is the number of classes.

The G-means before and after applying the CRF are shown in Fig.7. It can be seen that the CRF offers better predictions for 19 out of 26 attributes, sometimes providing large margins of improvement. For those attributes that the CRF fails to improve the performance, only very minor degradations are observed. Overall, better classification performance is achieved by applying the CRF on top of our attribute classifiers. Fig.8 shows the attribute predictions of 4 images sampled from our dataset. As can be seen in Fig.8, the CRF corrects some conflicting attributes that are generated by independent classifiers. More clothing attribute illustration examples can be found here [2].

5.3 Applications

Dressing Style Analysis With a collection of customer photos, our system can be used to analyze the customer's dressing style and subsequently make shopping recommendations. For each attribute, the system can tell the percentage of its occurrences in the group of photos. Intuitively, in order for an attribute class to be regarded as a personal style, this class has to be observed much more frequently compared to the prior of the general public. As an example, if a customer wears plaid-patterned clothing three days

[2] https://sites.google.com/site/eccv2012clothingattribute/

Fig. 9: With a set of person or event photos (sample images shown above), our system infers the dressing style of a person or an event. Most of our predicted dressing styles are quite reasonable. The wrong predictions are highlighted in red.

a week, then wearing plaid is probably his personal style, because the frequency that he is observed in plaid is much higher than the general public's prior of wearing plaid (the prior of plaid is 6.4% according to our clothing attribute dataset). In our analysis, we regard an attribute class as a dressing style if it is 20% higher than its prior. Of course, this threshold can be flexibly adjusted based on specific application requirements.

We perform dressing style analysis on both personal and event photos, as shown in Fig.9. Firstly, we analyze the dressing style of Steve Jobs, who is well known for wearing his black turtlenecks. Using 35 photos of Jobs from Flickr, our system summarizes his dressing style as "**solid pattern, men's clothing, black color, long sleeves, round neckline, outerwear, wear scarf**". Most of the styles predicted by our system are very sensible. The wrong inferences of "outerwear" and "wearing scarf" are not particularly surprising, since Steve Job's high-collar turtlenecks share visual similarities with outerwears and the presence of scarfs. Similarly, our system predicts the dressing style of Mark Zuckerberg as "**gray, brown, T-shirt, outerwear**", which is in agreement with his actual dressing style.

Apart from personal clothing style analysis, we can also analyze the dressing style for events. We downloaded 54 western-style wedding photos from Flickr, and consider the dressing styles for men and women using the "gender" attribute predicted by our system. The dressing style for men at wedding is predicted as: "**solid pattern, suit, long-sleeves, V-shape neckline, wearing necktie, wear scarf, has collar, has placket**", while the dressing style for women is "**high skin exposure, no sleeves, dress, other neckline shapes** (i.e. neither v-shape nor round)**, white,** $>$**2 colors, floral pattern**". Most of these predictions agree well with the dressing style of weddings, except "wearing scarf" for men, and ">2 colors" and "floral pattern" for women. These abnormal predictions can be understood, by considering the similarity between men's scarfs and neckties, as well as the visual confusion for including the wedding flowers in women's hands when describing the color and pattern of their dresses. Similar analysis was done to predict the clothing style of the event of basketball. Using 61 NBA photos,

the basketball players' dressing style is predicted as "**no sleeves, high skin exposure, other neckline shapes, tank top**".

Gender Classification Although there exist clothes that are unisex style, visual appearance of clothes often carries significant information about gender. For example, men are less likely to wear floral-patterned clothes, and women rarely wear a tie. Motivated by this observation, we are interested in combining the gender prediction from clothing information, with traditional gender recognition algorithms that use facial features. We adopt the publicly available gender-from-face implementation of [32], which outputs gender probabilities by projecting aligned faces in the Fisherface space. For the clothing-based gender classification, we use the graph potential in Equation 5 as the penalty score for assigning gender to a testing instance. The male and female prediction penalties are simply given by the CRF potentials, by assigning male or female to the "gender" node, while keeping all other nodes unchanged. An RBF-kernel SVM is trained to give gender predictions that combine both the gender probabilities from faces and penalties from clothing.

We evaluate the gender classification algorithms on our clothing attribute dataset. As shown Table 3, the combined gender classification algorithm offers a better performance than that of using each feature alone. Interestingly, clothing-only gender recognition outperforms face-only gender recognition.

	G-mean
Face only	0.715
Clothing only	0.810
Face + clothing	0.849

Table 3: Performance of gender classification algorithms.

6 Conclusions

We propose a fully automated system that describes the clothing appearance with semantic attributes. Our system demonstrates superior performance on unconstrained images, by incorporating a human pose model during the feature extraction stage, as well as by modeling the rules of clothing style by observing co-occurrences of the attributes. We also show novel applications where our clothing attributes can be directly utilized. In the future, we expect to observe even more improvements on our system, by employing the (almost ground truth) human pose estimated by Kinect sensors [23].

References

1. Kumar, N., Belhumeur, P.N., Nayar, S.K.: Facetracer: A search engine for large collections of images with faces. In: ECCV. (2008)
2. Anguelov, D., Lee, K., Gokturk, S.B., Sumengen, B.: Contextual identity recognition in personal photo albums. In: CVPR. (2007)
3. Lin, D., Kapoor, A., Hua, G., Baker, S.: Joint people, event, and location recognition in personal photo collections using cross-domain context. In: ECCV. (2010)
4. Gallagher, A.C., Chen, T.: Clothing cosegmentation for recognizing people. In: CVPR. (2008)

5. Cao, L., Dikmen, M., Fu, Y., Huang, T.S.: Gender recognition from body. In: ACM Multimedia. (2008)
6. Bourdev, L., Maji, S., Malik, J.: Describing people: Poselet-based attribute classification. In: ICCV. (2011)
7. Eichner, M., Marin-Jimenez, M., Zisserman, A., Ferrari, V.: Articulated human pose estimation and search in (almost) unconstrained still images. Technical Report 272, ETH Zurich, D-ITET, BIWI (2010)
8. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS. (2007)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
10. Russakovsky, O., Fei-Fei, L.: Attribute learning in large-scale datasets. In: ECCV. (2010)
11. Parikh, D., Grauman, K.: Interactively building a discriminative vocabulary of nameable attributes. In: CVPR. (2011)
12. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: ICCV. (2009)
13. Siddiquie, B., Feris, R.S., Davis, L.S.: Image ranking and retrieval based on multi-attribute queries. In: CVPR. (2011)
14. Berg, T.L., Berg, A.C., Shih, J.: Automatic attribute discovery and characterization from noisy web data. In: ECCV. (2010)
15. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Berg, A., Choi, Y., Berg, T.: Baby talk: Understanding and generating image descriptions. In: CVPR. (2011)
16. Farhadi, A., Hejrati, M., Sadeghi, A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.A.: Every picture tells a story: Generating sentences for images. In: ECCV. (2010)
17. Bourdev, L., Malik, J.: Poselets: Body part detectors trained using 3d human pose annotations. In: ICCV. (2009)
18. Song, Z., Wang, M., Hua, X., Yan, S.: Predicting occupation via human clothing and contexts. In: ICCV. (2011)
19. Yang, M., Yu, K.: Real-time clothing recognition in surveillance videos. In: ICIP. (2011)
20. Zhang, W., Begole, B., Chu, M., Liu, J., Yee, N.: Real-time clothes comparison based on multi-view vision. In: ICDSC. (2008)
21. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. PAMI (2009)
22. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR. (2011)
23. Shotton, J., Fitzgibbon, A., Cook, M., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR. (2011)
24. Viola, P., Jones, M.: Robust real-time object detection. IJCV (2001)
25. Rother, C., Kolmogorov, V., Blake, A.: Grabcut - interactive foreground extraction using iterated graph cuts. In: SIGGRAPH. (2004)
26. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
27. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. IJCV (2005)
28. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. (2006)
29. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans. on Intel. Sys. and Tech. (2011)
30. Xiao, J., Hays, J., Ehinger, K.A., Torralba, A., Oliva, A.: Sun database: Large scale scene recognition from abbey to zoo. In: CVPR. (2010)
31. Tappen, M.F., Freeman, W.T.: Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In: ICCV. (2003)
32. Gallagher, A.C., Chen, T.: Understanding images of groups of people. In: CVPR. (2009)