Magic Mirror: A Virtual Fashion Consultant

Yejun Liu[12], Jia Jia[1] *, Jingtian Fu[12], Yihui Ma[1], Jie Huang[1], Zijian Tong[3]
[1]Department of Computer Science and Technology, Tsinghua National Laboratory for Information Science and Technology (TNList), Tsinghua University, Beijing, China
[2]Academy of Arts & Design, Tsinghua University, Beijing, China
[3]Sogou Corporation, Beijing, China
jjia@mail.tsinghua.edu.cn

ABSTRACT

What should I wear? We present Magic Mirror, a virtual fashion consultant, which can parse, appreciate and recommend the wearing. Magic Mirror is designed with a large display and Kinect to simulate the real mirror and interact with users in augmented reality. Internally, Magic Mirror is a practical appreciation system for automatic aesthetics-oriented clothing analysis. Specifically, we focus on the clothing collocation rather than the single one, the style (aesthetic words) rather than the visual features. We bridge the gap between the visual features and aesthetic words of clothing collocation to enable the computer to learn appreciating the clothing collocation. Finally, both object and subject evaluations verify the effectiveness of the proposed algorithm and Magic Mirror system.

Keywords

Clothing Collocation, Aesthetics Learning

1. INTRODUCTION

You are what you wear. People reinforce their mood and express their feelings through their clothing. What to wear is inevitable, however, most people have no great aesthetics and appreciation for fashion and can not follow the fashion trend. So they usually struggle with what to wear.

To solve the problems, we need to figure out what makes dressing different. Instead of a single piece of clothing, the collocation of clothing plays a more important role in dressing. As well, in Nina Garcia's books, style [1] [2] is determinant of what to wear, which is the aesthetic feeling of wearing. Thus, the style of clothing collocation is the focus to appreciate the wearing.

Meanwhile, it is not appropriate to make a decision of what to wear for people, but it is helpful to provide the professional appreciation and consultation of wearing, then let people make their own decision.

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

MM '16, October 15–19, 2016, Amsterdam, The Netherlands.

DOI: http://dx.doi.org/10.1145/2964284.2970928

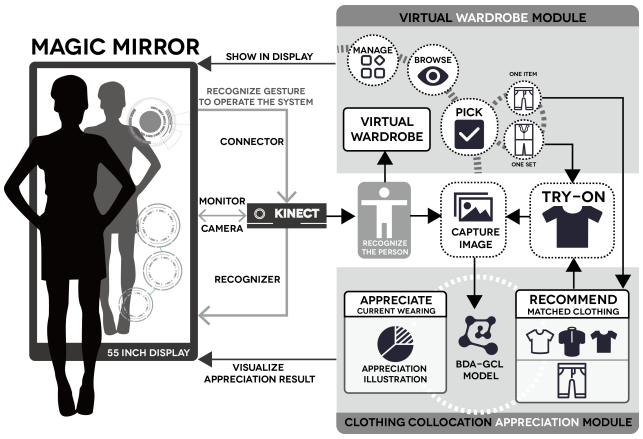

Figure 1: Overview of Magic Mirror.

So, in this paper, we designed a practical appreciation system, Magic Mirror, to be a virtual fashion consultant to parse, appreciate and recommend the clothing collocation. Magic Mirror is designed with the large display and Kinect, which simulates the real mirror and interacts with users in AR (augmented reality). Magic Mirror automatically appreciates the collocation of clothing and suggests the wearing. Through Magic Mirror, users can build their own virtual wardrobe and manage the clothing in it. As well, users can get the recommendation of matched collocation after picking one piece of clothing.

To appreciate clothing collocation, we built the association between the visual features of single-clothing and aesthetic words of collocation by proposing a three-level framework (VF-ISS-AWS). Then, the system is based on the model of Bimodal Deep Autoencoder Guided by Correlative Label (BDA-GCL), which can deeply combine visual features of tops and bottoms in the collocation and describe the mapping between combined visual features and image-scale space.

Finally, we conducted the experiments to verify the validity of algorithm and performed the user study to observe the satisfaction of appreciation and the effectiveness of recommendation.

2. MAGIC MIRROR

In normal life, people always use mirror to appreciate their own wearing directly in an other sight. To improve immersion and authenticity, Magic Mirror is not only designed to

simulate the real mirror, but also empowered to appreciate and recommend the fashion collocation in AR.

2.1 System Architecture

2.1.1 Hardware

Magic Mirror consists of a 55-inch large display, a Kinect and a computing device. The 55-inch display is placed on end like a true mirror and connected to the Kinect. Especially, Kinect is a monitor, a camera, a recognizer and a connector. It shows what is in front of the display like a real mirror, especially capturing some key image in which there is a person wearing clothing. Meanwhile, Kinect can recognize the skeleton and the gestures to optimize the interaction between users and Magic Mirror.(Figure 1)

2.1.2 Interactive experience with Kinect

In order to give users the natural experience, we designed the interaction by body languages and gestures. Users operate the Magic Mirror with the natural responding gestures. Additional, Magic Mirror can trigger the functions naturally by recognizing the appearance or disappearance of users(skeleton). Moreover, the appreciating results and suggestions are shown as tips in Magic Mirror, which are attaching to and going with the skeleton of users. Specifically, we will introduce the interactive details in chapter 2.2.1.

2.2 System Modules

To be a virtual fashion consultant, Magic Mirror, the appreciation system, has two modules: the *virtual wardrobe* module and the *appreciation* module. (Figure 1)

2.2.1 Virtual Wardrobe Module

In the wardrobe module, users can get into a virtual wardrobe and have a simple try-on looking in Magic Mirror.

First, Users can open the virtual closet by the gesture of pushing open, then browse the tops and bottoms in the wardrobe by opening or closing the hands. If users want to see the details of one piece of tops, they need to close the bottom drawer through holding and dragging down the handle of the bottom drawer. As well, the clothes in the wardrobe can be managed such as adding or deleting. Users add the clothing by hanging on one piece in display, and Magic Mirror will take a photo automatically as soon as users leaving from the mirror. Well, users delete the picked clothing by crossing gesture to confirm and tearing it apart.

In addition, the virtual wardrobe allows users to pick and match the clothes themselves and shows the simple try-on looking in the display. Hence, the virtual wardrobe module is the basic library for users to copy their real clothes to the virtual, supporting the appreciation module.

2.2.2 Appreciation Module

In the appreciation module, there are two main functions: appreciation function and recommendation function.

In appreciation, user can get the appreciating analysis of current wearing in mirror with vivid visualization. As soon as Kinect recognizes the person (skeleton) in the display, the system will capture the image and analyze the clothing collocation on it. There are several conclusions:(Figure 2)

1.Basic attributes, including color combination, suited seasons and appropriate occasions of the collocation.

2.Style distribution. It shows the distribution of relevant styles of the wearing.

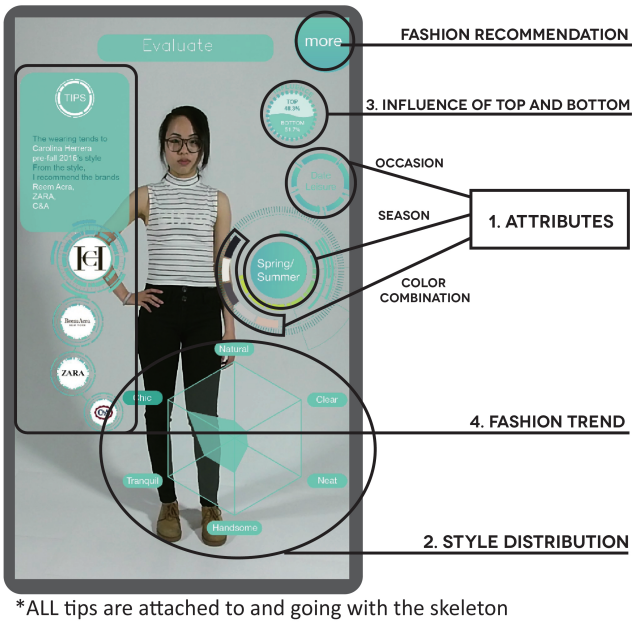

Figure 2: Introduction of Appreciation Illustration.

3.The influence of the top and bottom on style. The top and bottom of wearing will have different contribution to the style analysis.

4.Fashion trend. Based on the styles of the wearing, we can analyze which season of which brand is similar.

In recommendation, user can get several matched top/ bottom after picking certain piece of clothes in wardrobe, dividing into different relevant styles. First, user can pick one piece in virtual wardrobe and request the recommendation of matched one. Hence, Magic Mirror will give several relevant styles to choose and show the recommended collocation in different styles. Then, user will pick the willing recommended one. In the end, it will show a simple try-on looking of the combination and give the appreciation.

3. CLOTHING COLLOCATION APPRECIATION VIA BDA-GCL MODEL

How to appreciate and recommend the clothing collocation? Internally, we aimed at a clothing appreciation system, which uses the clothing photos as the input, then automatically parses the collocation of clothing.

3.1 Aesthetic Theory and Framework

In order to analyze the aesthetic style of collocation of tops and bottoms, we formulate the task to a three level framework: visual features (VF) - image-scale space (ISS) - aesthetic words space (AWS). We employ the ISS as an intermediate layer and map both visual features and aesthetic words space into it. In this way, we can label clothing images with aesthetic words automatically.

The visual features of clothing in this paper are defined from two aspects : the pattern features and color features. The pattern features of tops include collar shape, sleeve length and cut shape, of bottoms include waist height and both include sexy, length, patterns and materials. We extract pattern features by a deep neural network. The color

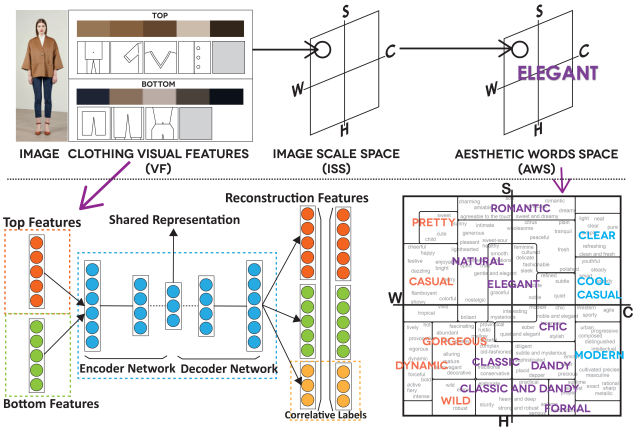

Figure 3: Framework and BDA-GCL Model.

features include five-color combination, brightness and its contrast, saturation and its contrast. ISS based on Kobayashi's aesthetic theory is a two-dimensional space (warm-cool and hard-soft), which has been used in appreciating menswear before and is proven to work effectively[3]. AWS consists of the words used to describe clothing in Amazon. Kobayashi proposed 180 keywords in 16 aesthetic categories and defined their coordinate values in ISS, considered as seed words. For a new word, we pick three of the most similar seed words depending on the semantic distances and calculate their mean value as the new coordinate value.(Figure 3)

3.2 Map visual feature to image scale space

In order to map visual features to the image-scale space (ISS), we divide the task into two steps. First, we propose a novel Bimodal Deep Autoencoder Guided by Correlative Labels (BDA-GCL) for feature learning. Second, we use several regression methods to make the new constructed features cast into two-dimensional coordinates in ISS.

A critical challenge of our task is how to calculate the specific influence of the aesthetic style by the relationship between tops and bottoms. We get several inspirations of multimodal learning[4]. More specifically, we consider the visual features of tops and bottoms as two different modals and put them into the autoencoder. We triple the training data and discard the top and bottom features separately as the other two parts data. In this way, our autoencoder can recover the whole clothing by only top or bottom features. The loss of the autoencoder can be counted as the indicator of the collocation. A smaller loss means that the tops and bottoms clothing match better.

Compared to classical autoencoder, we introduce correlative labels of clothing into the original symmetrical structure as another improvement. We choose the categories of clothing (e.g., sweater, shirt) as correlative labels. Then we use the network to regain the visual features and correlative labels at the same time. This method can be considered as a semi-supervised learning process.

As shown in the figure 3, the orange and green nodes present the top and bottom features of clothing, while the yellow nodes present the correlative labels. Through an encoder network and a decoder network indicated by the blue nodes, we regain the reconstruction features and correlative labels. And we consider the middle layer of the network as the results of feature learning.

Formally, we use the vector $x = [x_t, x_b]$ to indicate the visual features of a clothing image and c to indicate the correlative labels. The relationship between two adjacent hidden layers $h^{(l+1)}$ and h^l can be presented as:

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

where $W^{(l)}$ 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, $h^{(0)} = [x_t, x_b]$ and $[\hat{x_t}, \hat{x_b}, \hat{c}] = h^{(N_h+1)}$.

The loss function to evaluate the difference between $[x_t, x_b, c]$ and $[\hat{x_t}, \hat{x_b}, \hat{c}]$ is defined as:

$$J(W,b) = \frac{\lambda_1}{2m}\sum_{i=1}^{m}||x_t - \hat{x_t}||^2 + \frac{\lambda_2}{2m}\sum_{i=1}^{m}||x_b - \hat{x_b}||^2$$
$$+\frac{\lambda_3}{2m}\sum_{i=1}^{m}||c - \hat{c}||^2 + \frac{\lambda_4}{2}\sum_{l}(||W^{(l)}||_F^2 + ||b^{(l)}||_2^2) \tag{2}$$

where m is the number of samples, $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ is a regularization hyperparameter and $||\cdot||_F$ denotes the Frobenius norm.

We define $\theta = (W, b)$ as our parameters. The training of BDA-GCL is optimized to minimize the cost function:

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

To map visual features to the image-scale space, we further make the new constructed features produced by our model cast into two-dimensional coordinate. This step can be considered as a regression problem. We try using several regression models (e.g. K-Nearest Neighbor, Support Vector Machine, Linear Regression, Decision Tree).

3.3 Experiment

We conducted several objective experiments to validate the model by evaluating the mapping results between visual features(VF) of clothing images and coordinate values in image-scale space(ISS). In the experiment, we firstly compared the performances of different feature learning methods including denosing autoencoder(DA), bimodal strategy(B), guiding strategy(-GCL). Then we compare several regression algorithms with the output of BDA-GCL.

We employed clothing images from online clothing shop [1] and fashion website[2] as our experimental data. The dataset contains 100000 clothing images. Cooperating with Sougou company, we employ a group to manually label the dataset with the coordinate values in the ISS. In the evaluation of the mapping effects, we calculated the error between predicted coordinate values and labeled coordinate values. The error is formulated by mean squared error (MSE[3]). All the experiments are performed on five-folder cross-validation.

The experiment results are shown in Table 1 and Table 2. It is clear that our model BDA-GCL (MSE:0.1854) performs better than the baselines. And the results support the effectiveness of our guiding strategy and bimodal strategy. Comparing among several regression algorithms, we find that SVM and LR have the best results.

[1]http://www.Amazon.com
[2]http://www.style.com
[3]https://en.wikipedia.org/wiki/MSE

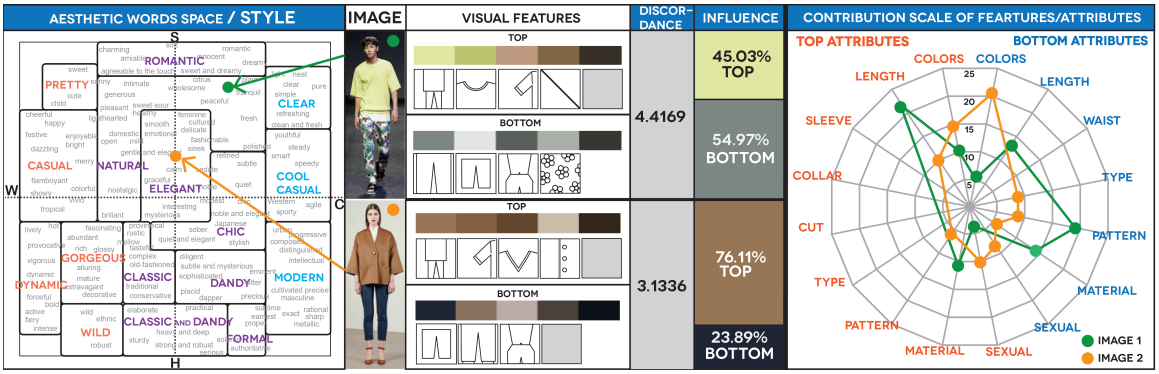

Figure 4: The Statement of Analysis Results.

Table 1: Comparison among different feature learning methods

Autoencoder	Regression	MSE
None	SVM	0.1927
DA		0.1928
DA-GCL		0.1860
BDA		0.1881
BDA-GCL		0.1854

Table 2: Comparison among different regression models

Autoencoder	Regression	MSE
BDA-GCL	D-Tree	0.3378
	KNN	0.3324
	LR	0.1854
	SVM	0.1854

3.4 Analysis Results

Using our framework, we can not only analyze the style of certain clothing image, but also conclude several interesting aesthetic rules to support our recommendation:(Figure 4)

1. The distribution of relevant styles of the wearing. The collocation of the top and bottom locates at a coordinate in the aesthetic words space. We can calculate the correlation between the coordinate and rounded styles.

2. The different roles of tops and bottoms clothing playing in the aesthetics effect. We compare the errors between using whole features and using single tops or bottoms features to analyze the different contributions of them.

3. The discordance of tops and bottoms. We use the value from zero to infinite to measure the discordance that if the value close to the infinite, the collocation is more discordant.

4. The different contributions of various visual features. For example, as shown in the Figure 4, the length of tops have larger influence in the Image 1 than that in Image 2, while the color features play a more important in role in the Image 2 than Image 1.

5. The indicator of the collocation. Our autoencoder can measure the harmony of tops and bottoms. So even with the loss of piece, we can get the matched one.

6. The aesthetic style of brands. We analyze the aesthetic style distribution of brands and take them into consideration when we recommend clothing.

4. USER STUDY

We employed 11 people, 5 males and 6 females, to participant the user study. First, participants try to use Magic Mirror to appreciate the current wearing. Then we investigated their satisfaction of appreciation that 10/11 participants thought Magic Mirror gave them great appreciation about wearing. Next, we provided them some pieces of top that let them choose the matched bottom in several limited styles from the wardrobe. We compared their choices with the first five recommended ones by Magic Mirror. 89 percent choices are in the first five. Hence, the recommendation of Magic Mirror is appropriate and rational. According to the study, we can conclude that the Magic Mirror truly gave the similar wearing suggestions of the participants and make them feel appreciated.

5. FUTURE WORK

In this work, Magic Mirror mainly focused on mining general rules based on the popular aesthetics and fashion trend to guide users wearing. In fact, personal factors have no small impact on wearing in daily life, such as user's taste and figure. In the future, we will investigate such personalization thoroughly.

6. ACKNOWLEDGMENTS

This work is supported by National Key Research and Development Plan(2016YFB1001200), the National Basic Research Program (973 Program) of China (2012CB316401) and National Natural, and Science Foundation of China (61370023).

7. REFERENCES

[1] N. Garcia. *The little black book of style*. Harper Collins, 2010.

[2] N. Garcia. *Nina GarciaâĂŹs Look Book: What to Wear for Every Occasion*. Harper Collins, 2010.

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

[4] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.