https://doi.org/10.2352/ISSN.2470-1173.2018.10.IMAWM-446

Use of Color Information in the Analysis of Fashion Photographs*

***Zhi Li*[a], *Gautam Golwala*[b], *Sathya Sundaram*[b], *Jan Allebach*[a];**
[a] *School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 47907, U.S.A;*
[b] *Poshmark Inc., 101 Redwood Shores Pkwy, 3rd Floor, Redwood City, CA 94065*

Abstract

The 21st century witnesses a blooming of online fashion retailing business, as well as the Peer-to-Peer (P2P) marketplaces. However, many listings on the P2P marketplace lack complete and accurate information. In this research, we target the fashion online marketplace, and try to retrieve garment color information to help sellers and buyers gain a more comprehensive knowledge of the fashion items. We focus on fashion product portraits, and propose a system that autonomously finds the garment region in the image and retrieves the color information and patterns of the item. This system contains three modules: First, image segmentation is deployed to partition the image into perceptually meaningful areas, and then some features are designed to differentiate the garment, region from non-garment region. Secondly, we propose a classifier module based on unsupervised clustering methods to select the garment region based on the feature vector. For the last module, we study current color naming systems, and color naming schemes on fashion websites, and propose a computational model that matches the color coordinates with the pre-defined color labels in the marketplace. Compared with other methods, thanks to the unsupervised learning methods that we use, our approach does not require a huge amount of training data labeled by human subjects.

Introduction

Online shopping is an exponentially growing market. Retail sales world-wide, including both in-store and internet purchases, totaled approximately $22.5 trillion in 2014, with $1.316 trillion of sales occurring online. By 2018, ecommerce retail spending is projected to increase to nearly $2.5 trillion† [1]. However, in recent times, public attention has switched from traditional Business-to-Customer (B2C) to Peer-to-Peer (P2P) business model.

The P2P online marketplace allows its users to post and sell their items to other users on the platform. This flexible and dynamic shopping platform has rapidly drawn many users and gained popularity. However, P2P marketplace has many issues with incomplete and unorganized information. Most P2P e-commerce websites only maintain the web services and provide marginal services to the users. This creates a more friendly and easy-to-use atmosphere for both shoppers and sellers, leading to

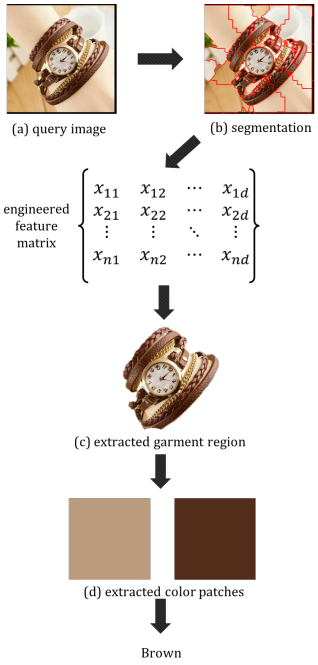


Figure 1: An overview of the Autonomous Garment Color Extraction System (AGCES). The system employs segmentation and engineered features to select the garment region. The final segmentation mask is used to find the color of the item.

*Research supported by Poshmark, Inc. Redwood City, CA, 94065

†`http://www.emarketer.com/Article/Retail-Sales-Worldwide-Will-Top-22-Trillion-This-Year/1011765`

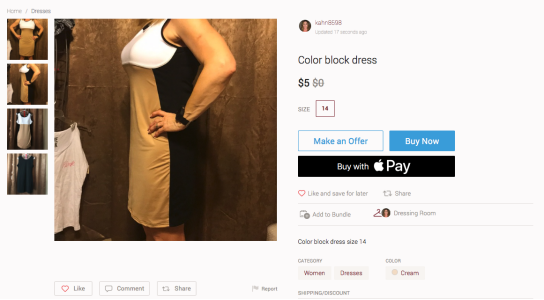


Figure 2: Example of the case of incomplete color input. In this listing, based on the images provided by the user, the colors input should be white, black, and cream. However, only the cream color is marked. Source:Poshmark.com.

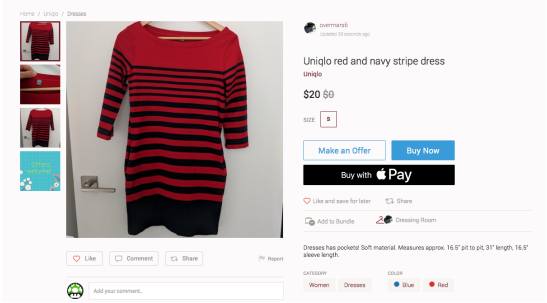


Figure 3: Example of the case of photography color inconsistency. In this listing, the seller (overmars6) is selling a dress of red and blue stripes. However, due to various reasons, the blue strips appear closer to black. Source:Poshmark.com.

a more dynamic shopping experience. However, it also creates some problems due to the lack of management and attention to details. Some of the information provided by average sellers is not correct, and some products are missing some very important information. This type of problems often causes miscommunication or publicity issues, leading to a slower sales performance or a worse shopping experience.

For the fashion market, problems on information organization also persist. For example, as one of the most defining features of a fashion product, color information on P2P websites is not always provided by the sellers; and sometimes the color information provided by the sellers can be inaccurate.

- **Incomplete or inaccurate user input.** Incomplete user input includes two kinds: First, sellers might simply forget or omit the part of entering the item's color(s). Another scenario is that sellers fail to report the full set of the colors contained in the garment. An example is given in Figure 2, in which only one of the three colors is labeled by the user. There are also some cases where the sellers provide the wrong color. Inaccurate input is less likely to happen compared with incomplete input. But both of these two cases are frequent enough to influence the overall site searching accuracy, and hence to affect the selling/shopping experience.
- **Photography color inconsistency.** This means that the color shown in the images does not agree with the color information provided by the seller, or the real colors of the item. Fig. 3 shows an example of color inconsistency. This can be caused by an undesirable lighting condition, the device's color reproduction capabilities, or additional photo post-processing done by the user. Note that some of the colors have very similar appearance, for example, black vs. navy blue, pink vs. cream white, and gold vs. yellow. In this case, it is hard for human viewers to distinguish the fine differences between colors by looking at a single color patch. A reference color is usually recommended to provide a more accurate color description.

Hence, there are several factors that come to our attention to improve the quality of the color information. First of all, a more user-friendly photography guide is developed to assess and potentially improve the quality of the fashion photography. Previously we have done research to develop an SVM based aesthetic quality predictor to assess the aesthetic quality of an image. Later, we proposed a theory to improve the fashion image aesthetic quality by modeling the garment and using a background of simpler and non-cluttered color[2]. This serves as a guidelines for the users to provide better and more accurate product images.

Secondly, we would like to design a fashion image understanding system that can autonomously identify the garment region from a fashion image and extract the garment color. This result can be used to fill in some missing information on the website, and also allows user to reflect how accurate their photos and information are. And this is the targeted problem we study in this paper.

We propose the Autonomous Garment Color Extraction System to extract color from the fashion portrait image (AGCES) shown in Figure 1.

Related Work

As stated in the previous section, in order to find the color(s) of the garment, we need to find the garment region in the image first. This task can been generalized as a semantic segmentation problem, where each pixel of an image is labeled with a physical meaning corresponding an object class. After we retrieve the pixel values of the garment region, we will be able to use color matching algorithm to assign the correct label to

the color, hence name the color of the garment. In this section, we provide an overview of previous works in related research areas, namely image semantic segmentation, fashion image understanding, and color naming analysis.

Semantic Segmentation

Semantic segmentation, also known as *scene parsing*, is one of the most critical image understanding problems. Unlike unsupervised image segmentation, which aims to partition an image into coherent small regions according to the low-level cues including color or other metrics [3], semantic segmentation requires the algorithm to understand the image, and classify each pixel of an image into one of the several predefined object classes [4]. It has been studied for a long time, due to the fact that more applications benefit from inferring knowledge from imagery [5], including autonomous driving [6, 7], and medical imaging [8].

Prior to the recent explosion of computer vision applications of Convolution Neural Networks (CNNs), the most common approach to the semantic segmentation problem was to use Conditional Random Fields (CRFs) [4, 9, 10]. However, these traditional computer vision scene parsers are limited by the structure of the classifiers; And it is difficult to engineer capable image features to work on all the scenes. CNNs and recent high-performance computing system developments enable image recognition research to achieve new state-of-art performance. Krizhevsky *et al.* [11] train an ImageNet network with 1.2 million images to do classification for 1000 different classes. Many research efforts have been done to develop many variation of the network and to improve its performance [12, 13, 14].

Long *et al.* [15] re-structured the Convolutional Network to direct, dense prediction of semantic segmentation. This Fully Convolutional Network (FCN) structure provides an end-to-end learning solution for semantic segmentation problems. However, like any other supervised learning method, these neural network based methods are limited by the ground truth and the predefined class. Although most of the popular datasets include many instances of humans, the datasets that include different clothing information are relatively rare.

Fashion Image Understanding

Fashion understanding has been a very popular topic recently, not only because of the growing fashion industry that is estimated to be worth 2.5 trillion dollars in next four or five years [16], but also because clothing understanding can be an important clue in many human-centric research and analysis. Some of the current advanced methods include clothing semantic attribute summarization [17, 18], fashion landmark detection [19, 16], and trend discovery[20, 21, 22]. One of the most active topics is clothing retrieval.

Recognizing clothing is a challenging problem because of the wide variation of clothing appearance, layering, style, and interaction with the human body. However, some progress has been made to pursue this goal. Recent practices combine visual cues from human body detection and fashion landmark localization. A popular approach is to train a bounding box detector to find the garments in the image [1, 23]. In addition, for pixel-wise clothing parsing, Yamaguchi *et al.* [24] developed a clothing parser based on pose estimation and a CRF based on the Fashionista dataset. More recently, Yamaguchi *et al.* [25] proposed a framework to produce pixel-wise labeling by analyzing images that are similar to a query image.

One of the challenges is that there is not a sufficiently generic public fashion dataset available for fashion semantic segmentation. Most of the datasets are similar: street shots with full or part of a human body[19, 24]. This is not necessarily the case for fashion online shopping.

Color Naming System

Many computer vision research efforts and applications have used color information of an image or video. Yet, the ability or complete theories to name individual colors, pinpoint objects of specific colors, and communicate the impression of a certain color composition is still under development [26]. Color interpretation is highly intertwined with the image content, and other factors such as human perception, culture, and linguistics. Therefore a flexible computational model for color categorization is desired.

Color Naming Standard

Although many well defined numerical color spaces have been developed, and these methods have been proved to be effective in terms of most color-related image processing and computer graphics tasks, in everyday life the most common way to communicate about color is through verbal description. Thus, a color naming system is desired to transform information from color spaces to color names. So far there are several proposed work mapping color coordinates to a verbal description. One of the most commonly used color naming system is the Munsell color standard [27]. Munsell Color System was developed in the late 1800s by Albert Henry Munsell[‡]. It has been widely used in paint and textile production [26]. However, this proposed system lacks a color vocabulary and an exact transform from any color space to Munsell. Derived from the Munsell system, The National Bureau of Standards(NBS) developed ISCC-NBS dictionary of color names according to the recommendation of Inter-Society Council [28].

[‡]http://munsell.com/color-blog/munsell-color-order-system-what-is-it-and-how-is-it-used/

Some other alternative systems have been proposed as well. An extension of the CNS model is the Color Naming Method (CNM). This was originally proposed by Tominaga [29]. This method utilizes a predefined set of color names in the Munsell color space, and proposes a method to map pixel values to specific color names using an optical measurement system. The color names in this method are specified at one of four accuracy levels (fundamental, gross, medium, and minute) so that names from the higher accuracy level correspond to smaller color regions in the Munsell space [26]. Belpaeme [30] built a another color categorization framework based on the notion of color primitives surrounded by color regions with fuzzy boundaries and modeling via adaptive radial basis function networks [26]. Mojsilovic [26] studied the *National Bureau of Standards'* color recommendation for color names, and developed a new vocabulary and syntax. He also proposed a new perceptually based color-naming metric to match an arbitrary input color to a color name.

Fashion Color Analysis

For color management in the fashion industry, unlike other industries, where color names usually are given based on the color appearance, most of the fashion color naming is done by the manufacturers individually before products are released. Other than that, seasonal trending color names are also predefined every season by the team at the Pantone Color Institute in the Pantone Fashion Color Trend Report [31]. However, these color naming practices do not use any standard color naming dictionary. They only consider trending colors, which change constantly by the season. Besides, these names are commonly used among fashion experts and high-end fashion brands. Yet it is not a standard color naming system that is universally recognized by all fashion manufacturers.

However, several research projects have been conducted to do color mapping from color coordinates to color names defined by researchers that are more specific and distinct [32, 17]. In our case, color labels are more artistically descriptive than scientifically accurate.

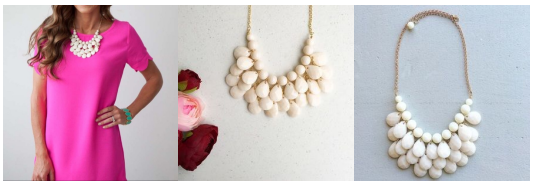
System Overview

Bound the Problem

When it comes to fashion image analysis, it is no surprise that there is a huge variety of fashion images in online fashion marketplaces. Some of the images have very complex combinations of fashion items, many layers of clothing and other accompanying objects in the image. These images are usually designed and produced professionally or delicately; and they usually have higher aesthetic quality. However, for this type of image, as discussed in the previous section, traditional feature based machine learning has difficulty achieving the semantic segmentation to identify different garments within pixel level accuracy. Therefore, in this case, only product portraits are considered in this research.

Product portraits are images that only feature the target items, in our case, the items that sellers want to sell. This allows the user to include some other items to "decorate" the image to a limited extent, as long as the item as sale is staged for the highlight of the image.

Some examples of product portraits and counterexample are given in Fig.4. Our system is design to process images like (b) and (c).



(a) (b) (c)

Figure 4: Source: Poshmark.com. This is a set of example images of fashion photos. All three images are posted by the same Poshmark seller in the same listing to promote the necklace, but in (a) the promoted item is overshadowed by the pink dress. The other two, (b) and (c), directly highlight the target item.

System Overview

We propose a system that autonomously extracts the color of the featured garment from a fashion image. As shown in Fig.5, this autonomous color extraction system can be divided into three modules: the segmentation module, the grouping module, and the color processing module. The segmentation module takes the input image and cuts it into smaller segments that share similar characteristics. The grouping module uses the segmentation information, calculates the features of each segments, and groups them into two classes: garment and non-garment regions. The last module is deployed to extract the color information, and to transfer numeric color coordinates into descriptive words, such as "blue", "pink", or "white". All modules will be explained in the following sections.

Image Segmentation

In this module, we decompose a given image into smaller perceptually homogeneous segments. By doing this, the computation and complexity of the algorithm is reduced. We propose the image segmentation structure that is prone to preserve edges in the image, and which is sensitive towards both color and texture.

We use SLIC superpixels [33] to abstract the image into perceptually uniform elements. The SLIC algorithm calculates

pixel perceptual distance in LABXY space, and deploys K-means to group similar pixels together as a superpixel. By combining both color and spatial information, SLIC superpixels are local, compact and edge aware. Because of the nature of the SLIC superpixels, the algorithm tends to render noisy superpixels where lines and edges congregate. Therefore, we adopt the preprocessing steps to the image proposed by Wang et al. [34] to smooth the image before superpixel segmentation. Some comparison results are given later.

We further apply graph-based image segmentation techniques to group superpixels together, namely the Region Adjacency Graph (RAG) [35]. In more generic situations, the graph representation usually considers color difference between two neighboring superpixels as the only aspect of the edge weight. For fashion images, however, with the goal of accentuating the different fabrics or materials, we update the weight using a new method developed by Wang at al.[34]

This method combines texture difference and color difference by calculating Local Binary Pattern (LBP) within 3×3 neighborhood and CIE ΔE. The standard 1976 CIE ΔE is used to calculate the color similarity measurement between the average colors of two neighboring superpixels. For texture measurement, the LBP an be calculated based on every pixel in each superpixel. For pixel p and its corresponding superpixel sp, if all the pixels surrounding of p are also in sp, we do following:

1. Compare the L values of each surrounding pixel with the central pixel in a clockwise order. If the neighboring value is greater or equal to the central value, then the output is 1, otherwise it is 0.
2. Concatenate the binary results and form an 8-digit binary number.
3. Convert the 8-digit binary number into a decimal number.

After getting all the decimal numbers of the eligible pixels within the superpixel, a histogram is sampled into 9 bins and normalized. The final vector of normalized occurrences from the histogram is the LBP vector.

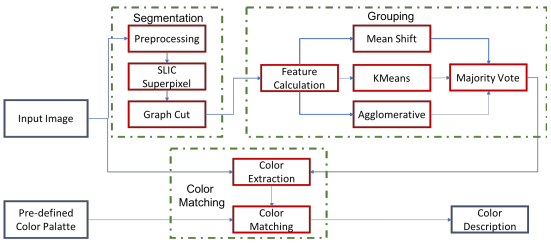


Figure 5: The pipeline of the Autonomous Garment Color Extraction System (AGCES).

To obtain a single value of edge weight, the CIE ΔE value and χ^2 distance in LBP vector space are calculated and mapped to a probability estimate, and the edge weight is assigned using criteria proposed by Wang [34].

After combining both texture and color measurements, we are able to produce the new edge weights, and hence deploy Normalized Cut [36] to composite the graph into segments. The segmentation results and comparisons are given in Fig.6.

Figure 6: The segmentation results. Row (a) is the original image; row (b) is the images with preprocessing and LBP as texture cue; row (c) is the images without preprocessing; row (d) is images with regular RAG, without preprocessing and LBP as texture cue.

In the picture of jeans, some noisy residual segments are rendered right below the folded part as can be seen in Row (c) and (d) of Fig.6. And the preprocessing (shown in the second row) eliminates such noisy regions. However, there are also some drawbacks from the smoothing. For example, for the picture of bra, the segmentation results without smoothing shows more fidelity to the original object edges. As for the texture, In the picture of the lady with sunglasses, the texture of the background has been preserved thanks to the LBP algorithm.

Feature Engineering and Segment Grouping

The group module is used to group all the segments into two groups by determining whether each segment should be in background or in the garment region. This can be further divided into two steps: to design discriminative features to differentiate garment and background as much as possible; and to use the calculated features to make decision.

Feature Engineering

After we investigate many fashion product portraits, we find that there are some features that can be generally applied to garment regions in all fashion product portraits: a high volume

of high-frequency details; great contrast from the background; and dominant, prominent positioning.

Therefore, as shown in Table.1 we design four sets of measurements: a Laplacian feature set, a Perrazi contrast feature set, a positioning feature set, and a low level feature set. Among all four feature sets, the low level features are easy to obtain by averaging pixel values in CIEL*a*b space and XY space, respectively. Thus these two features will not be covered in full detail.

Table 1: Feature Symbols

Feature Sets	Feature Names	Symbol
Laplacian	1st Order Laplacian	$\mathbb{L}^{(1)}$
	2nd Order Laplacian	$\mathbb{L}^{(2)}$
Perrazi	Color Uniqueness	$\mathbb{U}$
	Distribution	$\mathbb{D}$
Positioning	1st Boundary	$\mathbb{F}^{(1)}$
	2nd Boundary	$\mathbb{F}^{(2)}$
	Standard Deviation	$\mathbb{E}$
Low Level	Mean Color Values	–
	Centroid Position	–

Laplacian Feature Set

Fashion product portraits are shot to present the fine details of the garment. Therefore, capturing constructive high-frequency details of the image can help determine the garment region.

Since the Laplacian pyramid [37] can also catch image details at different scales, we adopt the Laplacian power summation features proposed by Wang at al[37]. A 2-level Laplacian pyramid is built and the pixels of Laplacian images are represented by their absolute values to focus on the detail strength. One example is shown in Fig.7, where a black or darker pixel mean the Laplacian difference at that position is low.

First, we calculate the Laplacian power images $\mathbb{L}^{(1)}$ and $\mathbb{L}^{(2)}$ by generating first and second Laplacian pyramid layers of the gray scale image and taking the absolute value. For each pixel p, its corresponding Laplacian power in the i-th layer $\mathbb{L}^{(i)}$ is $l_p^{(i)}$. Thus, for the i-th segment, its k-th order Laplacian power $\mathbb{L}_i^{(k)}$ can be calculated by

$$\mathbb{L}_i^{(k)} = \frac{1}{|C_i^{(k)}|} \sum_{p \in C_i^{(k)}} l_p^{(k)}, \tag{1}$$

where $C_i^{(k)}$ is the set of pixels in k-th layer that belong to segment i, and $|C_i^{(k)}|$ is the number of elements of this set. Note that for second layer, we also down sample the segmentation map to match the size of the Laplacian image.

Results of Laplacian features for for some images are shown in fig.8 column (c) and column (d) respectively.

Perrazi Feature Set

For fashion product portraits, the garment should be highlighted in a situation where great contrast is created. Perrazi et al.[38] proposed an algorithm that produces image saliency maps by evaluating color contrast and distribution contrast. However, the Perrazi contrast is calculated at a superpixel level. Therefore, a transform from superpixel level features to segment level features is required. The color uniqueness U_i for superpixel i can be calculated as

$$U_i = \sum_{j=1}^{N_s} w_{i,j}^{\mathrm{p}} |\mathrm{c}_i - \mathrm{c}_j|^2, \tag{2}$$

where N_s is the number of SLIC superpixels in the image, $w_{i,j}^{\mathrm{p}}$ is a Gaussian weight related to the spatial correlation between superpixel i and j, and c_i is the mean color coordinates of the superpixel i.

The distribution of superpixel i is given by

$$D_i = \sum_{j=1}^{N_s} w_{i,j}^{\mathrm{c}} |\mathrm{p}_i - \mathrm{p}_j|^2, \tag{3}$$

where $w_{i,j}^{\mathrm{c}}$ is a Gaussian weight related to the color correlation between two superpixels. Both $w_{i,j}^{\mathrm{p}}$ and $w_{i,j}^{\mathrm{c}}$ are intended to yield local contrast term, which brings more sensitivity to superpixels that are similar color-wise or position-wise , respectively. These terms are given by

$$w_{i,j}^{\mathrm{p}} = \frac{1}{Z_j} \exp\left(-\frac{|\mathrm{p}_i - \mathrm{p}_j|^2}{,} 2\sigma_p^2 \right) \tag{4}$$

and

$$w_{i,j}^{\mathrm{c}} = \frac{1}{Z_j'} \exp\left(-\frac{|\mathrm{c}_i - \mathrm{c}_j|^2}{,} 2\sigma_c^2 \right) \tag{5}$$

(a) (b) (c)

Figure 7: Visualization of Laplacian pyramid. (a) is the original image, and (b) is the 1st layer of the Laplacian pyramid. Likewise (c) is the 2nd layer.

Note that Z and Z' are scalars for the purpose of normalization, which means $\sum_{j=1}^{N_s} w_{i,j}^{(\mathrm{p})} = 1$ and $\sum_{j=1}^{N_s} w_{i,j}^{(\mathrm{c})} = 1$. Empirically, we choose the parameter $\sigma_p = 0.25$, and $\sigma_c = 20$.

Once the superpixel-level color uniqueness and distribution are obtained, we compute the segment-level features. For segment i and its corresponding set of superpixels C,

$$\mathbb{U}_i = \sum_{s \in C} \frac{m_s}{M_i} U_s, \tag{6}$$

and

$$\mathbb{D}_i = \sum_{s \in C} \frac{m_s}{M_i} D_s, \tag{7}$$

where m_s is the number of pixels in superpixel s, and M_i is the total number of pixels in segment i. The sample results for color uniqueness and distribution features are shown in Fig.8 column (e) and (f).

Positioning Feature Set

The first feature that we would like to look at is how close the segment is to the image boundary. Although it is not a hard line, the garment region tends to be located at the center of the image, or at least according to the Rule Of Third guidelines. Therefore, a frame or boundary feature is developed to show how many pixels of a given segment are close to the borders. For a segment i, the frame feature $\mathbb{F}_i$ can be described as

$$\mathbb{F}_i = 1 - \frac{|F_i|}{M_i}, \tag{8}$$

where F_i is the set of frame pixels in the segment i, and M_i is the total number of pixels in segment i. When $\mathbb{F}_i \to 0$, more pixels inside of segment i are on the boundary, and it is less likely to be a garment segment. When $\mathbb{F}_i \to 1$, fewer pixels are on the boundary, and the segment is more likely to be in the garment region. However, the unanswered question is how to define the boundary pixel.

We use two different definitions of boundary for this task:

1. Circle boundary $\mathbb{F}^{(1)}$: a pixel p is a boundary pixel if its Euclidean distance between p and the center of the image is larger than a preset threshold r.
2. Box boundary $\mathbb{F}^{(2)}$: a pixel p is a boundary pixel if its distance to the closest border of the image is smaller than a preset threshold h.

The parameter $\mathbb{F}^{(1)}$ is focused on the central area of the image, and it is prone to overlook the pixels close to the corners; $\mathbb{F}^{(2)}$ is closer to the image frames in real life, but less selective. We use $r = 0.3 \cdot \min\{W,H\}$ and $h = 0.1 \cdot \min\{W,H\}$, where W and H are the width and the height of the image, respectively. Our experiments show that a combination of both $\mathbb{F}^{(1)}$ and $\mathbb{F}^{(2)}$ yields a better result.

We also want to use the standard deviation of each segment to measure its 2D dispersion. For the i-th segment, and the set of its pixels S_i,

$$\mathbb{E}_i = \sqrt{\frac{1}{|S_i|} \sum_{p \in S_i} \left(x_p - x_c^{(i)}\right)^2 + \left(y_p - y_c^{(i)}\right)^2} \tag{9}$$

where $(x_c^{(i)}, y_c^{(i)})$ is the coordinates of the centroid of the i-th segment. Compare with distribution contrast $\mathbb{D}$ which considers the superpixel distribution and the color difference, the standard deviation feature $\mathbb{E}$ is more focused on how the pixels of the segment are distributed regardless of other visual cues. Results for these cues for some fashion product images are shown in Fig.8 column (g), (h), and (i).

Processing Features

The First step is to normalize the feature vectors. The original feature values have different ranges. For example, as one of the feature, the average L channel value in CIE L*a*b* can range from 0 to 100, while the a*, b* values can be negative. Other than that, the range can also be a variable. For example, the range of the centroid position, which is the average (x,y) coordinate, is determined by the size of the image itself. Hence, it would be unwise to input raw feature values without any normalization into the next step.

For a given feature vector $\mathbb{V} = [v_1, v_2, ..., v_N]$, where v_i is the feature value of i-th segment. The normalized feature vector $\mathbb{V}'$ can be expressed as

$$\mathbb{V}' = \frac{\mathbb{V} - \min(\mathbb{V})}{\max(\mathbb{V}) - \min(\mathbb{V})}. \tag{10}$$

Therefore, all the feature values are restrained to the range [0,1], and their individual impact to our grouping system is unified. Different weights can be multiplied by the normalized feature vectors to adjust their significance to achieve an optimal selection result. Our experiments show that in general the most desirable result occurs when only the standard deviation feature $\mathbb{E}$ is scaled by 0.7.

As stated in the previous section about the positioning feature set, a combination of two boundary features is used in the final feature set. The circle boundary $\mathbb{F}^{(1)}$ tends to be more progressive than the box boundary $\mathbb{F}^{(2)}$: it works well when the item is located at the center of the image. But for images with large garment pieces, where the garment spread horizontally or vertically in the image, it cuts through the garment region. Therefore, in the final feature set we only use the normalized $F^{(2)}$ feature to do the clustering, as well as the inner product of both. The

new inner product feature $\mathbb{F}$ is defined as

$$\mathbb{F}_i = \mathbb{F}_i'^{(1)} \cdot \mathbb{F}_i'^{(2)}, \tag{11}$$

where $\mathbb{F}'$ means the normalized feature value. For central segments and marginal segments, $\mathbb{F}$ performs similarly as $\mathbb{F}^{(1)}$ or $\mathbb{F}^{(2)}$, since central segments always have fewest boundary pixels and marginal segments always have the most boundary pixels. In the transition area between the image center and the boundary, $\mathbb{F}$ behaves more conservatively. Based on our experiments and testing, the weight of this feature is set to be 0.6.

Segment selection

Our final goal in this module is to produce a mask that only recognizes the garment region. This can be seen as a pixel level binary classification task: every pixel in the image should be labeled as either garment or non-garment. As stated in the introduction section, pixel level image segmentation has been mostly studied using supervised statistical learning especially deep neural networks. However, by utilizing the features introduced above, we are able to approximate the segmentation results by clustering algorithms.

This approach seems not as "smart" as neural networks and other supervised classification algorithm, and learning with ground truth heuristically performs better as they utilize each data point individually. However, clustering algorithms allow us to avoid collecting a very comprehensive dataset for training and testing. For a single image, we can feed clustering algorithm varying number of image segments as input, as our segmentation algorithm does not produce a fixed number of segments. This also enable us to consider the correlation between segments within a given image.

However due to the nature of clustering algorithms, the clustering result is less predictable and stable. Therefore, we propose a semi-supervised learning structure. First, we choose three different clustering algorithms to perform clustering generating three different clustering results $\Theta^{(1)}$, $\Theta^{(2)}$, and $\Theta^{(3)}$.

$$\Theta^{(k)} = \begin{bmatrix} \theta_1 & \theta_2 & \cdots & \theta_N \end{bmatrix}, \tag{12}$$

where N is total number of segments in the image, and for i-th segment, the prediction from k-th algorithm is binary: 1 means garment region, and 0 means other. Then, we process all the clustering results with certain criteria.

Here, we employ three different algorithms: the meanshift algorithm, the K-means algorithm, and the agglomerative clustering algorithm. K-means algorithm is widely used the algorithm and it serves very general purposes. The agglomerative algorithm performs a hierarchical clustering procedure by a bottom up approach. The linkage criteria is set to be ward distance, which minimizes the sum of squared differences within all clusters. The meanshift clustering aims to discover blobs in a smooth density of samples[39]. However, the meanshift algorithm is generally used where cluster sizes are similar; so multiple iterations should be used to secure the clustering result. In Fig.10, we show some results from our semi-unsupervised segment selection method. Pseudo colors, red and blue, are overlaid on the image to distinguish the two different labels.

Unlike supervised classification algorithms, the labels produced by the clustering algorithm usually only serve the purpose of separating different clusters; and they do not have a practical meaning. Therefore, the actual labels produced by the three algorithms might have different meanings and may not be consistent. We make all the clustering results consistent by doing following steps: first, find a pixel that can be used as a reference, which should be labeled as background, and then examine all the labels to see if the produced label matches the reference label. If not, we take the complement of the original labels as the new label.

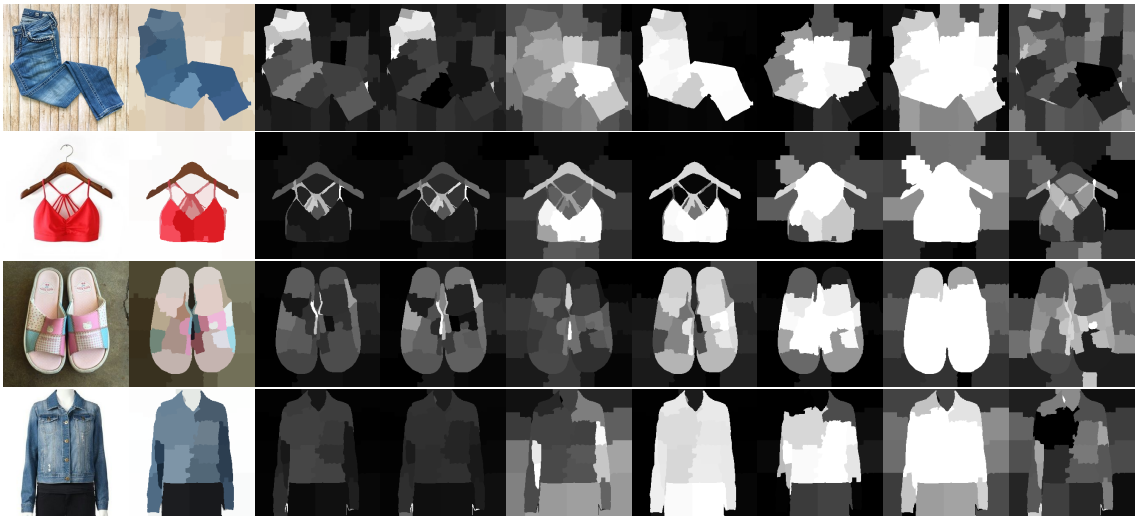
To choose the reference pixel that is in most cases a background pixel, we investigated many portraits photographs and found that for image composition, the most important objects are aligned with or on the baroque diagonal, which is from the lower left to the upper right corner. The baroque diagonal is said to provide a more pleasing and positive viewing experience, while the sinister diagonal has negative connotations. So most product portraits avoid arranging items on the sinister diagonal [40]. In Fig.9, some examples are given to demonstrate object alignment in portrait photography. In the other fashion portrait examples shown in the earlier figures in this paper, it is rare to see items that follow the sinister diagonal. Thus, we use the top left pixel as the non-garment reference pixel to calibrate all three clustering results. Some examples are shown in Fig.10.

Then, we use a majority voting scheme to finalize the results. The algorithm is given below. When disagreement exists, $\mathbb{F}^{(2)}$ is used to eliminate one or more border segments, making the selection more conservative.

Algorithm 1 Majority Vote

for $i = 1$ to N **do**
  if $\theta_i^{(1)} = \theta_i^{(2)} = \theta_i^{(3)} = \theta_i$, **then**
    $\theta^* = \theta_i$
  else if $\mathbb{F}_i > 0.5$, **then**
    $\theta^* = 0$
  else
    $\theta^* = \sum_{k=1}^{3} \theta_i^k - 1$
  end if
end for

The final segmentation results are shown in the last column



(a) (b) (c) (d) (e) (f) (g) (h) (i)

Figure 8: Features for segment grouping. (a) original image, (b) RAG segmentation results, (c) 1st order Laplacian power $\mathbb{L}^{(1)}$, (d) 2nd order Laplacian power $\mathbb{L}^{(2)}$, (e) color contrast $\mathbb{U}$, (f) distribution $\mathbb{D}$, (g) circle boundary indicator $\mathbb{F}^{(1)}$, (h) box boundary indicator $\mathbb{F}^{(2)}$, (i) standard deviation $\mathbb{E}$.

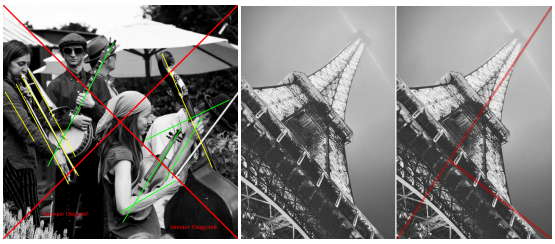
in Fig.10. In general, the algorithm produces robust results.

Color Processing

The last module of the Autonomous Garment Extraction System finds the color from the garment region, and maps it to a certain color description.

The Gaussian Mixture Model is widely used in color imaging to extract the mean color from a population of color data points. Here we use the Gaussian Mixture Model and the Expectation Maximization algorithm to obtain the estimated average color(s) from given color vectors contained in the garment region. We set the number of Gaussian distributions to 2 to generate at least two colors for each image. If these two colors are very similar to each other ($\Delta E < 30$), then we can say that this garment only has one color. A sample result is shown in the Fig,11. The GMM-based summarization matches the human visual perception.

The next step is to match the color coordinates with the



(a) (b)

Figure 9: Examples of Baroque and Sinister diagonal analysis in photography. In (a), most object align along or on the two diagonals [40]. Based on (b) [41], the baroque diagonal is more visually pleasing or important than the sinister diagonal.

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

Figure 10: The intermediate clustering results and the final clustering result. Column (a) contains the original images; Column (b) shows the results from applying the meanshift algorithm; Column (c) contains the results of the agglomerative clustering with Ward distance; Column (d) contains the results of applying the K-means algorithm; the rightmost column shows the final result.

pre-defined color palette on the fashion retail website, shown in Fig.12(b). We extract the RGB values from the website. Then, we calculate and compare the color differences between extracted mean colors and all the reference colors. This is a very simple solution. It generally works well when the colors are bright and highly saturated, for example red and green. But it doesn't work well in other situations. Therefore, we conduct some research on color naming schemes on retailer websites, and also on the users' end.

Fig.12 shows two different color definitions from two different online fashion retailers, Neiman Marcus and Poshmark. There are some findings:

1. Although most websites share some basic colors like red, green, yellow, the full color palette can be very different for different websites.
2. Some color families have a more granular color description. For example, for the yellow color family, websites usually have yellow, orange, gold and more, while there is only one green or blue color.
3. For the same color, the look might be different. Silver on the Neiman Marcus website is much darker than it is on the Poshmark website.
4. Sometimes, the color definition is not just about the color itself. For example. the silver color at the Neiman Marcus also has a metallic and reflective material feature. So the website is attempting to convey a sense of the surface appearance, including attributes that go beyond the color.

Figure 11: Sample color extraction results

(a) (b)

Figure 12: The color palette on two different fashion websites. (a) the Neiman Marcus website; (b) the Poshmark website.

Also, we find that users generally match the color of the item with the imaginary color associated with the word description first. In other words, they pay less or virtually no attention to the sample color patches that the website provides. Figure 13 shows three different items sold on poshmark.com, and all of which are listed as pink. However only the left one is actually close to the pink color shown at the website. Based on all the reasons above, a perception based color categorization should be developed.

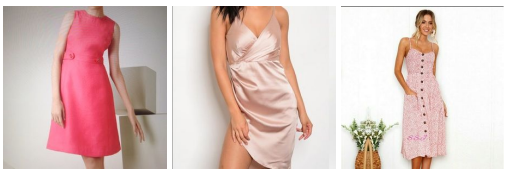


Figure 13: Examples of different pink dresses

The new color matching method first groups all the reference colors into three kinds:

- Neutral colors – black, white, silver, and gray. These color are strongly associated with low chroma appearance, and it is extremely sensitive towards any chroma changes. Therefore, although the lightness of the color can vary, the chroma value of the neutral color is relatively consistent.
- Saturated colors – red, yellow, green, blue, purple, orange, pink. Unlike the neutral colors, saturated colors typically have a very high chroma value, and preserve a certain hue angle. Saturated colors are also more saturated than colors in other categories.
- Off-neutral colors – cream, brown, tan. Off-neutral colors are in between neutral colors and saturated colors. They have a larger variation based on human viewer's interpretation. Therefore, it is harder to classify off-neutral colors as they don't have a certain feature.

Then, we categorize the extracted mean color into its corresponding kinds by the following criteria:

- A color c is a neutral color if its chroma is smaller than 10.
- A color c is a saturated color if its chroma is higher than 20.

Note that the chroma value can be calculated as

$$c* = \sqrt{(a*)^2 + (b*)^2}, \tag{13}$$

where $a*, b*$ are the CIE L*a*b* color coordinates, and hue can be defined as

$$h = \arctan(\frac{b*}{a*}). \tag{14}$$

All the colors whose chroma is between 10 and 20 are the off-neutral color. Therefore, each color only has to compare with its corresponding kind of colors, instead of the entire set of reference colors. For both the off-neutral color and the neutral color set, we use the reference color with the smallest ΔE as final color results. And for saturated colors, in order to accentuate the significance of hue, we use the color with smallest cosine similarity. The new color classification emphasizes the significant of the chroma and hue of the color appearance, instead of the lightness. This helps us get better accuracy by lowering the cross category prediction, where, for example, blue colors are classified as black.

Conclusion

In this paper, we aim to develop an autonomous garment color extraction system to help the user fill in the color information or identify incorrect entries. A three-stage system is introduced: The image segmentation module partitions the given image into smaller perceptually uniform segments. The Grouping module calculates designed nine features to differentiate the garment region from the background region and uses a grouping method to generate a binary mask indicating the garment region. The third module uses a Gaussian Mixture Model to extract the mean color values from the garment region, and finds a proper predefined color description on the fashion retailing website. The proposed system is able to achieve a pixel level segmentation result for fashion product portraits using a semi-supervised learning scheme. Compared with the state-of-art semantic segmentation algorithms powered by deep neural networks, because only clustering algorithms are used in the grouping step, it requires virtually no training data, hence has less training complexity. In addition, we study the color naming system used by online fashion retailers, and propose a simple color classification model to map color coordinates to a verbal description.

Acknowledgments

We thank Poshmark Inc. for their continued support of our research project, and the following people for their contributions and suggestions regarding this work: Kendal G. Norman and Yang Cheng.

References

[1] M. H. Kiapour, X. Han, S. Lazebnik, A. C. Berg, and T. L. Berg, "Where to buy it: Matching street clothing photos in online shops," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 3343–3351.

[2] Z. Li, S. Lin, Y. Cheng, N. Yan, G. Golwala, S. Sundaram, and J. Allebach, "Aesthetics of fashion photographs: Effect on user preferences," in *Imaging and Multimedia Analytics in a Web and Mobile World 2017, (Part of IS&T Electronic Imaging 2017)*, Z. F. J. Allebach and Q. Lin, Eds., vol. 2017, San Francisco, CA, USA, Jan. 2017, pp. 65–69.

[3] G. Csurka, D. Larlus, F. Perronnin, and F. Meylan, "What is a good evaluation measure for semantic segmentation?" in *British Machine Vision Conference*, vol. 27, 2013.

[4] R. Mohan, "Deep Deconvolutional Networks for Scene Parsing," *ArXiv e-prints*, Nov. 2014.

[5] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, and J. G. Rodríguez, "A review on deep learning techniques applied to semantic segmentation," *CoRR*, vol. abs/1704.06857, 2017. [Online]. Available: http://arxiv.org/abs/1704.06857

[6] A. Ess, T. Mueller, H. Grabner, and L. V. Gool, "Segmentation-based urban traffic scene understanding," in *British Machine Vision Conference*, 2009.

[7] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 3354–3361.

[8] Z. Yi, A. Criminisi, J. Shotton, and A. Blake, "Discriminative, semantic segmentation of brain tissue in mr images," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009*, G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 558–565.

[9] L. Ladick, C. Russell, P. Kohli, and P. H. S. Torr, "Associative hierarchical CRFs for object class image segmentation," in *2009 IEEE 12th International Conference on Computer Vision*, Sept 2009, pp. 739–746.

[10] J. Shotton, J. Winn, C. Rother, and A. Criminisi, "Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context," *International Journal of Computer Vision*, vol. 81, no. 1, pp. 2–23, Jan 2009. [Online]. Available: https://doi.org/10.1007/s11263-007-0109-1

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[13] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S.

Bernstein, A. C. Berg, and F. Li, "Imagenet large scale visual recognition challenge," *CoRR*, vol. abs/1409.0575, 2014. [Online]. Available: http://arxiv.org/abs/1409.0575

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. E. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *CoRR*, vol. abs/1409.4842, 2014. [Online]. Available: http://arxiv.org/abs/1409.4842

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, vol. abs/1411.4038, 2014. [Online]. Available: http://arxiv.org/abs/1411.4038

[16] S. Yan, Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Unconstrained fashion landmark detection via hierarchical recurrent transformer networks," in *Proceedings of the 2017 ACM on Multimedia Conference - MM '17*, ser. the 2017 ACM. ACM Press, pp. 172–180.

[17] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool, "Apparel classification with style," in *Proceedings of the 11th Asian Conference on Computer Vision - Volume Part IV*, ser. ACCV'12. Berlin, Heidelberg: Springer-Verlag, 2013, pp. 321–335. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-37447-0_25

[18] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Computer Vision – ECCV 2012*, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 609–623.

[19] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion landmark detection in the wild," vol. pt.II, Cham, Switzerland, 2016//, pp. 229 – 45.

[20] M. Kiapour, K. Yamaguchi, A. Berg, and T. Berg, "Hipster wars: Discovering elements of fashion styles," vol. pt.I, Cham, Switzerland, 2014//, pp. 472 – 88.

[21] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "Neuroaesthetics in fashion: Modeling the perception of fashionability," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, Boston, MA, United states, 2015, pp. 869 – 877.

[22] K. Yamaguchi, T. L. Berg, and L. E. Ortiz, "Chic or social: Visual popularity analysis in online fashion networks," in *Proceedings of the 22Nd ACM International Conference on Multimedia*, ser. MM '14. New York, NY, USA: ACM, 2014, pp. 773–776. [Online]. Available: http://doi.acm.org/10.1145/2647868.2654958

[23] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Los Alamitos, CA, USA, June 2012, pp. 3330 – 7.

[24] K. Yamaguchi, "Parsing clothing in fashion photographs," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, ser. CVPR '12. Washington, DC, USA: IEEE Computer Society, 2012, pp. 3570–3577. [Online]. Available: http://dl.acm.org/citation.cfm?id=2354409.2355126

[25] K. Yamaguchi, M. H. Kiapour, and T. L. Berg, "Paper doll parsing: Retrieving similar styles to parse clothing items," in *2013 IEEE International Conference on Computer Vision*, Dec 2013, pp. 3519–3526.

[26] A. Mojsilovic, "A computational model for color naming and describing color composition of images," *IEEE Transactions on Image Processing*, vol. 14, no. 5, pp. 690–699, May 2005.

[27] E. R. Heider, "Universals in color naming and memory." *Journal of Experimental Psychology*, vol. 93, no. 1, p. 10, 1972.

[28] D. Nickerson and S. M. Newhall, "Central notations for iscc-nbs color names," *J. Opt. Soc. Am.*, vol. 31, no. 9, pp. 587–591, Sep 1941. [Online]. Available: http://www.osapublishing.org/abstract.cfm?URI=josa-31-9-587

[29] S. Tominaga, "A colour-naming method for computer color vision," in *Proceedings of the 1985 IEEE International Conference on Cybernetics and Society*, vol. 573, 1985, p. 577.

[30] T. Belpaeme, "Simulating the formation of color categories," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 1*, ser. IJCAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 393–398. [Online]. Available: http://dl.acm.org/citation.cfm?id=1642090.1642144

[31] L. Pressman, "Fashion color trend report london fashion week autumn/winter 2018," Feb 18AD. [Online]. Available: https://www.pantone.com/fashion-color-trend-report-london-autumn-winter-2018

[32] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[33] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 8. 2274 – 2282, 2012, a previous version of this article was published as a EPFL Technical Report in 2010: http://infoscience.epfl.ch/record/149300. Supplementary material can be found at:

http://ivrg.epfl.ch/research/superpixels.

[34] Y. Wang, *Learning Based Image Analysis with Application in Dietary Assessment and Evaluation*, 2017. [Online]. Available: http://search.proquest.com.ezproxy.rit.edu/docview/1975367006?accountid=108

[35] A. Trmeau and P. Colantoni, "Regions adjacency graph applied to color image segmentation," *IEEE Transactions on Image Processing*, vol. 9, pp. 735–744, 2000.

[36] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000. [Online]. Available: https://doi.org/10.1109/34.868688

[37] J. Wang and J. Allebach, "Automatic assessment of online fashion shopping photo aesthetic quality," in *Proceedings of the 22nd IEEE International Conference on Image Processing (ICIP)*, Sept. 2015, pp. 2915–2919.

[38] Y. P. F. Perazzi, P. Krahenbuhl and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, ser. 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012, pp. 733–740.

[39] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 603–619, 2002.

[40] K. Skjrven. Street photographer's toolbox: Baroque and sinister diagonals. [Online]. Available: https://streetphotographerstoolbox.wordpress.com/2013/01/06/baroque-and-sinister-diagonals/

[41] C. Knight. Fstoppers: The ultimate guide to composition - part one: Just say 'no'keh. [Online]. Available: https://fstoppers.com/architecture/ultimate-guide-composition-part-one-just-say-nokeh-31359

Author Biography

Zhi Li is a fourth year doctoral student and teaching/research assistant in the School of Electrical and Computer Engineering at Purdue University, West Lafayette. He works with Prof. Jan Allebach and Poshmark primarily on fashion photography analysis, including fashion photography aesthetics and the autonomous garment color extraction system. He has also involved multiple research projects such as fashion textural analysis, category/style classification based on CNN and more. Beyond academics, Zhi is an active member of the Purdue University Choir since 2014, and the Eta Kappa Nu (HKN) Beta Chapter since 2016. He is currently serving as the HKN volunteer director.