Towards Better Understanding the Clothing Fashion Styles: A Multimodal Deep Learning Approach

Yihui Ma,[1] Jia Jia,[1*] Suping Zhou,[3] Jingtian Fu,[12] Yejun Liu,[12] Zijian Tong[4]
[1]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology (TNList)
[2]Academy of Arts & Design, Tsinghua University, Beijing, China
[3]Beijing University of Posts and Telecommunications, Beijing, China
[4]Sogou Corporation, Beijing, China
jjia@mail.tsinghua.edu.cn

Abstract

In this paper, we aim to better understand the clothing fashion styles. There remain two challenges for us: 1) how to quantitatively describe the fashion styles of various clothing, 2) how to model the subtle relationship between visual features and fashion styles, especially considering the clothing collocations. Using the words that people usually use to describe clothing fashion styles on shopping websites, we build a *Fashion Semantic Space (FSS)* based on Kobayashi's aesthetics theory to describe clothing fashion styles quantitatively and universally. Then we propose a novel fashion-oriented multimodal deep learning based model, *Bimodal Correlative Deep Autoencoder (BCDA)*, to capture the internal correlation in clothing collocations. Employing the benchmark dataset we build with 32133 full-body fashion show images, we use BCDA to map the visual features to the FSS. The experiment results indicate that our model outperforms (+13% in terms of MSE) several alternative baselines, confirming that our model can better understand the clothing fashion styles. To further demonstrate the advantages of our model, we conduct some interesting case studies, including fashion trends analyses of brands, clothing collocation recommendation, etc.

1 Introduction

What are the most popular clothing fashion styles of this season? Reported by Vogue[1], romantic, elegant and classic are the top trends during the Fall 2016 Couture collection. Exploring the fashion runway images, these styles rely heavily on some specific visual details, such as nipped waist, lapel collar, matched with high-waistlines dress or pencil trousers. Since clothing fashion styles benefit a lot from visual details, can we bridge the gap between them automatically? Many efforts have been made towards this goal. For example, (Wang, Zhao, and Yin 2014) presents a method to parse refined texture attribute of clothing, while (Yang, Luo, and Lin 2015) builds an integrated application system to parse a set of clothing images jointly. Besides, people try to analyse visual features by adding occasion and scenario elements. (Liu et al. 2012) considers occasions in dressing and focus on scenario-oriented clothing recommendation. Although a latest work (Jia et al. 2016) proposes to appreciate the aesthetic effects of upper-body menswear, it still lacks universality and ignores that the collocation of top and bottom has a significant impact on fashion styles. Thus, there still remain two challenges for us: 1) how to quantitatively describe the fashion styles of various clothing, 2) how to model the subtle relationship between visual features and fashion styles, especially considering the clothing collocations.

In this paper, we aim to better understand the clothing fashion styles and propose our solutions from two aspects. First, we build a Fashion Semantic Space (FSS) based on the Image-Scale Space in aesthetic area proposed by (Kobayashi 1995). By computing the semantic distances using WordNet::Similarity (Pedersen, Patwardhan, and Michelizzi 2004), we coordinate the most often used 527 aesthetic words on clothing section of Amazon to FSS. Second, we propose a fashion-oriented multimodal deep learning based model, Bimodal Correlative Deep Autoencoder (BCDA), to capture the correlation between visual features and fashion styles by utilizing the intrinsic matching rules of tops and bottoms. Specifically, we regard the tops and bottoms as two modals of clothing collocation, and leverage the shared representation of multimodal deep learning to learn the relationship between the modalities. In addition, we improve the process of feature learning by taking the clothing categories (e.g. suit, coat, leggings, etc.) as correlative labels. Connecting BCDA to a regression model, we finally map the visual features to the FSS. Employing 32133 full-body fashion images downloaded from fashion show websites as our experimental data, we conduct several experiments to evaluate the mapping effects between visual features and coordinate values in FSS. The results indicate that the proposed BCDA model outperforms (+13% in terms of MSE) several alternative baselines. Besides, we also show some interesting cases

*Corresponding author: J. Jia (jjia@mail.tsinghua.edu.cn)

[1]http://www.vogue.com/fashion-shows

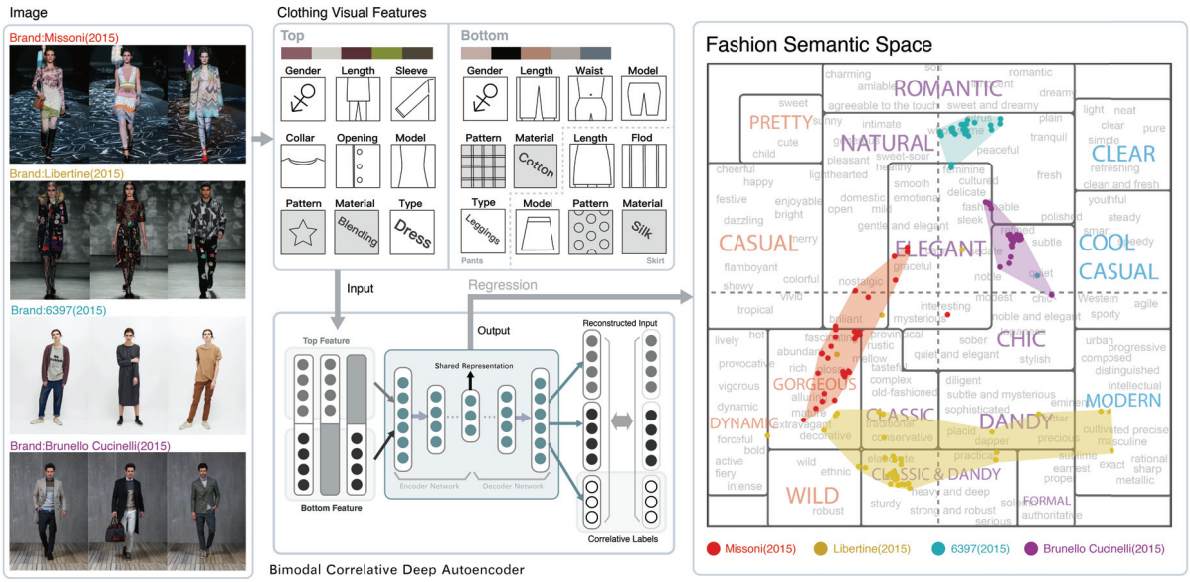


Figure 1: The workflow of our framework.

to demonstrate the advantages of our model. The illustration of our work is shown in Figure 1.

We summarize our contributions as follows:

- We construct a benchmark clothing fashion dataset containing 32133 full-body fashion show images from Vogue in the last 10 years. The collected dataset is labeled with complete visual features (e.g. collar shape, pants length, color theme, etc.) and fashion styles (e.g. casual, chic, elegant, etc.). We are willing to make our dataset open to facilitate other people's research on clothing fashion[2].
- We build a universal Fashion Semantic Space (FSS) to describe clothing fashion styles quantitatively. It is a two-dimensional image-scale space containing hundreds of words that people often use to describe clothing on shopping websites. Based on the FSS, we can not only do quantitative evaluation on fashion collocation, but also analyze the dynamic change of fashion trends intuitively.
- We propose a fashion-oriented multimodal deep learning based model, Bimodal Correlative Deep Autoencoder (BCDA), connected with regression to implement the task of mapping visual features to FSS. Specifically, leveraging the shared representation learned by the multimodal strategy, BCDA can make full use of the internal correlation between tops and bottoms, and resolve the issue of clothing collocation.

The rest of paper is organized as follows. Section 2 lists related works. Section 3 formulates the problem. Section 4 presents the methodologies. Section 5 introduces the experiment dataset, results and case studies. Section 6 is the conclusion.

[2]https://pan.baidu.com/s/1boPm2OB

2 Related Works

Clothing parsing. Clothing parsing is a popular research topic in the field of computer vision. (Wang, Zhao, and Yin 2014) parses refined clothing texture by exploiting the discriminative meanings of sparse codes. (Yamaguchi et al. 2015) tackles the clothing parsing problem using a retrieval-based approach. Benefiting from deep learning, the performance of clothing parsing is promoted in recent years. For example, (Wang, Li, and Liu 2016) use Fast R-CNN for a more effectively detecting of human body and clothing items.

Clothing recommendation. As people pay more attention to clothing fashion, clothing recommendation becomes a hot topic. (Liu et al. 2012) considers occasions in dressing and focus on scenario-oriented clothing recommendation. (Jagadeesh et al. 2014) proposes a data-driven model which has large online clothing images to build a recommendation system. (Hu, Yi, and Davis 2015) proposes a functional tensor factorization to build a model between user and clothing. However, since people select clothing by words "romantic" or "elegant" rather than visual details, how to bridge the gap between the visual features and fashion styles is a issue to be resolved.

Fashion style modeling. A latest work (Jia et al. 2016) proposes to appreciate the aesthetic effects of upper-body menswear. However, clothing variety and fashion collocation are significant elements of clothing fashion that we cannot ignore. How to universally describe fashion styles on clothing collocation is still a open problem.

3 Problem Formulation

Given a set of fashion images V, for each image $v_i \in V$, we use an N_{x^t} dimensional vector $x_i^t = \langle x_{i1}^t, x_{i2}^t, \ldots, x_{iN_{x^t}}^t \rangle$ $(\forall x_{ij}^t \in [0, 1])$ to indicate v_i's top

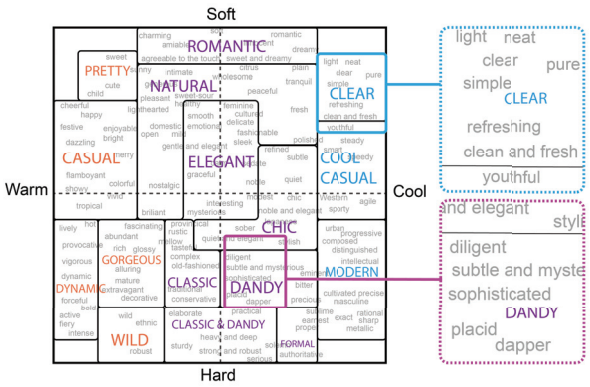


Figure 2: The fashion semantic space.

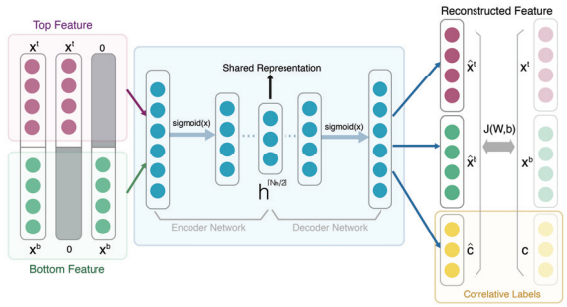


Figure 3: The structure of BCDA.

(upper-body) visual features, an N_{x^b} dimensional vector $x_i^b = \langle x_{i1}^b, x_{i2}^b, \ldots, x_{iN_{x^b}}^b \rangle \quad (\forall x_{ij}^b \in [0,1])$ to indicate v_i's bottom (lower-body) visual features, an N_{c^t} dimensional vector $c_i^t = \langle c_{i1}^t, c_{i2}^t, \ldots, c_{iN_{c^t}}^t \rangle \quad (\forall c_{ij}^t \in [0,1])$ to indicate v_i's top clothing categories, and an N_{c^b} dimensional vector $c_i^b = \langle c_{i1}^b, c_{i2}^b, \ldots, c_{iN_{c^b}}^b \rangle \quad (\forall c_{ij}^b \in [0,1])$ to indicate v_i's bottom clothing categories. In addition, X^t is defined as a $|V| * N_{x^t}$ feature matrix with each element x_{ij}^t denoting the jth top visual features of v_i. The definitions of X^b, C^t and C^b are similar to X^t.

Definition. **The Fashion Semantic Space** Y is a two-dimensional space (warm-cool and hard-soft) based on the image-scale space, denoted as $Y(wc, hs) \quad (\forall wc, hs \in [-1, +1])$. 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. Focusing on fashion styles, a series of fashion descriptors (e.g., youthful, casual, chic) that people usually use to describe the styles of clothing are coordinated in Y. Thus, we build a fashion semantic space, which can be further clustered into n sections to meet the demand of actual applications.

Problem. **Modeling the fashion styles of clothing.** We aim to find a corresponding style descriptor in Y for a clothing image automatically. Therefore, a prediction model $M : (V, X^t, X^b, C^t, C^b) \Rightarrow Y$ needs to be learned. For an input image $v_i \in V$, we calculate $Y_i(wc_i, hs_i)$ by model M, thus the fashion style of v_i is determined.

4 Methodologies

In order to understand clothing fashion better, we formulate it to two tasks: 1) building a Fashion Semantic Space (FSS) to describe clothing fashion styles universally and quantitatively, 2) proposing a fashion-oriented multimodal deep learning based model, Bimodal Correlative Deep Autoencoder (BCDA), connected with regression models to build the mapping from visual features to FSS.

4.1 Building the Fashion Semantic Space

For art design, Kobayashi proposed 180 keywords in 16 aesthetic categories and defined their coordinate values in the image-scale space (Kobayashi 1995). In order to describe the fashion styles of various clothing, 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), only adjectives are retained. Next, we manually remove those not often used to describe clothing, like "happy" or "sad", getting 527 aesthetic words representing fashion styles. To determine the coordinates of these words, we calculate the semantic distances between keywords and aesthetic words using WordNet::Similarity (Pedersen, Patwardhan, and Michelizzi 2004). For an word to be coordinated, we choose three keywords with the shortest distances, the weighted arithmetic mean of which can be regarded as the coordinate value. In this way, we build the fashion semantic space, illustrated in Figure 2.

4.2 Fashion-Oriented Multimodal Deep Learning Based BCDA

Intuition. Although the traditional Deep Autoencoder (DA) is an approach to feature learning, it cannot make use of the internal correlation between top and bottom clothing. Thus we take two strategies to resolve this issue: 1) We adopt a multimodal deep learning model, Bimodal Deep Autoencoder (BDA) (Ngiam et al. 2011), and adapt it for fashion-oriented feature learning. Regarding tops and bottoms as two modals of clothing collocations, the shared representation learned from BDA can be regarded as the new feature representation, which serves as the input of the regression model. 2) It is observed that for different categories (e.g. suit, coat, leggings, etc.), even clothing images with similar visual features can present different fashion effects, showing that fashion styles are influenced by clothing categories (Jia et al. 2016). So we take clothing categories as correlative labels, and promote the BDA to a novel structure named Bimodal Correlative Deep Autoencoder (BCDA), shown in Figure 3.

The structure of BCDA. Given an image $v_i \in V$, the initial input vector $x_i = \{x_i^t, x_i^b\}$ represents the visual feature vector and $c_i = \{c_i^t, c_i^b\}$ represents the clothing category vector. We use a multilayer neural network to rebuild 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 BCDA contain encoder network and decoder network as illustrated in the blue box of Figure 3. The relationship between two adjacent layers depends on model

Algorithm 1 Bimodal Correlative Deep Autoencoder

Input: $X = \{X^t, X^b\}$, a preprocessed feature matrix.
$C = \{C^t, C^b\}$, a clothing category matrix.

Output: Middle hidden layer feature $h^{\lceil N_h/2 \rceil}$, the shared representation of the input visual features.

Initialize model parameters $\theta^{(l)}, \alpha, \lambda_1, \lambda_2, \lambda_3$
repeat
  $W^{(l)} = W^{(l)} - \alpha \frac{\delta}{\delta W^{(l)}} J(W, b)$
  $b^{(l)} = b^{(l)} - \alpha \frac{\delta}{\delta b^{(l)}} J(W, b)$
until convergence (Gradient Descent)
for $l = 2$ to $\lceil N_h/2 \rceil$ **do**
  $h^{(l)} = sigmoid(W^{(l-1)} h^{(l-1)} + b^{(l-1)})$
end for
return $h^{\lceil N_h/2 \rceil}$

parameters. After training, we determine the parameters and learn the intermediate representation as the output of this step.

Compared to classic autoencoder, we introduce correlative labels c_i into the original symmetrical structure as shown in yellow box of Figure 3. We use the neural network to regain the correlative labels c_i, reconstructing x_i in parallel. In this way, the clothing categories can be leveraged to help to discover the correlation between various visual features and make facilitate the training process.

In order to capture the internal correlation between top and bottom in clothing collocation, we influence the training process of BCDA by preprocessing the dataset. Concretely, we treat top and bottom as two different modals of fashion collocation. Tripling the original dataset, we get $X_1 = \{X^t, X^b\}$, $X_2 = \{X^t, X^b\}$, and $X_3 = \{X^t, X^b\}$. Then we set the bottom features of X_2 and the top features of X_3 to zero. Now we get a new dataset $X = \{X_1, X_2', X_3'\}$, where $X_2' = \{X^t, 0\}$, and $X_3' = \{0, X^b\}$. When training the autoencoder, we still expect it to recover all the three small datasets into full features (i.e. $\hat{X} = \{X_1, X_1, X_1\}$). In this way, the BCDA learns the hidden rules of fashion collocation automatically.

Formally, supposing the BCDA has N_h layers, the recursion formula between two adjacent layers is:

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

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 $sigmoid$ is the sigmoid function ($sigmoid(x) = \frac{1}{1+e^{-x}}$). Specially, $h_i^{(0)} = x_i$ and $z_i = h_i^{(N_h+1)}$.

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

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

TOP	
Gender	Female / Male / General
Length	Extra-Short / Short / Mid / Long / Extra-Long
Sleeve Length	Suspender / Sleeveless / Short / Mid / Long
Collar Shape	Round / Lapel / High / Stand / V / Bateau / Fur / Hoodie
Opening	Single-breasted / Double-breasted / Half-breasted / Zipper/ Half-zipper / Pullover / Open
Model	Tight / Straight / Loose / Waisted / Cloak
Pattern	Pure / Grid / Dot / Floral / Vertical stripe / Cross stripe / Number&Letter / Focus / Repeat
Material	Cotton / Chemical fiber / Blending / Woolen / Silk / Denim / Leather / Flax / Knit
Type	Dress / T-shirt / Sweater / Shirt / Suit / Jacket / Vest / Hoodie / Coat / Sportswear / Down-Jacket / Fur / Leather / Cheongsam / Mountaineering jacket

BOTTOM	PANTS
Gender	Female / Male / General
Length	Short / Mid / Long / Cropped
Waist	Low / Normal / High
Model	Tight / Straight / Loose
Pattern	Pure / Grid / Dot / Floral / Vertical stripe / Cross stripe / Number&Letter / Focus / Repeat
Material	Cotton / Chemical fiber / Blending / Woolen / Silk / Denim / Leather / Flax / Knit
Type	Leggings / Sport pants / Hot pants / Harem pants / Bell-bottoms / Suspender / Jeans / Suit pants / Casual pants

BOTTOM	SKIRT
Length	Short / Mid / Long
Fold	With / Without
Model	A-shape / Packet Hip
Pattern	Pure / Grid / Dot / Floral / Vertical stripe / Cross stripe / Number&Letter / Focus / Repeat
Material	Cotton / Chemical fiber / Blending / Woolen / Silk / Denim / Leather / Flax / Knit

Figure 4: The annotation details.

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

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

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

The optimization method we adopt is Stochastic Gradient Descent Algorithm (Bottou 2010). For each iteration, we perform updates as following:

$$W = W - \alpha \frac{\delta}{\delta W} J(W, b) \tag{4}$$

$$b = b - \alpha \frac{\delta}{\delta b} J(W, b) \tag{5}$$

where α is the step size in gradient descent algorithm.

After training, the middle layer $h^{\lceil N_h/2 \rceil}$ is a shared representation of the input features, considered as the output of BCDA. The complete algorithm for BCDA is summarized in Algorithm 1.

Regression Model. To build a correlation between visual features and fashion semantic space, we further make the shared representation $h^{\lceil N_h/2 \rceil}$ produced by BCDA cast into $Y(wc, hs)$. Specifically, we choose one of the 527 style words in FSS which has the shortest Euclidean distance with $Y(wc, hs)$ as the fashion style label of the input image. This

Table 1: (a)Comparison among different autoencoders (b)Comparison among different regression models

(a)

Autoencoder	SVM
None	0.2083
DA	0.1927
BDA	0.1881
CDA	0.1860
BCDA	0.1841

(b)

Regression	BCDA
D-Tree	0.3378
KNN	0.3324
DNN	0.2591
LR	0.1854
SVM	0.1841

step can be considered as a regression problem. We will compare the experimental results of using different regression models specifically in Section 5.

5 Experiments

In this section, we first conduct several objective experiments to validate the BCDA by evaluating the mapping effects between visual features of clothing images and coordinate values in the FSS. Then we show the effectiveness of our model through some interesting demonstrations.

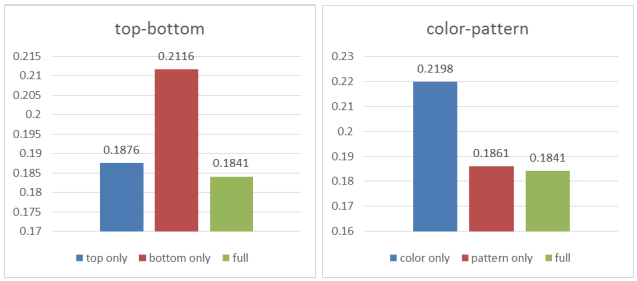
5.1 Dataset

We build a large full-annotated benchmark dataset, which employs 32133 full-body fashion show images in the last 10 years downloaded from Vogue. It covers both men and women clothing, and contains 550 fashion brands.

Annotation details of our dataset. 1) *Clothing visual features.* We define the visual features of clothing from two aspects: color features and pattern features. For the color features, we extract the five color theme of clothing images adopting the algorithm proposed by (Wang et al. 2013). For the pattern features, we invite people to annotate them, and the annotation details are listed in Figure 4, in which the "type" feature corresponds to the clothing category in our algorithm. In practice, using these annotated features as training data, we can train CNN models (Szegedy et al. 2015) to detect the pattern features. 2) *Fashion style feature.* For the annotation of fashion styles, we provide the FSS space for the annotators, who are well trained with the annotation method. For each image, they choose a style section in FSS at first, and then determine a specific coordinate for the image according to the distribution of fashion style words. For coordinates in the FSS, both warm-cool and soft-hard coordinate values range in [-1, +1] and the granularity of each dimension is 0.001. For all the annotation features above, 20 annotators (10 males and 10 females) are invited to finish the annotation task. Each attribute of each image is annotated by 3 different annotators, and the final results are voted or averaged by the original 3 results.

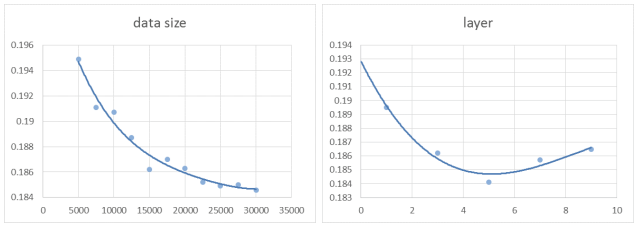
5.2 Metric

To evaluate the mapping effects between visual features and coordinate values in FSS, we calculate the error between predicted coordinate values and annotated coordinate values. The error is measured by mean squared error (MSE). All the experiments are performed on 5-folder cross-validation.



(a) (b)

Figure 5: Feature contribution analyses.



(a) (b)

Figure 6: Parameter sensitivity analyses. (a)Training data size. (b)Hidden layer number.

5.3 Results and Analyses

Performance of different autoencoders. Using the same regression model Support Vector Machine (SVM) (Rebentrost, Mohseni, and Lloyd 2014), we compare the proposed BCDA with other different autoencoder settings (None: no feature learning, DA: Deep Autoencoder, BDA: Bimodal Deep Autoencoder, CDA: Correlative Deep Autoencoder). The results are shown in Table 1(a). We can find the bimodal strategy (with/without "B") of learning the shared representation takes effect, which indicates that top and bottom do have a correlation at feature level. Besides, the correlative-label strategy (with/without "C") contributes to the result too, proving that clothing categories have impact on fashion styles indeed. Moreover, we compare the methods taking clothing categories as correlative labels and takes them as features. The performance of the former method (MSE: 0.1841) is better than the latter (MSE: 0.1876), also supporting the effectiveness of correlative labels.

Performance of different regression models. Using the proposed BCDA, we also make several comparisons among different regression models including Decision Tree (D-Tree) (Trendowicz and Jeffery 2014), K-Nearest Neighbors (KNN) (Li et al. 2012), Deep Neural Network (DNN) (Bengio 2009), Linear Regression (LR) (Ho and Lin 2012) and Support Vector Machine (SVM). As shown in Table 1(b), D-Tree and KNN have the worst performance. It can be inferred that the regression models leveraging all the samples simultaneously fit our BCDA better than those relying on

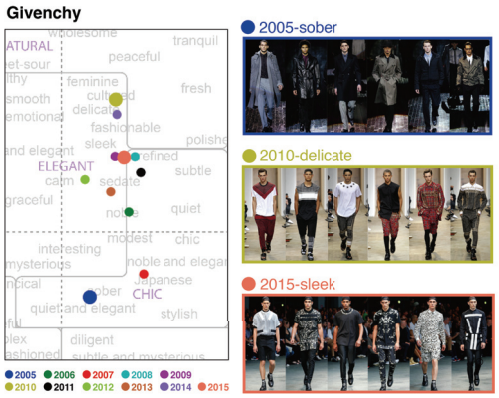


Figure 7: Fashion style trend of Givenchy.

just a few samples. In the following experiments and demonstrations, we take the best performing SVM as the regression model.

Feature contribution analyses. First, we discuss the contributions of top and bottom features separately. As shown in Figure 5(a), top features contribute more than bottom features, which is in accordance with our ordinary feelings that tops count more in clothing collocation. Then we compare the contributions of color features and pattern features in Figure 5(b). It shows that pattern features perform better than color features, probably because patterns' combination has more influence on fashion collocation.

Parameter sensitivity analyses. We further test the parameter sensitivity about two key parameters with different values. 1) Training data size. From Figure 6(a), we can find that as the scale of training data increases the performance gets better. With the size over 25000, the performance almost reaches convergence. Therefore, the size of our whole dataset (32133) is a proper number. 2) Hidden layer number. Theoretically, the description ability of BCDA can be improved by more layers. According to Figure 6(b), the performance do increase with layer number less than 5, but get worse after the number become larger because of overfitting. Therefore, we take 5 layers in our experiments. On this condition, the experiment lasts for about 20 minutes in a quad-core 2.80GHz CPU, 16GB memory environment.

5.4 Demonstration

With the proposed FSS and BCDA, we are capable of understanding clothing fashion styles better. Employing our fashion dataset, we conduct some interesting demonstrations to further show the advantages and universality of our method.

Fashion distribution comparison among different brands. At fashion shows, different brands tend to present different styles of clothing. Thus we compare the fashion style distribution of different brands published in 2015, shown in the rightmost part of Figure 1. We observe that Missoni's main style is gorgeous, Libertine tends to present a classic style, 6397 falls in the natural section intensively, and Brunello Cucinelli focuses on the chic style. The sample images in the leftmost part verify these observations.

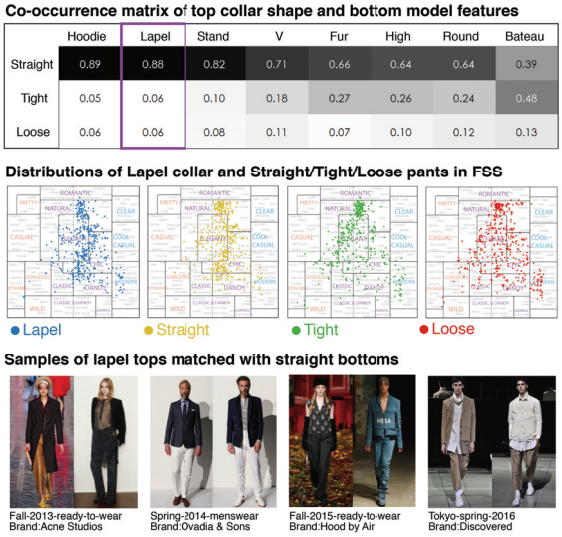


Figure 8: Feature collocation rules exploration.

Different years' fashion comparison of the same brand. Fashion trends of famous brands has always been a hot issue for fashion people. Focusing on all the fashion images of one brand (Givenchy) in the last eleven years, we show the center point of each year and find that the style of this brand has changed as shown in Figure 7. In 2005, the brand has a sober style, but moves to a softer delicate section in 2010. In the most recent year 2015, it tends to present a sleek style.

Mining feature collocation rules. Feature collocation is definitely an important element about fashion. We observe an interesting case that what kind of top collar and bottom model collocation follows the fashion. In Figure 8, the co-occurrence matrix of top collar and bottom model is calculated. The matrix elements stand for the probabilities of straight/tight/loose pants collocated with each kind of collar, thus every column sums to 1.00. The following two facts are observed: 1) From the first row of the matrix, we can find that straight pants match almost every collar in fashion collocations. Select the lapel column as example, we compare the distribution of lapel collar and straight/tight/loose pants in FSS. Among the three kinds of pants, straight presents a most similar shape with lapel, which is in accordance with the probability. 2) Although straight matches well with most collar shapes, fur/high/round/bateau matched with tight are also good choices. Moreover, bateau-tight has a even higher probability than bateau-straight. Thus bateau-tight is also a classic collocation in fashion shows.

In addition, we apply our model to build a practical application named Magic Mirror (Liu et al. 2016), which is capable of analysing people's clothing fashion styles automatically.

6 Conclusion

In this paper, we make an intentional step towards better understanding clothing fashion. The fashion semantic space

and bimodal correlative deep autoencoder turn out to be effective. In future work, we will carry on our work in two aspects: 1) Taking users' different preferences about clothing fashion into consideration, 2) Leverage our model to build various applications.

7 Acknowledgments

This work is supported by the National Key Research and Development Plan (2016YFB1001200, 2016IM010200), and National Natural and Science Foundation of China (61370023, 61602033).

References

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

Bottou, L. 2010. *Large-Scale Machine Learning with Stochastic Gradient Descent*. Physica-Verlag HD.

Ho, C. H., and Lin, C. J. 2012. Large-scale linear support vector regression. *Journal of Machine Learning Research* 13(1):3323–3348.

Hu, Y.; Yi, X.; and Davis, L. S. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *ACM International Conference on Multimedia*, 129–138.

Jagadeesh, V.; Piramuthu, R.; Bhardwaj, A.; Di, W.; and Sundaresan, N. 2014. Large scale visual recommendations from street fashion images. *Eprint Arxiv* 1925–1934.

Jia, J.; Huang, J.; Shen, G.; He, T.; Liu, Z.; Luan, H.; and Yan, C. 2016. Learning to appreciate the aesthetic effects of clothing. In *Thirtieth AAAI Conference on Artificial Intelligence*.

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

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

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 *ACM Multimedia*, 1333–1334.

Liu, Y.; Jia, J.; Fu, J.; Ma, Y.; Huang, J.; and Tong, Z. 2016. Magic mirror: A virtual fashion consultant. In *Proceedings of the 2016 ACM on Multimedia Conference*.

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* 72:1–19.

Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; and Ng, A. Y. 2011. Multimodal deep learning. In *International Conference on Machine Learning, ICML 2011, Bellevue, Washington, Usa, June 28 - July*, 689–696.

Pedersen, T.; Patwardhan, S.; and Michelizzi, J. 2004. Wordnet: similarity - measuring the relatedness of concepts. In *National Conference on Artifical Intelligence*, 1024–1025.

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

Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; and Rabinovich, A. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1–9.

Trendowicz, A., and Jeffery, R. 2014. *Classification and Regression Trees*. Springer International Publishing.

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

Wang, F.; Li, Z.; and Liu, Q. 2016. Coarse-to-fine human parsing with fast r-cnn and over-segment retrieval. In *IEEE International Conference on Image Processing*.

Wang, F.; Zhao, Q.; and Yin, B. 2014. Refined clothing texture parsing by exploiting the discriminative meanings of sparse codes. In *IEEE International Conference on Image Processing*, 5946–5950.

Yamaguchi, K.; Kiapour, M. H.; Ortiz, L. E.; and Berg, T. L. 2015. Retrieving similar styles to parse clothing. *IEEE Transactions on Pattern Analysis & Machine Intelligence* 37(5):1028–40.

Yang, W.; Luo, P.; and Lin, L. 2015. Clothing co-parsing by joint image segmentation and labeling. In *Computer Vision and Pattern Recognition*, 3182 – 3189.