Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Learning to Appreciate the Aesthetic Effects of Clothing

Jia Jia[1], **Jie Huang**[1], **Guangyao Shen**[1], **Tao He**[2], **Zhiyuan Liu**[1*], **Huanbo Luan**[1] and **Chao Yan**[3]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Tsinghua National Laboratory for Information Science and Technology (TNList)
Key Laboratory of Pervasive Computing, Ministry of Education
[2]School of Computer Science, Sichuan University, Chengdu 610065
[3]Beijing Samsung Telecom R& D center
{jjia, liuzy}@mail.tsinghua.edu.cn

Abstract

How do people describe clothing? The words like "formal" or "casual" are usually used. However, recent works often focus on recognizing or extracting visual features (e.g., sleeve length, color distribution and clothing pattern) from clothing images accurately. How can we bridge the gap between the visual features and the aesthetic words? In this paper, we formulate this task to a novel three-level framework: visual features (VF) - image-scale space (ISS) - aesthetic words space (AWS). Leveraging the art-field image-scale space served as an intermediate layer, we first propose a Stacked Denoising Autoencoder Guided by Correlative Labels (SDAE-GCL) to map the visual features to the image-scale space; and then according to the semantic distances computed by WordNet::Similarity, we map the most often used aesthetic words in online clothing shops to the image-scale space too. Employing upper-body menswear images downloaded from several global online clothing shops as experimental data, the results indicate that the proposed three-level framework can help to capture the subtle relationship between visual features and aesthetic words better compared to several baselines. To demonstrate that our three-level framework and its implementation methods are universally applicable, we finally present some interesting analyses on the fashion trend of menswear in the last 10 years.

1 Introduction

Apparel makes the man. It has been documented that people reinforce their mood and express their feelings through their clothing (Kang, Johnson, and Kim 2013). A variety of researches have made it possible to extract or recognize the visual features (e.g., sleeve length, color distribution and clothing pattern) from clothing images accurately (Yang and Ramanan 2011; Yamaguchi, Kiapour, and Berg 2013; Yang, Luo, and Lin 2014). But how do people describe clothing? The aesthetic words like "formal" or "casual" are usually used rather than comments like the sleeves are long or the collar is round. These aesthetic words are obviously related to the visual features. For example, suits with more than three buttons look formal, while tank tops seem casual. If it is possible to understand the aesthetic effects of clothing based on the visual features automatically, there will be significant progress in many applications such as clothing recommendation systems. That means we can enable the computer to learn to appreciate the aesthetic effects of clothing.

However, fulfilling the task is not a trivial issue. Focusing on clothing segmentation and recognition, (Hasan and Hogg 2010) presents a method for segmenting the parts of multiple instances using deformable spatial priors. (Wang and Ai 2011) studies on simultaneous clothing segmentation for grouping images. (Yamaguchi et al. 2015) proposes to tackle the clothing parsing problem using a retrieval-based approach. In recent years, scenario-oriented and occasion-oriented clothing recommendation have attracted increasing attentions, which shows that people begin to focus on higher level semantics related to clothing rather than low-level visual features. (Shen, Lieberman, and Lam 2007) proposes the scenario-oriented recommendation to satisfy users' personal preference. (Yu-Chu et al. 2012) proposes a system to recommend an appropriate combination from the user's available clothing options according to the current situation (e.g., the weather and the user's schedule). Both works above focus on making users wear properly. However, it is also quite significant for people to wear aesthetically. (Liu et al. 2012) takes two criteria, wearing properly and aesthetically into consideration. (Kouge et al. 2015) obtains the associated rules from color combinations to derive impressions. In the field of aesthetics, their works train some matching rules (e.g., a red T-shirt matches white pants better than green ones) to ensure that there is no strange collocation. Nevertheless, the matching rules cannot reveal the aesthetic effects holistically and lack interpretability.

In this paper, we aim to bridge the gap between the visual features and the aesthetic words of clothing. In order to capture the intrinsic and holistic relationship between them, we introduce an intermediate layer and form a novel three-level framework. The low level is visual features (VF) of clothing images, including color features and pattern features. The middle level is the image-scale space (ISS) based on the aesthetic theory proposed by (Kobayashi 1995), which is a two-dimensional space (warm-cool and hard-soft) well applied in art design. The high level is the aesthetic words space (AWS) consisting of words like "formal" and "casual". Specifically, we propose a Stacked Denoising Autoencoder Guided by

[*]Corresponding author: Z. Liu (liuzy@mail.tsinghua.edu.cn)

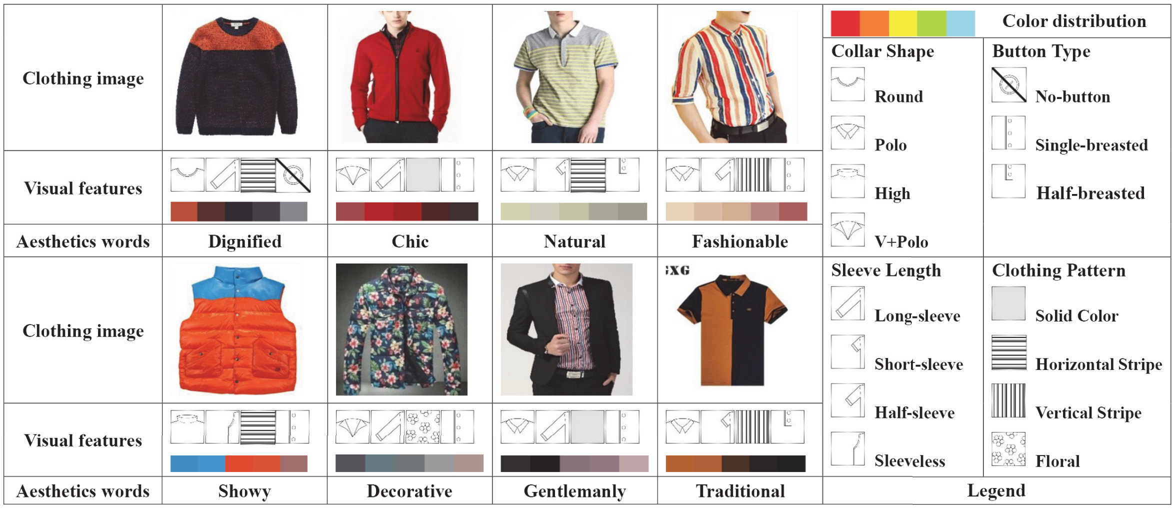


Figure 1: Examples of clothing images and their corresponding aesthetic words.

Correlative Labels (SDAE-GCL) to map the visual features to the image-scale space. Then, by computing the semantic distances using WordNet::Similarity, we map the most often used aesthetic words in online clothing shops to the image-scale space too. Thus, we implement our three-level framework by mapping both the low level and high level to the middle level. Employing upper-body menswear images downloaded from several global online clothing shops as our experimental data, we first conduct several experiments to evaluate the mapping effects between visual features and coordinate values in the image-scale space. The results indicate that the proposed SDAE-GCL can reduce 4% to 10% in terms of MSE (Mean squared error) and MAE (Mean absolute error) than baselines. We then present some examples of clothing images and their corresponding aesthetic words to demonstrate the effectiveness of the whole three-level framework (Shown in Figure 1). Finally some interesting cases of fashion trend analysis are shown to prove the proposed framework is universally applicable.

We summarize our contributions as follows:

- We build the association between clothing images and aesthetic words by proposing a three-level framework (VF-ISS-AWS). This framework introduces a novel notion of using the two-dimensional continuous image-scale space as an intermediate layer with a strong ability of description, thus facilitating the deep and high-level understanding of aesthetic effects.
- We propose a Stacked Denoising Autoencoder Guided by Correlative Labels (SDAE-GCL) to implement the mapping from visual features to the image-scale space. Specifically, the SDAE-GCL can amend the random error existing in initial input and make full use of the information of both labeled and unlabeled data. Moreover the stacked methods improve the representation capability of the model by adding more hidden layers.

2 Problem Formulation

Given a set of clothing images V, we divide it into two sets V^L (labeled data) and V^U (unlabeled data). For each image $v_i \in V$, we use a N_q dimensional vector $q_i = \langle q_{i1}, q_{i2}..q_{iN_q} \rangle$ $(\forall q_{ij} \in \mathbb{R})$ to indicate v_i's color features (e.g., brightness, saturation), a N_p dimensional vector $p_i = \langle p_{i1}, p_{i2}..p_{iN_p} \rangle$ $(\forall p_{ij} \in \mathbb{R})$ to indicate v_i's pattern features (e.g., button type, collar shape, clothing pattern), and a N_c dimensional vector $c_i = \langle c_{i1}, c_{i2}..c_{iN_c} \rangle$ $(\forall c_{ij} \in \mathbb{R})$ to indicate v_i's clothing categories (e.g., suit, sweater, shirt). In addition, Q is defined as a $|V| * N_q$ feature matrix with each element q_{ij} denoting the jth color feature of v_i. The definitions of P and C are similar to Q.

Definition 1. **The Middle Level Image-Scale Space** D is a two-dimensional space (*warm-cool* and *hard-soft*), denoted as $D(wc, hs)$ $(\forall wc, hs \in [-1, +1])$. The horizontal axis represents warmest to coolest with coordinate value wc varying from -1 to +1, while the vertical axis represents hard-soft with hs.

Definition 2. **The High Level Aesthetic Words Space** Y contains a series of aesthetic words (e.g., graceful, casual, sporty), which can be further divided into n clusters. Each cluster consists of several synonymic aesthetic words and n depends on the actual need of applications.

Problem. **Labeling the clothing images with aesthetic words**. The proposed three-level framework is implemented through 2 steps. 1) Learning a prediction model $M : (V^L, V^U, Q, P, C) \Rightarrow D$. 2) Determining a function f:$Y \Rightarrow D$. For an input image $v_i \in V$, we calculate $D_i(wc_i, hs_i)$ by model M and select a word y_i, whose two-dimensional coordinate value $D_{y_i}(wc_{y_i}, hs_{y_i})$ has least Euclidean distance to $D_i(wc_i, hs_i)$, as the aesthetic label of v_i.

3 Methods

In this section, we present the proposed three-level framework (VF-ISS-AWS) in detail. We first illustrate how we extract the low-level visual features from clothing images. Then we propose a Stacked Denoising Autoencoder Guided by Correlative Labels (SDAE-GCL) to map the visual features to the middle-level image-scale space. Finally, we introduce how we map the words, which are related to high-level aesthetic effects, to the image-scale space.

3.1 Feature Extraction

We define the visual features of clothing from two aspects: the color features and the pattern features. The color features reflect the overall perception of clothing images, including five-color combination (Wang et al. 2014), saturation and its contrast, brightness and its contrast, warm or cool color and clear or dull color. The pattern features describe the local style of clothing images, including collar shape, clothing pattern, sleeve length and button type.

We first extract the mask to separate clothing from its background in the image using the method proposed by (Liu et al. 2015). Then, for extracting the color features, we adopt the interpretable aesthetic features and extracting algorithms proposed by (Wang et al. 2013). For extracting the pattern features, we use the method proposed by (Szegedy et al. 2014) and train a CNN model to get these features.

3.2 Mapping Visual Features to Image-scale Space

In order to map visual features to the image-scale space, we divide the task into two steps. Firstly, we propose a Stacked Denoising Autoencoder Guided by Correlative Labels (SDAE-GCL) for feature learning. Secondly, we make the new constructed features cast into two-dimensional coordinates in image-scale space, which can be considered as a regression problems.

The Motivation of SDAE-GCL. Although the above-mentioned feature extraction algorithms perform well, the visual features cannot avoid errors. Therefore, we first consider to adopt Denoising Autoencoder (DAE) to make learned feature representations robust to partial corruption of the input (Vincent et al. 2008). The traditional DAE first maps a corrupted input vector to a hidden representation and then maps it back to a reconstructed feature vector, which is optimized to be similar to the initial input vector. The hidden representation can be regarded as the new feature representation, which serves as the input of the regression model. Based on the traditional DAE, we further design specific extensions, according to the following two aspects:

1) **Deep Stacked Denoising Autoencoder (SDAE)**: A critical challenge of our task is how to unveil the complicated relationship between the visual features and the aesthetic effects. The traditional DAE has only one hidden layer. Obviously, increasing the number of hidden layers can enhance the modeling ability to handle complicated work in sophisticated feature learning. Therefore, we import extra hidden layers in the DAE to get the SDAE.

2) **Denoising Autoencoder Guided by Correlative Labels (DAE-GCL)**: Our investigation finds out that for different categories, even clothing images with similar visual features can present different aesthetic effects, meaning that aesthetic effects are strongly correlated with clothing category. However, the clothing category is a kind of high-level semantics, which has intrinsic difference with the visual features. If we join the categories and visual features together as the input features of DAE directly, the training process will be interfered to produce a unexpected performance. In order to take full advantage of clothing categories, we consider clothing categories as *correlative labels* and promote the DAE to a novel structure named DAE-GCL.

Combining the two extensions above, we propose a deep Stacked DAE-GCL (SDAE-GCL) as shown in Figure 2.

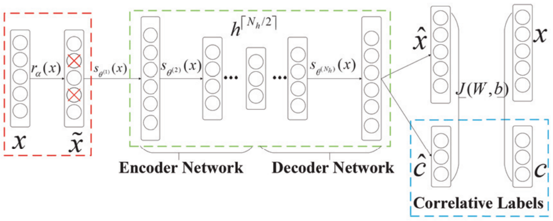


Figure 2: The structure of the SDAE-GCL model. In the left red box, the corrupted features are got from initial input through $r_\alpha(x)$, and the dimensions with red cross in $\widetilde{x}$ are forced to 0. The N_h hidden layers contain the encoder network and the decoder network, which are in middle green box. To reduce the difference between $\{\hat{x}, \hat{c}\}$ and $\{x, c\}$, we minimize the cost function $J(W, b)$ by tuning the parameters. The correlative labels are added to influence the training process. After training, the middle layer $h^{\lceil N_h/2 \rceil}$ is considered as output of our SDAE-GCL.

The Structure of SDAE-GCL. Given an image $v_i \in \{V^L, V^U\}$, the initial input vector c_i represents the clothing category vector and x_i represents the visual feature vector consisting of two parts: the color features q_i and pattern features p_i. For denoising, we set some dimensions to 0 randomly in x_i to get the partial corrupted features $\widetilde{x}_i = r_\alpha(x_i)$, where α is the proportion of the corrupted dimensions. Then we use a multilayer neural network to rebuild $\widetilde{x}_i$ into $z_i = \{\hat{x}_i, \hat{c}_i\}$, where $\hat{x}_i$ and $\hat{c}_i$ are optimized to be similar to the initial input vector x_i and c_i specifically. The hidden layers of the SDAE-GCL contain encoder network and decoder network as illustrated in the green box of Figure 2. The relationship between two adjacent layers depends on model parameters. After training, we determine the parameters and learn the intermediate representation as the output of this step.

Compared to classical autoencoder, we introduce correlative labels c_i into the original symmetrical structure as shown in blue box of Figure 2. We use the neural network to regain the correlative labels c_i, while reconstructing x_i in parallel using a shared representation. In this way, the clothing categories can be leveraged to help to discover the correlativity between various visual features and make the training process more targeted.

Formally, supposing the SDAE-GCL has N_h layers, the recursion formula between two adjacent layers is:

$$h_i^{(l+1)} = s(W^{(l)} h_i^{(l)} + b^{(l)}) \tag{1}$$

Algorithm 1 Stacked Denoising Autoencoder Guided by Correlative Labels

```
Input: $X = \{x_1, x_2 ... x_m\}$, a feature matrix of all samples.
    $C = \{c_1, c_2 ... c_m\}$, a clothing category set of all samples.
Output: The middle hidden layer features $h_i^{(\lceil N_h/2 \rceil)}$
Initialize model parameters $\theta^{(l)}, \alpha, \lambda_1, \lambda_2, \lambda_3, \beta$
Destroy the initial input partial, $h_i^{(1)} = \widetilde{x}_i = r_\alpha(x_i)$
repeat
    $W^{(l)} = W^{(l)} - \beta \frac{\partial}{\partial W^{(l)}} J(W, b)$
    $b^{(l)} = b^{(l)} - \beta \frac{\partial}{\partial b^{(l)}} J(W, b)$
until convergence (Gradient Decline)
for $l$=2 to $\lceil N_h/2 \rceil$ do
    $h_i^{(l)} = s(W^{(l-1)} h_i^{(l-1)} + b^{(l-1)})$
end for
return $h^{(\lceil N_h/2 \rceil)}$
```

where $h_i^{(l)}$ denotes the vector of lth hidden layers for v_i, $W^{(l)}$ and $b^{(l)}$ are the parameters between lth layer and $(l+1)$th layer and s is the sigmoid function ($s(x) = \frac{1}{1+e^{-x}}$). Specially, $h_i^{(0)} = \widetilde{x}_i$ and $z_i = h_i^{(N_h+1)}$.

The cost function to evaluate the difference between $\{x_i, c_i\}$ and $\{\hat{x}_i, \hat{c}_i\}$ is defined as:

$$J(W,b) = \frac{\lambda_1}{2m} \sum_{i=1}^{m} ||x_i - \hat{x}_i||^2 + \frac{\lambda_2}{2m} \sum_{i=1}^{m} ||c_i - \hat{c}_i||^2 + \frac{\lambda_3}{2} \sum_{l} (||W^{(l)}||_F^2 + ||b^{(l)}||_2^2) \quad (2)$$

where m is the number of samples, $\lambda_1, \lambda_2, \lambda_3$ is a regularization hyperparameter and $|| \cdot ||_F$ denotes the Frobenius norm.

The first and second terms in Equation 2 indicate average error of x_i and c_i. The third term is a weight decay term for decreasing the values of the weights W and preventing overfitting (Ng 2011). The hyperparameters $\lambda_1, \lambda_2, \lambda_3$ control the relative importance of the three terms. We define $\theta = (W, b)$ as our parameters to be determined. The training of SDAE-GCL is optimized to minimize the cost function:

$$\theta^* = arg \min_{\theta} J(W, b) \quad (3)$$

The optimization methods used in this paper is *Stochastic Gradient Descent Algorithm* (Bottou 2010). For each iteration, we perform updates as following:

$$W = W - \beta \frac{\partial}{\partial W} J(W, b) \quad (4)$$

$$b = b - \beta \frac{\partial}{\partial b} J(W, b) \quad (5)$$

where β is the step size in gradient descent algorithm.

After training, the middle layer $h^{\lceil N_h/2 \rceil}$ is considered as output of our SDAE-GCL. The complete algorithm for SDAE-GCL is summarized in Algorithm 1.

Regression Model. To map visual features to the image-scale space, we further make the new constructed features $h^{(\lceil N_h/2 \rceil)}$ produced by SDAE-GCL cast into $D_i(wc_i, hs_i)$. This step can be considered as a regression problem. We will compare the experimental results of using different regression models specifically in Section 4.

3.3 Mapping Aesthetic Words to Image-scale Space

For art design, Kobayashi proposed 180 keywords in 16 aesthetic categories and defined their coordinate values in the image-scale space (Kobayashi 1995). But some of the words like "alert" and "robust" are seldom used to describe clothing. How to build a library of aesthetic words describing clothing specifically? We first observe all the comments in the last three years from the clothing section of Amazon and split them by words. Then, using WordNet (Miller 1995), we only retain those adjectives. Next, we manually remove those not often used to describe clothing, like "happy" or "sad". Finally, we establish the aesthetic words space Y for clothing, containing 527 words.

Then, we will illustrate how to map the aesthetic words y_i ($\forall y_i \in Y$) to the image-scale space D. To determine the coordinate value $D_{y_i}(wc_{y_i}, hs_{y_i})$ of an aesthetic word $y_i \in Y$, we first define the 180 keywords proposed by Kobayashi as $keyword_j$ ($j = 1, 2, \cdots, 180$) and calculate the semantic distances between y_i and each $keyword_j$ using WordNet::Similarity (Pedersen, Patwardhan, and Michelizzi 2004). Then we choose three keywords with the shortest distances d_{i_1}, d_{i_2} and d_{i_3}, marking the coordinate values of these three keywords as $D_{i_1}(wc_{i_1}, hs_{i_1})$, $D_{i_2}(wc_{i_2}, hs_{i_2})$, $D_{i_3}(wc_{i_3}, hs_{i_3})$. Taking the reciprocals of distances rec_{i_1}, rec_{i_2}, rec_{i_3} as weights (e.g. $rec_{i_1} = \frac{1}{d_{i_1}}$), the weighted arithmetic mean[1] of D_{i_1}, D_{i_2} and D_{i_3} can be regarded as the coordinate value $D_{y_i}(wc_{y_i}, hs_{y_i})$ of y_i. The formula is shown as follows:

$$wc_{y_i} = \frac{\sum_{k=1}^{3} wc_{i_k} \cdot rec_{i_k}}{\sum_{k=1}^{3} rec_{i_k}}, hs_{y_i} = \frac{\sum_{k=1}^{3} hs_{i_k} \cdot rec_{i_k}}{\sum_{k=1}^{3} rec_{i_k}} \quad (6)$$

In this way, for each $y_i \in Y$, we can calculate its coordinate value D_{y_i} in the image-scale space as (wc_{y_i}, hs_{y_i}). To label an input clothing image v with an aesthetic word, we first use the proposed SDAE-GCL to predict its coordinate value $D_v(wc_v, hs_v)$ in D. Then, we find a word $y_v \in Y$ whose corresponding coordinate value D_{y_v} has the shortest Euclidean distance to the D_v. Thus, y_v can be regarded as the aesthetic word of image v.

4 Experiments

In this section, we first conduct several objective experiments to validate the SDAE-GCL by evaluating the mapping effects between visual features of clothing images and coordinate values in the image-scale space. Then we show the effectiveness of the proposed VF-ISS-AWS framework through some interesting demonstrations.

[1] https://en.wikipedia.org/wiki/Weighted_arithmetic_mean

Table 1: Comparison among different autoencoders

Autoencoder	Regression	MSE	MAE
None	SVM	0.3578	0.2808
AE[8]		0.3462	0.2502
DAE[9]		0.3398	0.2485
SDAE		0.3395	0.2481
DAE-GCL		0.3365	0.2467
SDAE-GCL		**0.3256**	**0.2366**

Table 2: Comparison among different regression models

Autoencoder	Regression	MSE	MAE
SDAE-GCL	KNN	0.3807	0.2734
	BLR	0.3271	0.2412
	DNN	0.3270	0.2439
	SVM	**0.3256**	**0.2366**

4.1 Dataset

We employ upper-body menswear as our experimental data for the following two considerations. First, compared to various kinds of women's dress, menswear has more clear categories and simple features. Second, by focusing on upper-body but not full-body menswear, we can avoid the deviation produced by the matches of tops and bottoms.

D1: Labeled Dataset. This dataset contains 5500 images downloaded from *Amazon*[2]. There are 11 common categories of upper-body menswear in *Amazon*: suit, sweater, padding, shirt, tee, windbreak, mountainwear, fur, hoodies, jacket and vest. We randomly select 500 images for each category and manually label the coordinate values in the image-scale space. Both warm-cool and soft-hard coordinate values range in [-1, +1] and the granularity of each dimension is 0.1. We invite 5 annotators (2 males and 3 females) who are well trained with the image-scale space to use the annotation tool. For each image, its final coordinate value is averaged over the five coordinate values annotated. If the Euclidean distance of two values given by different annotators is larger than 0.3, they will discuss to get an unanimous result.

D2: Unlabeled Dataset. Since one of the advantages of the proposed SDAE-GCL is that both labeled and unlabeled data can be incorporated to improve the model accuracy, we establish an unlabeled upper-body menswear dataset. In order to make our model more applicable in different data sources, we select another online shopping website *JD*[3] as the data source and fetch 130,316 images of the 11 categories same as those in D1. The category distribution is: suit: 10.0%, sweater: 9.4%, padding: 15.7%, shirt: 5.7%, tee: 6.4%, windbreak: 9.1%, mountainwear: 6.1%, fur: 9.4%, hoodies: 5.9%, jacket: 8.4%, vest: 7.7%.

D3: Demonstration Dataset. This dataset consists of two parts: 1) 76360 menswear images containing 277 brands and covers the Spring menswear and Fall menswear in the last 10 years from *Style* [4]; 2) 800 menswear images appearing in 2016 New York fashion week from *The Fashion Is To* [5].

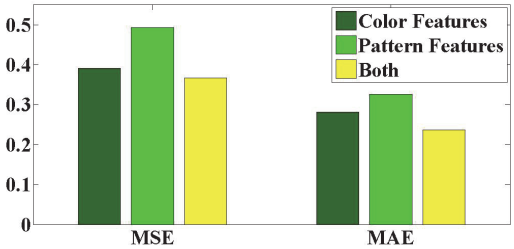


Figure 3: Feature contribution analyses

4.2 Metrics

In the evaluation of the mapping effects between visual features of clothing images and coordinate values in the image-scale space, we calculate the error between predicted coordinate values and labeled coordinate values. The error is formulated by mean squared error (MSE[6]) and mean absolute error (MAE[7]) (Jia et al. 2011). All the experiments are performed on five-folder cross-validation.

4.3 Results and Analyses

Performance of different autoencoders. Using the same regression model Support Vector Machine (SVM) (Rebentrost, Mohseni, and Lloyd 2014), we compared the proposed SDAE-GCL with other different autoencoders. The results are shown in Table 1. The following results are observed: 1) Autoencoders perform better than simple SVM regression, while DAE has even better performance because it works well on the noisy data which is significant as the features extracted cannot avoid errors; 2) The proposed SDAE-GCL outperforms traditional DAE and both the guiding strategy and stacking strategy contribute to the performance. Moreover, we compare the methods taking clothing categories as correlative labels and takes them as features. The results show that the performance of the former method (MSE: 0.3256, MAE: 0.2366) is better than the latter (MSE: 0.3458, MAE: 0.2507), also supporting the effectiveness of the guiding strategy which takes categories as correlative labels.

Performance of different regression models. Using the proposed SDAE-GCL, we make several comparisons among different regression models including Support Vector Machines (SVM), Bayesian linear regression (BLR) (Wang, Sun, and Lu 2015), K-Nearest Neighbors (KNN) (Li et al. 2012) and deep neural network (DNN) (Bengio 2009). As shown in Table 2, all of BLR, DNN and SVM have satisfactory results, which indicates that the SDAE-GCL can well capture the intrinsic and holistic relationship between the visual features and aesthetic effects. In our experiments

[2] http://www.amazon.com

[3] http://www.jd.com

[4] http://www.style.com

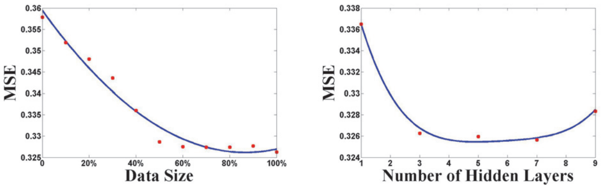
[5] http://www.thefashionisto.com/

[6] https://en.wikipedia.org/wiki/MSE

[7] https://en.wikipedia.org/wiki/Mean_absolute_error

[8] https://en.wikipedia.org/wiki/Autoencoder

[9] https://en.wikipedia.org/wiki/Deeplearning4j



(a) (b)

Figure 4: (a) The influence of data size of unlabeled data. (b) The influence of the number of hidden layers.

and demonstrations, we take the best performing SVM as the regression model.

Feature contribution analyses. On the condition of SDAE-GCL and SVM, we discuss the contributions of color features and pattern features. As shown in Figure 3, all the features contribute to the mapping effects. Moreover, color features (MSE: 0.4212, MAE: 0.3007) contribute significantly more than pattern features (MSE: 0.5111, MAE: 0.3734), which shows the color features can affect people's judgement in a greater degree.

Parameter sensitivity analyses. For the proposed SDAE-GCL, we further test the parameter sensitivity about two key parameters with different values. 1) Training data size. Since the size of labeled part is constant, we change the size of unlabeled data to evaluate the performance. From Figure 4(a), we can find that as the scale of unlabeled data increases the performance gets better. With the size over 1.04×10^5, the performance reaches convergence. Therefore we use 1.04×10^5 unlabeled data as our training data. 2) Hidden layer number. Theoretically, the description ability of SDAE-GCL can be improved by more layers. The performance do increase with layer number less than five, but get worse after the number become larger because of overfitting caused by limited size of training data. Therefore we take 5 layers in our experiments. Moreover, on this condition the experiment lasts for about two hours on an environment with dual-core 2.10GHZ CPU, 64GB memory.

4.4 Demonstration

As it is hard to validate the accuracy of mapping aesthetic words to the image-scale space through objective experiments, we show the results of labeling the clothing images with aesthetic words in Figure 1 to demonstrate the effectiveness of the proposed framework and methods. Furthermore, we would like to show several interesting case studies about fashion trend of menswear as supplementary:

1. In the last ten years, the general style of fashion menswear of 277 brands is "gentlemanly", "formal" and "chic", shown in Figure 5(a). The majority of menswear (66.2%) during this decade is distributed in the fourth quadrant with cooler and harder aesthetic effects. Figure 5(b) further presents that the fashion of menswear keeps several classic styles and has minor alterations.
2. Despite the overall steadiness, some brands present significant changes in the past decades. Taking *Jean Paul Gaultier* as an example, its aesthetic effects change from "decorative" to "simple" as shown in Figure 5(c).
3. Different brands have their own styles with different aesthetic effects. For example, as figure 5(d) compares three different brands in 2016 New York fashion week: the style of *Detroit* is usually "simple", the style of *Jeffrey Rudes* is more likely to be "dignified" and "traditional" and the style of *Eponymous* is "gentilemanly" and "diginified".

5 Conclusion

In this paper, we make an intentional step on understanding the aesthetic effects of clothing images automatically. By introducing the image-scale space as the intermediate level, the proposed three-level framework follows the aesthetic and psychological principles. In future work, we will carry on our work in two aspects: 1) Discover the relationship between clothing collocation and aesthetic effects; 2) Apply the framework to analyse the various women's dress.

6 Acknowledgments

This work is supported by the National Basic Research Program (973 Program) of China (2012CB316401), National Natural, and Science Foundation of China (61370023, 61303075). This work is partially supported by the major project of the National Social Science Foundation of China(13& ZD189), the National High Technology Research and Development Program("863" Program) of China (2012AA011602). We also thank Microsoft Research Asia-Tsinghua Univertity Joint Laboratory for its support (FY14-RES-SPONSOR-111).

References

Bengio, Y. 2009. Learning deep architectures for ai. *Foundations and trends® in Machine Learning* 2(1):1–127.

Bottou, L. 2010. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*. Springer. 177–186.

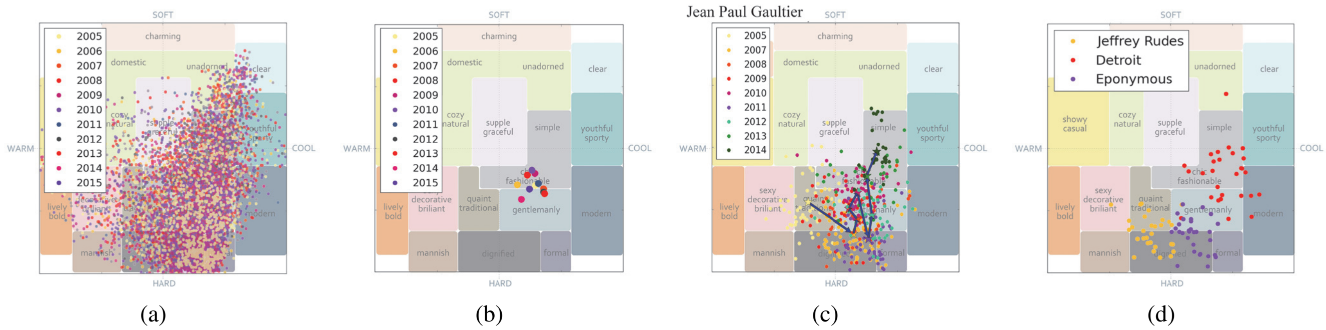
Hasan, B., and Hogg, D. 2010. Segmentation using deformable spatial priors with application to clothing. In *BMVC*, 1–11.

Jia, J.; Zhang, S.; Meng, F.; Wang, Y.; and Cai, L. 2011. Emotional audio-visual speech synthesis based on pad. *Audio, Speech, and Language Processing, IEEE Transactions on* 19(3):570–582.

Kang, J.-Y. M.; Johnson, K. K.; and Kim, J. 2013. Clothing functions and use of clothing to alter mood. *International Journal of Fashion Design, Technology and Education* 6(1):43–52.

Kobayashi, S. 1995. Art of color combinations. *Kosdansha International*.

Kouge, Y.; Murakami, T.; Kurosawa, Y.; Mera, K.; and Takezawa, T. 2015. Extraction of the combination rules of colors and derived fashion images using fashion styling data. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1.



(a) (b) (c) (d)

Figure 5: (a) The overall distribution of 76,360 menswear in the last ten years in the image-scale space. (b) The barycentric coordinate of each year is calculated and shown in the image-scale space. (c) The fashion trend of Jean Paul Gaultier in the last ten years. (d) Comparison among three brands in 2016 New York fashion week.

Li, C.; Liu, Q.; Liu, J.; and Lu, H. 2012. Learning distance metric regression for facial age estimation. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, 2327–2330. IEEE.

Liu, S.; Feng, J.; Song, Z.; Zhang, T.; Lu, H.; Xu, C.; and Yan, S. 2012. Hi, magic closet, tell me what to wear! In *Proceedings of the 20th ACM international conference on Multimedia*, 619–628. ACM.

Liu, S.; Liang, X.; Liu, L.; Shen, X.; Yang, J.; Xu, C.; Lin, L.; Cao, X.; and Yan, S. 2015. Matching-cnn meets knn: Quasi-parametric human parsing. *arXiv preprint arXiv:1504.01220*.

Miller, G. A. 1995. Wordnet: a lexical database for english. *Communications of the ACM* 38(11):39–41.

Ng, A. 2011. Sparse autoencoder. *CS294A Lecture notes* 72.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration papers at hlt-naacl 2004*, 38–41. Association for Computational Linguistics.

Rebentrost, P.; Mohseni, M.; and Lloyd, S. 2014. Quantum support vector machine for big data classification. *Physical review letters* 113(13):130503.

Shen, E.; Lieberman, H.; and Lam, F. 2007. What am i gonna wear?: scenario-oriented recommendation. In *Proceedings of the 12th international conference on Intelligent user interfaces*, 365–368. ACM.

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2014. Going deeper with convolutions. *arXiv preprint arXiv:1409.4842*.

Vincent, P.; Larochelle, H.; Bengio, Y.; and Manzagol, P.-A. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, 1096–1103. ACM.

Wang, N., and Ai, H. 2011. Who blocks who: Simultaneous clothing segmentation for grouping images. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 1535–1542. IEEE.

Wang, X.; Jia, J.; Yin, J.; and Cai, L. 2013. Interpretable aesthetic features for affective image classification. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, 3230–3234. IEEE.

Wang, X.; Jia, J.; Tang, J.; Wu, B.; Cai, L.; and Xie, L. 2014. Modeling emotion influence in image social networks.

Wang, M.; Sun, X.; and Lu, T. 2015. Bayesian structured variable selection in linear regression models. *Computational Statistics* 30(1):205–229.

Yamaguchi, K.; Kiapour, M. H.; Ortiz, L. E.; and Berg, T. L. 2015. Retrieving similar styles to parse clothing. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 37(5):1028–1040.

Yamaguchi, K.; Kiapour, M. H.; and Berg, T. 2013. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, 3519–3526. IEEE.

Yang, Y., and Ramanan, D. 2011. Articulated pose estimation with flexible mixtures-of-parts. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 1385–1392. IEEE.

Yang, W.; Luo, P.; and Lin, L. 2014. Clothing co-parsing by joint image segmentation and labeling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, 3182–3189. IEEE.

Yu-Chu, L.; Kawakita, Y.; Suzuki, E.; and Ichikawa, H. 2012. Personalized clothing-recommendation system based on a modified bayesian network. In *Applications and the Internet (SAINT), 2012 IEEE/IPSJ 12th International Symposium on*, 414–417. IEEE.