Learning and Recognition of Clothing Genres From Full-Body Images

Shintami C. Hidayati, Chuang-Wen You, Wen-Huang Cheng, and Kai-Lung Hua

***Abstract*—According to the theory of clothing design, the genres of clothes can be recognized based on a set of visually differentiable style elements, which exhibit salient features of visual appearance and reflect high-level fashion styles for better describing clothing genres. Instead of using less-discriminative low-level features or ambiguous keywords to identify clothing genres, we proposed a novel approach for automatically classifying clothing genres based on the visually differentiable style elements. A set of style elements, that are crucial for recognizing specific visual styles of clothing genres, were identified based on the clothing design theory. In addition, the corresponding salient visual features of each style element were identified and formulated with variables that can be computationally derived with various computer vision algorithms. To evaluate the performance of our algorithm, a dataset containing 3250 full-body shots crawled from popular online stores was built. Recognition results show that our proposed algorithms achieved promising overall precision, recall, and *F*-score of 88.76%, 88.53%, and 88.64% for recognizing upperwear genres, and 88.21%, 88.17%, and 88.19% for recognizing lowerwear genres, respectively. The effectiveness of each style element and its visual features on recognizing clothing genres was demonstrated through a set of experiments involving different sets of style elements or features. In summary, our experimental results demonstrate the effectiveness of the proposed method in clothing genre recognition.**

***Index Terms*—Classification, clothing genre, style element.**

I. Introduction

CLOTHING analysis and understanding is an essential element to enrich context-aware applications, e.g., automatic clothing tagging for online shopping systems [2]. According to the theory of clothing design [3], clothing genres are determined by the combination of various style elements that exhibit consistent and differentiable visual properties. Therefore, in this paper, we propose a novel classification technique to determine the clothing genres through recognizing fundamental style elements of clothing design.

Clothing genre recognition is part of the wider task of visual object recognition [4]–[8]. It poses significant challenges because of the rich garment silhouettes and variations in design features, such as lengths, print styles, and pleats details. Previous studies [9]–[11] described clothing appearance only by image-based low-level features, such as global color histograms and textures, for the purpose of similarity search, without capturing the clothing features of visual appearance to recognize and classify the clothing genres as well. To address these problems, our proposed scheme instead identifies fundamental style elements of clothing design, including collars, front buttons, print styles, and sleeves. By determining discriminative representations of style elements, we can define features for recognizing clothing genres.

Clothes are worn by people in their everyday lives and reflect meaningful contextual information. In this paper, a novel clothing genre classification technique is proposed to identify the genre of clothes worn by people in full-body shots. To infer positions of body parts, our technique first estimates the spatial layout of humans body parts. After that, the location-based features describing style elements are extracted for learning and recognizing the clothing genres in the training and recognition phases, respectively.

The main contributions of this paper can be summarized as follows.

1) We design, implement, and evaluate a clothing genre classification technique to identify popular upperwear and lowerwear genres based on their style elements.
2) We identify a set of computable features, e.g., shapes of printed patterns or color variances, extracted from various parts of clothes to describe the representations of different clothing genres. By formulating clothing as representative features, our system can efficiently classify a cloth into one of the adopted genres of upperwear or lowerwear using machine learning algorithms.
3) We built a large image dataset, with full-body shots crawled from several online shopping stores. By experimenting our system with this dataset, the average precision and recall achieve promising values of more than 80% for classifying the clothing genres of both upperwear and lowerwear. Further, this dataset can also be applied for facilitating future clothing analysis studies.

Manuscript received December 26, 2016; revised March 26, 2017; accepted May 27, 2017. Date of publication June 19, 2017; date of current version April 13, 2018. This work was supported by the Ministry of Science and Technology of Taiwan under Grant MOST104-2221-E-011-091-MY2 and Grant 105-2628-E-001-003-MY3. A preliminary version of this paper was presented at the 20th ACM International Conference on Multimedia [1]. This paper was recommended by Associate Editor D. Goldgof. *(Corresponding author: Kai-Lung Hua.)*

S. C. Hidayati and K.-L. Hua are with the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106, Taiwan (e-mail: d10115802@mail.ntust.edu.tw; hua@mail.ntust.edu.tw).

C.-W. You is with the Intel-NTU Laboratory, National Taiwan University, Taipei 106, Taiwan (e-mail: cwyou@ntu.edu.tw).

W.-H. Cheng is with the Research Center for Information Technology Innovation, Academia Sinica, Taipei 115, Taiwan (e-mail: whcheng@citi.sinica.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2017.2712634

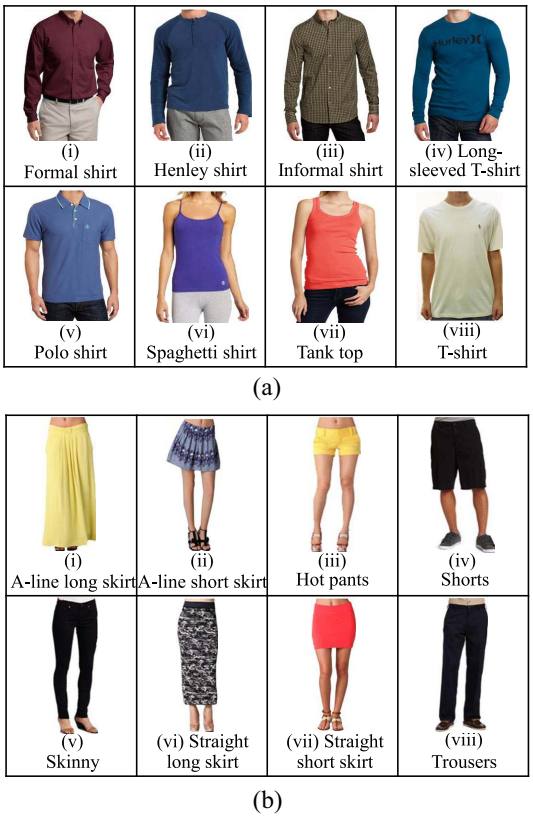

Fig. 1. Defined clothing genres in our proposed framework. (a) Upperwear taxonomy. (b) Lowerwear taxonomy.

The rest of this paper is organized as follows. Section II gives a brief review of the related works. Section III provides overview of the overall system. Section IV introduces the set of features proposed to capture style elements of clothes. Section V describes the learning and classification algorithms of our framework for recognizing clothing genres of full-body shots from online clothing stores. Section VI describes how we prepare the dataset and evaluates the effectiveness of the propose technique. Finally, Section VII concludes this paper.

II. Related Work

There have been many works on clothing recognition. Early work on clothing segmentation attempts to solve the localization of clothing region [12], [13]. In this paper, we introduce clothing genre recognition problem, where the goal is to localize and categorize specific style of clothing items. In the past, clothing retrieval [14], [15] were proposed to address the task of finding similar clothing queried by an input clothing image. To model the discrepancy between user photographs and clothing images from online shopping stores, street-to-shop image retrieval task has been explored [11], [16]. Vittayakorn *et al.* [17] presented an approach to learn outfit similarity on the runway and in the real-world settings based on specific descriptors for color, texture, and shape. The works in [11] and [14]–[17] aim to measure the similarity of clothing images, without capturing specific instances that correspond to the unique meanings reflected by clothing genres. Different from their works, we propose to identify fundamental style elements of clothing genres and formulate their corresponding visual features with variables that can be computationally derived with various computer vision algorithms. By learning the style elements of clothes, our proposed system is able to recognize the genre of clothes as well as to describe their appearance. For example, the description "clothes with collar, short sleeve, and half front button" refers to "polo shirt."

Further, clothing recognition and parsing studies are performed to understand and classify the clothing genres. Zhang *et al.* [18] proposed a clothing recognition system for three kinds of upperwear taken in a well controlled fitting room environment with human poses in frontal view. Yang and Yu [19] presented a video content analysis system to locate human figures, segment the clothing regions, and recognize the category of that clothing. Hidayati *et al.* [20] presented an algorithm that automatically discovers visual style elements for representing the fashion trends. Different with the recognition approaches of aforementioned methods, this paper learns style elements of clothing genres, such as the presence of collar and the sleeve type, to analyze the clothing categories. In the works of fashion parsing, classification is done to the level of pixel or superpixel. Yamaguchi *et al.* [21] proposed to assign a semantic label to each pixel in the image using a retrieval-based approach. Specifically, they used the retrieved similar styles from a large database of tagged fashion images to recognize clothing items in the query. Simo-Serra *et al.* [22] proposed to segment different garments worn by a person using a conditional random field (CRF), which takes into account the dependencies between clothing and human pose. Liang *et al.* [23], [24] are also working on the clothing-related analysis. In their work, they address the human parsing task. Given an input image, they predict every pixel with 18 labels: 1) face; 2) sunglass; 3) hat; 4) scarf; 5) hair; 6) upper-clothes; 7) left-arm; 8) right-arm; 9) belt; 10) pants; 11) left-leg; 12) right-leg; 13) skirt; 14) left-shoe; 15) right-shoe; 16) bag; 17) dress; and 18) null. The proposed work in [23] and [24] are not able to classify the parsed regions into different types. On the other hand, we focus on recognition of different categories.

For recent advances in clothing attributes recognition, Yamaguchi *et al.* [25] and Chen *et al.* [26] proposed style rules model based on a CRF over the combination of clothing attributes. The proposed method in [25], however, does not consider the issue of pose variation. In contrast, we propose to exploit the importance of pose information. Chen *et al.* [26] proposed a set of clothing attributes to describe the upperwear appearances. For an input image, they first perform human pose estimation, and then extract features from the entire torso and arm regions. Although they consider pose variation, however, their method is limited for straight-frontal pose images. Yet, not all images have photographed models with straight-frontal pose. Unlike their method, we consider more practical condition by converting the nonstraight-frontal body pose into a straight-frontal body pose before extracting the corresponding features. Consequently, our proposed method does not have nonstraight-frontal pose limitation. In addition, instead of extracting features from the entire torso and arm regions, we propose that features representing clothing attributes are

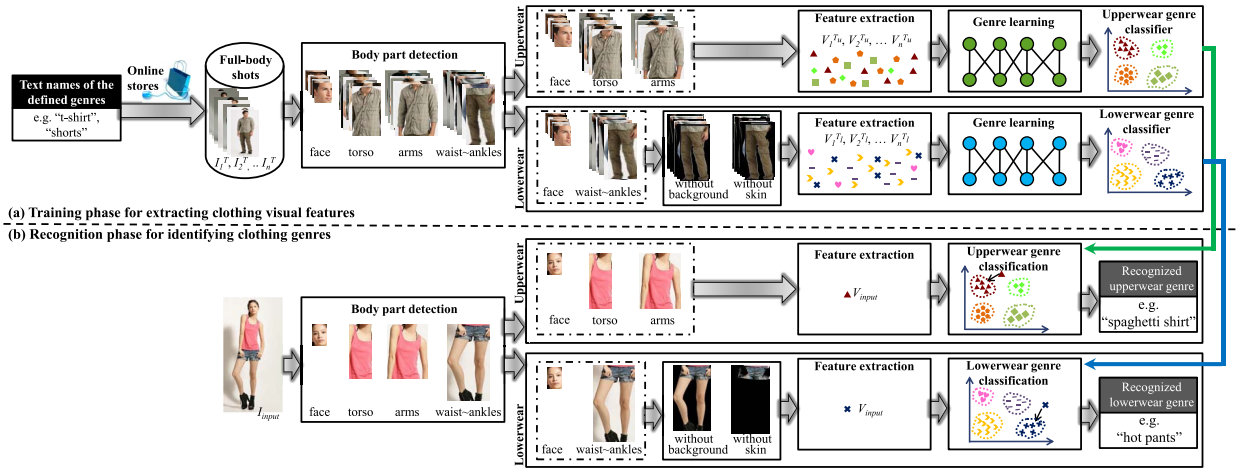

Fig. 2. Overview of our proposed framework for clothing genre classification.

extracted from certain human body parts. For example, it is more reasonable that features for describing collar types are extracted at collar region instead of the entire region of torso. By maintaining the location consistency between feature and human body part, we can identify more discriminative features of clothing attributes based on its consistent representations.

III. System Overview

Fig. 2 shows the overview of the proposed framework that consists of the following components: 1) data collection; 2) body parts identification; 3) feature extraction; and 5) learning and recognition. Having full-body shots crawled from online stores, which photographed models wearing upper- and lower-wear spanning across all interested genres, we first estimate positions of body parts, including the face, torso, arms, and legs. After that, the features describing visual representations of style elements located at known places relative to different human body parts are extracted and then used to prepare a unified feature vectors. Further, we use these feature vectors for learning and recognizing the clothing genres in the training and recognition phases, respectively.

To describe the representations of different clothing genres, a set of computable features are extracted from various parts of clothing region. In each image, the torso and arms are exploited to locate nearby upperwear and extract the thirteen features of upperwear style elements. These features are formulated according to the five visually differentiable upperwear style elements, including collar, print style, shoulder skin, front button, and sleeve types. Similarly, legs are exploited to locate nearby lowerwear and extract the thirteen visually salient features of seven lowerwear style elements, including leg gap, length, print style, side, pleat, wrinkle, and width. The details of the design of corresponding visual features of each style element will be described later in Section IV.

By concatenating the extracted features into a unified feature vector, our framework further learns models to describe upper- and lower-wear genres based on a supervised learning algorithm. Then, based on the learned classifiers, the classification framework recognizes the input feature vectors extracted from the upperwear area (or the lowerwear area) as one of the adopted upperwear (or lowerwear). We will describe the details of our learning and classification phases in Section V.

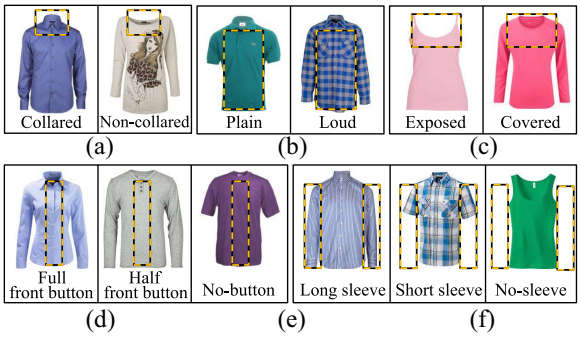

Fig. 3. Visual illustrations of style elements. (a) Collar type. (b) Print style type. (c) Shoulder skin type. (d) Front button type. (e) Sleeve type.

IV. Style Elements of Clothes

According to the theory of clothing design [3], a set of style elements (or semantic elements) of clothes dressed upper and lower bodies, i.e., *upperwear* and *lowerwear*, can be identified to describe clothing genres. The design of corresponding visual features of upper- and lower-wear are described as follows.

A. Style Elements of Upperwear

Table I summarizes the composition of each upperwear genre in terms of style elements (see Fig. 3). Eight most popular upperwear genres [27] are included in our upperwear taxonomy, as shown in Fig. 1(a). The proposed upperwear classification framework classifies the upperwear of each full-body shot into one of the eight genres by examining five types of upperwear style elements as described as follows.

1) Collar Types: A collar is a part of garment that fastens around or frames the neck. Fig. 3(a) illustrates the difference between collared clothes and noncollared ones. Comparing to noncollared clothes, corners of collared clothes appear more frequently and distribute closely in the upper part of the torso area. Therefore, to detect the presence of collar points, a corner detector [28] is exploited to identify corners located within the upper one-third part of torso τ, namely collar region $\Re_c$.

TABLE I
COMPOSITION OF EACH UPPERWEAR GENRE IN TERMS OF FIVE STYLE ELEMENTS

Genre	Collar	Front button	Print style	Shoulder skin	Sleeve
Formal shirt	collared	full front button	plain	covered	long sleeve
Henley shirt	non-collared	half front button	loud or plain	covered	long or short sleeve
Informal shirt	collared	full front button	loud	covered	long or short sleeve
Long-sleeved T-shirt	non-collared	no-button	loud or plain	covered	long sleeve
Polo shirt	collared	half front button	loud or plain	covered	long or short sleeve
Spaghetti shirt	non-collared	no-button	loud or plain	exposed, thin strap	no-sleeve
Tank top	non-collared	no-button	loud or plain	exposed, thick strap	no-sleeve
T-shirt	non-collared	no-button	loud or plain	covered	short sleeve

Having the locations of corners, $\mathcal{C}_c = [x_{c_1}, y_{c_1}; \ldots; x_{c_n}, y_{c_n}]$, where (x_{c_i}, y_{c_i}) represents the normalized location[1] of the ith detected corner and n is the number of detected corners, the first feature $f_{c_1} = (n/\Re_c)$, is proposed. The classification framework classifies upperwear as a collared clothing if the value of f_{c_1} is larger than a threshold determined by a supervised learning algorithm [29] described in the next section. To differentiate collar corners from other pattern corners, e.g., printed patterns, we then calculate variance of all x coordinates of corners, i.e., $\text{var}(x_{c_1}, \ldots, x_{c_n})$ defined as f_{c_2}, and variance of all y coordinates of corners, i.e., $\text{var}(y_{c_1}, \ldots, y_{c_n})$ defined as f_{c_3}. By calculating values of f_{c_2} and f_{c_3}, the classification framework can eliminate cases with larger values of f_{c_2} and f_{c_3} to filter out false alarms caused by patterns.

2) Print Style Types: Fig. 3(b) shows plain and loud print styles of clothes, which determine the pattern complexity of a garment. Because of colorful print styles on clothes, patterned clothes have more corners and colors than plain clothes. Therefore, two kinds of features are extracted from the torso area τ, including: 1) the density of detected corner f_{p_1} and 2) the overall significance of the color variances f_{p_2}.

Given m number of the detected corners in τ, the first feature is defined as the density of detected corner, $f_{p_1} = (m/A_\tau)$, where A_τ is the number of pixels in τ. For the second feature, f_{p_2}, since we are evaluating the colorfulness of τ, the edges caused by wrinkles need to be neglected. Since wrinkles can be identified based on the luminance changes, the Y channel of input image in the YCbCr color space is then ignored. To estimate the number of colors in τ, we evenly partition τ into a 5×5 grid of blocks. In each block b, we detect edges by Canny operator [30] and aggregate the number of pixels that are detected as edges as n_{p_b}. Given two predefined threshold values, T_1 and T_2, if there are a majority of blocks, $N \geq T_2$, exhibiting higher color variances with $n_{p_b} > T_1$, the overall significance of color variances f_{p_2} is estimated by the average number of colors among all blocks using the frequency image method [31]. Otherwise, f_{p_2} is set as zero to indicate an insignificant value of color variances. An upperwear is claimed to be loud printing style when the torso area has both high density of corners and high color variances.

3) Shoulder Skin Types: Fig. 3(c) shows two different skin types of clothes. The exposed shoulder skin type usually appears to be a part of spaghetti shirt and tank top. In contrast, the covered shoulder skin type covers skin in the shoulder area with a part of clothes, e.g., long or short sleeves.

[1]We normalize the coordinates to [0–1] to cope with different resolutions of the testing images.

Based on these observations, we first identify potential straps by detecting edge pixels located on the upper half part of the torso τ. To identify skin color, we quantify the similarity between pixels in τ and pixels extracted from skin sample. A subpart of the detected face is used to obtain the skin sample. We empirically choose a rectangle centered at $c_\mathcal{F} = (x_\mathcal{F}, y_\mathcal{F})$ and expanded the area with its width $w_\mathcal{F}$ (height $h_\mathcal{F}$) equal to $x_\mathcal{F} + 0.7w_\mathcal{F}$ ($y_\mathcal{F} + 0.7h_\mathcal{F}$) outward from the center along the x (y)-axis by an value of $x_\mathcal{F} + 0.35w_\mathcal{F}$ ($y_\mathcal{F} + 0.35h_\mathcal{F}$). After that, we compose a set of skin "seed" points consisting of pixels located in the torso area, whose Euclidean distance in the RGB color space to a skin sample pixel is below 15 distance units [32]. A clustering technique of [33] is then employed to cluster the pixels in τ. For each pixel in τ, the classification framework labels a pixel as skin-color when it belongs to a cluster with pixels of skin seed points. Once we identify all skin-color and nonskin-color pixels, only edges of nonskin-color pixels located on the upper half part of τ are indicated as the straps. The framework identifies upperwear as exposed shoulder skin type when the strap is identified. Otherwise, it classifies upperwear as covered shoulder skin type.

Given the locations of detected straps, three features for this style element are proposed, including: 1) the ratio between the number of skin pixels to the torso area, $f_{e_1} = n_s/A_\tau$; 2) the ratio of the average strap width to the width of the torso area, $f_{e_2} = (w_s/w_\tau)$; and 3) the ratio of the distance between left- and right-most straps to the width of the torso area, $f_{e_3} = (w_{d_s}/w_\tau)$. These ratios jointly guide the classification framework to understand the type of a detected strap.

4) Front Button Types: A front button is a small piece sewn to a piece of clothing to fasten one part of clothing to another. Fig. 3(d) shows three front button types of clothes differentiated based on the length. For shirts with front button(s), we can notice a vertical (or nearly vertical) line, called as a placket.

Identifying small buttons appearing in regions of front button is not easy due to the similarity in color to the fabric. We therefore instead detect the salient visual pattern of a placket in the center of the torso area τ to locate regions of front button types. Furthermore, people within the view range of the camera tend to slightly angle their bodies away from the camera to get flatting full-body images, instead of standing up straight and facing the camera, as shown in Fig. 4(a). To make the line pattern of a placket vertical before performing subsequent analyzing tasks, we rotate the original angled torso area τ' shown in Fig. 4(a) to an rotated torso area τ by θ^o shown

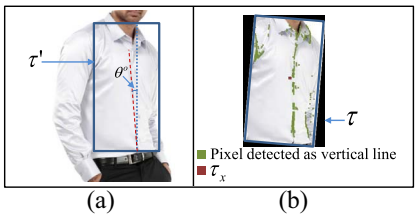

Fig. 4. In (a), the angle between a line formed in shirts with front buttons (i.e., the dotted red line) and a vertical line (i.e., the dotted blue line) is $\theta°$. In (b), the line of the placket (i.e., green pixels located near the center of the rotated torso area) becomes vertical in the rotated torso area τ.

in Fig. 4(b) using the following formula:

$$\theta^o = \arctan\left(\frac{y_{\min(x'_\tau)} - y_{\max(x'_\tau)}}{w_{x'_\tau}}\right) \times \frac{180^o}{\pi} \tag{1}$$

where $y_{\min(x'_\tau)}$ ($y_{\max(x'_\tau)}$) is the y-coordinate of the leftmost (rightmost) point of τ' and $w_{x'_\tau}$ is the width of the rectangle bounding the τ'. By detecting salient visual pattern of a placket, i.e., vertical lines, in the rotated torso area τ using Hough transform method, the detected coordinate of a placket along the x-axis, τ_x, is identified by

$$\tau_x = \frac{\sum_{i=1}^{w_\tau} \mathcal{H}(i) \times i}{\sum_{i=1}^{w_\tau} \mathcal{H}(i)} \tag{2}$$

where $\mathcal{H}(i)$ is the frequency of pixels detected as vertical lines falling within the ith bin along the x-axis. In this paper, the difference between x_i and x_{i+1} is 1 pixel.

Once the framework estimates the potential location of placket, it defines a front button area $\flat$ to approximately bound the placket. We center the front button area $\flat$ at the location of τ_x and set the width of $\flat$ as one-forth of τ since this portion is sufficient to cover all parts of plackets. To determine how the detected pixels on (nearly) vertical lines distribute, four regions are defined: 1) the upper-half part of torso, τ_u; 2) the lower-half part of torso, τ_l; 3) the upper-half part of front button area, $\flat_u$; and 4) the lower-half part of front button area, $\flat_l$. By calculating the ratio of pixels detected as vertical line falling in τ_u, τ_l, $\flat_u$, and $\flat_l$, respectively, the framework can classify upperwear as full front button type if most of pixels detected as vertical line distribute in $\flat_l$ and $\flat_u$, without only concentrating in one area, or as half front button type if most of pixels detected as vertical line mainly aggregate in the $\flat_u$ area. Otherwise, it classifies upperwear as nonbutton type.

5) Sleeve Types: Sleeve is a piece of clothing that covering the arm. Depending on the amount of arm coverage, sleeve can be classified into three types, as shown in Fig. 3(e). The sleeve length is inversely proportional to the amount of arm skin exposed. Through steps similar to those used for skin-color detection described in Section IV-A3, the skin-color pixels within arms are identified based on the color similarity between pixels of people's arms and pixels of subpart area of face $\mathcal{F}$. By measuring the skin-color pixels within arms, i.e., $f_s = (n_\mathcal{A}/A_\tau)$, as a feature, the classification framework can differentiate the sleeve type of upperwear based on thresholds determined by a supervised learning algorithm [29].

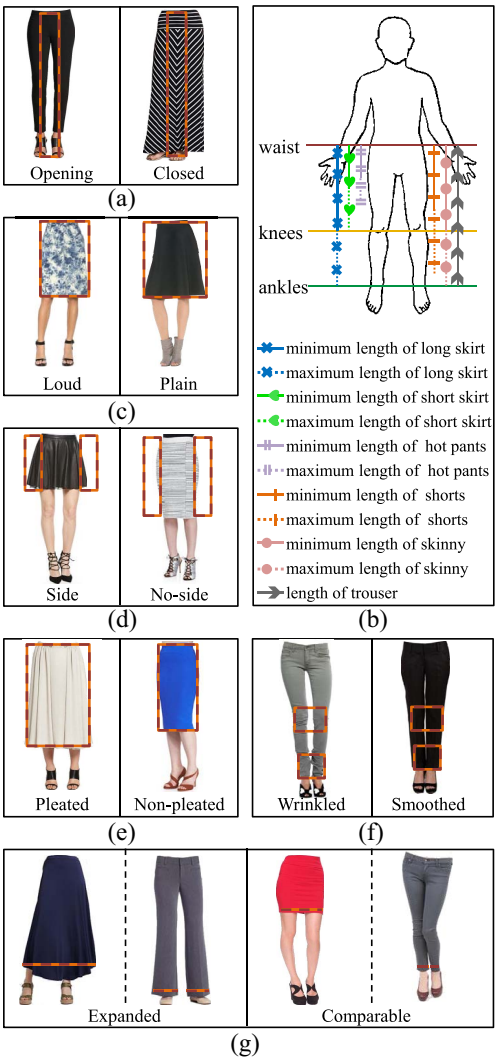

Fig. 5. Differences between seven style elements. (a) Leg gap type. (b) Length type. (c) Print style type. (d) Side type. (e) Pleat type. (f) Wrinkle type. (g) Width type.

B. Style Elements of Lowerwear

Fig. 1(b) shows eight most-popular lowerwear genres [27] that included in our proposed framework. Table II summarizes what composes these lowerwear genres in terms of the differentiable style elements illustrated in Fig. 5. Our lowerwear classification framework can classify each lowerwear in full-body shots as one of eight lowerwear genres.

The style elements of lowerwear in an input full-body shot is detected in three steps: 1) locating the human body part; 2) filtering out skin color pixels; and 3) extracting potential areas, i.e., *lowerwear areas* ℓ, containing pixels of lowerwear. First, a human body part detector [34] is employed to locate the position of legs in each full-body shot, as shown in Fig. 6(b). Second, the skin-color pixels are identified based on the color similarity between pixels of the detected legs to pixels of the skin sample described earlier. To measure color similarity between pixels, the framework calculates Euclidean distance on RGB color space between pixels of the detected legs and pixels of the skin sample using the same technique described in Section IV-A3. Once skin-color pixels are removed, the

TABLE II
COMPOSITION OF EACH LOWERWEAR GENRE IN TERMS OF SEVEN STYLE ELEMENTS

Genre	Leg gap	Length (ends at)	Print style	Side	Pleat	Width ratio (wider lower part)	Wrinkles
A-line long skirt	no	below knees	loud or plain	billowed out	yes	yes	no
A-line short skirt	no	above knees	loud or plain	billowed out	yes	yes	no
Hot pants	yes	above the half part of thighs	loud or plain	not billowed out	no	no	no or yes
Shorts	yes	above ankles and below the half part of thighs	loud or plain	might billow out	no	maybe	no or yes
Skinny	yes	below knees	loud or plain	not billowed out	no	no	yes
Straight long skirt	no	below knees	loud or plain	might billow out	no	no	no
Straight short skirt	no	above knees	loud or plain	might billow out	no	no	no
Trousers	yes	ankles	plain	billowed out	no	maybe	no

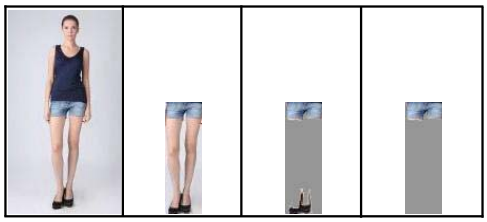

Fig. 6. Results of four lowerwear detection steps, including (a) Input image. (b) Detected body parts, i.e., legs. (c) Lowerwear and footwear areas after performing skin elimination. (d) Final lowerwear area.

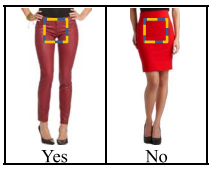

Fig. 7. Crotch areas only appear when people wear pants. There is no crotch area within the lowerwear area when wearing skirts because skirts do not tightly cover legs separately.

framework can extract pixels of lowerwear within regions detected as legs. The extracted pixels of lowerwear, however, might contain the footwear, as shown in Fig. 6(c). Since the footwear area appears in the lowest position, the framework discards the footwear area to filter out pixels in the lowest region boundary. The lowerwear area is located in a location above and directly adjacent to the footwear area and used to detect visually salient features of seven types of lowerwear style elements, which are described as follows.

1) Leg Gap Types: Fig. 5(a) shows two leg gap types, which are: 1) the opening type with a leg gap separating left leg from right leg and 2) the closed leg gap due to a hidden leg gap covered by lowerwear. The leg gap either is completely covered by fabric of some genres of lowerwear, e.g., skirts, or appears with a length that varies with the length of the box bounding the lowerwear. In addition, we notice an area formed by combination of vertical and oblique lines in the upper part of pants as shown in Fig. 7, called as a crotch area, when people are wearing pants and putting their feet together.

Based on these observations, the proposed features used to characterize the leg gap types given lowerwear ℓ are as follows.

1) The direction of the grain line that determines the shape of lowerwear, which is parallel to selvages, f_{g_1}.
2) The ratio between gap's length to the lowerwear's length, $f_{g_2} = (l_{\mathcal{G}}/l_\ell)$.
3) The ratio of pixels detected as vertical and oblique lines inside the area of crotch in lowerwear to the lowerwear area, $f_{g_3} = (n_{\mathcal{C}}/A_\ell)$.

To detect the direction of grain lines, the first feature, f_{g_1}, is extracted using histogram of oriented gradients descriptor [35]. For the cases that people wear tight-fitting lowerwear, e.g., a skinny, but stand with their feet slightly separated, detecting the direction of grain line (or selvage) instead of detecting the obscured gap can help to determine the leg gap type as opening. For the second feature, the length of a leg gap, $l_{\mathcal{G}}$, is identified based on the length of background pixels that separate the left and right legs. Longer pants might have higher feature values of f_{g_2}, whilst the value of f_{g_2} for skirts is zero because the leg gap is covered by lowerwear. And as for the third feature, f_{g_3}, Hough transform method is exploited to identify vertical and oblique lines located inside the region of crotch, which typically occupies the upper one-fourth area of the detected legs. Based on these features, the classification framework can jointly determine the leg gap type of lowerwear.

2) Length Types: Fig. 5(b) shows various length types of the eight lowerwear genres studied in this paper. Depending on the clothing genres, the length of each lowerwear genre, represented by a unique symbol labeled on the arrows, is typically limited by a minimal length, i.e., the length indicated by a solid line, and a maximal length, i.e., the total line length of both solid and dotted lines. For example, the blue line with cross symbols indicated that long skirts hit around the knee but above the ankle, whereas the red lines with heart symbols indicate that short skirts fall above knees. To measure the area of the portion covered by lowerwear, the classification framework calculates the length of lowerwear l_ℓ and the length of left and right legs $l_{\mathcal{L}_l}$ and $l_{\mathcal{L}_r}$. To differentiate the length type of lowerwear, a feature f_l is defined as ratio between the lowerwear length l_ℓ to the maximum of all leg lengths, $f_l = [l_\ell/\max(l_{\mathcal{L}_l}, l_{\mathcal{L}_r})]$. Based on the feature f_l, the classification framework classifies lowerwear as one of length types based on a model trained by a supervised learning algorithm [29].

3) Print Style Types: Fig. 5(c) shows two print style types of lowerwear, which determine the complexity of visual patterns of lowerwear. Among eight lowerwear genres, trousers have plain colors and are patterned with plain or clean-line styles. The print styles of trousers might obviously differ from those of loud print styles of other lowerwear genres as summarized in Table II. Therefore, to characterize the pattern complexity of lowerwear, the framework exploits similar

techniques used to extract the features of upperwear's print style types (Section IV-A2). Two features also are extracted, including: 1) the density of detected corner f_{n_1} and 2) the overall significance of the color variances f_{n_2}. The classification framework classifies lowerwear as having plain print style type if it detects lower values of f_{n_1} and f_{n_2}. Otherwise, it classifies lowerwear as loud print style type.

4) Side Types: The left subfigure of Fig. 5(d) shows two *sides* of loose-fitting lowerwear, e.g., an A-line short skirt. For loose-fitting lowerwear, the loose-fitting bottom billows out, and therefore prolongs the bottom edge of the sides [i.e., the lower edge of the sides within rectangles in left subfigure of Fig. 5(d)]. These prolonged bottom edges of the sides are longer than others of tight-fitting lowerwear, as shown in the right subfigure of Fig. 5(d). Thus, we calculate $f_o = [(w_{\mathcal{L}_l^t} + w_{\mathcal{L}_r^t})/w_{\ell_b}]$ to differentiate the side and nonside types, where $w_{\mathcal{L}_l^t}$ and $w_{\mathcal{L}_r^t}$ are the width of top edge of left and right legs uncovered by lowerwear, respectively, and w_{ℓ_b} is average width of bottom edges of two sides of lowerwear. The classification framework classifies lowerwear as having sides (i.e., the side type) if the value of f_o is greater than a threshold empirically determined through training a model using a supervised learning algorithm [29] described in the next section. Otherwise, the framework classifies lowerwear as having no sides (i.e., the nonside type).

5) Pleat Types: The left subfigure of Fig. 5(e) illustrates a pleated skirt that made by doubling the material upon itself and then pressing or stitching it in place. Comparing with nonpleated skirt as shown in the right subfigure of Fig. 5(e), a pleated lowerwear exhibits (nearly) vertical lines appearing on pleated skirts. To identify the presence of (nearly) vertical lines on pleated skirts, we employ Hough transform method for vertical line detection. The extracted feature is thus the ratio between pixels detected on (nearly) vertical lines in the lowerwear area to the lowerwear area, $f_t = (n_v/A_\ell)$. The classification framework identify ℓ_b as a pleated lowerwear when the value of f_t is greater than a learned threshold. Otherwise, it is classified as nonpleated one.

6) Wrinkle Types: In the left subfigure of Fig. 5(f), the rectangles indicate regions containing *wrinkles*, which are small furrows, ridges, or creases on a normally smooth surface of fabric due to folding. Wrinkles frequently appear in the fabric around knees and ankles of people wearing tight-fitting pants (e.g., skinny). The presence of wrinkles on the lowerwear's surface generates edges with different light intensities. Fig. 5(f) illustrates the differences between wrinkled lowerwear in the left and nonwrinkled lowerwear in the right.

To detect presences of wrinkles, the color values of lowerwear's pixels are first converted to the YCbCr color space. To detect edges with different light intensities, i.e., wrinkles, the change of luminance in the Y channel is measured by Canny operator [30]. Then, the framework calculates the number of pixels, $n_{\mathcal{W}}$, that fall on those identified wrinkles. The extracted feature for detecting the wrinkle type is defined as $f_r = (n_{\mathcal{W}}/A_\ell)$, where A_ℓ is total number of pixels detected in ℓ. The classification framework classifies lowerwear as wrinkled when the value of f_r is greater than a learned threshold. Otherwise, it classifies lowerwear as smoothed.

TABLE III
SUMMARY OF STYLE ELEMENTS OF UPPERWEAR GENRES

Style element (#)	Feature description
1. Collar (3)	• Percentage of corners inside the region of collars • Variance of all x coordinates of corners inside collar region • Variance of all y coordinates of corners inside collar region
2. Print style (2)	• Density of corner inside torso area • The significance of color variances in torso area
3. Shoulder skin (3)	• Ratio between number of skin pixels to torso area • Ratio of the average strap width to the width of torso area • Ratio of the distance between left- and right-most straps to the width of torso area
4. Front button (4)	• Percentage of vertical line in the upper-half part of torso • Percentage of vertical line in the lower-half of torso • Percentage of vertical line in the upper-half of front button area • Percentage of vertical line in the lower-half of front button area
5. Sleeve (1)	• Percentage of skin-color pixels within arms

TABLE IV
SUMMARY OF STYLE ELEMENTS OF LOWERWEAR GENRES

Style element (#)	Feature description
1. Leg gap (3)	• Direction of the grain line, which is parallel to selvages • Percentage of gap's length over lowerwear's length • Percentage of vertical and oblique lines inside crotch area
2. Length (1)	• Ratio of the lowerwear length over the maximum of leg lengths
3. Print style (2)	• Density of corner inside the lowerwear area • The significance of color variances in lowerwear area
4. Side (1)	• Ratio of the width of top edge of left and right legs uncovered by lowerwear over the average width of bottom edges of two sides of lowerwear
5. Pleat (1)	• Percentage of (nearly) vertical lines in lowerwear area
6. Wrinkle (1)	• Percentage of wrinkles in lowerwear area
7. Width (4)	• Ratio of the width of the waist zone to: - 1st part is located at $\frac{2}{9}$ of the lowerwear's length - 2nd part is located at $\frac{4}{9}$ of the lowerwear's length - 3rd part is located at $\frac{6}{9}$ of the lowerwear's length - 4th part is located at $\frac{8}{9}$ of the lowerwear's length

7) Width Types: Fig. 5(g) shows different *width* types. For loose-fitting lowerwear, as shown in the left subfigure of Fig. 5(g), the both sides billow out, and thereby increase the width of its lower part. The resultant width of the lower part of loose-fitting lowerwear is wider than that near the waist zone. In contrast, for tight-fitting lowerwear, as shown in the right subfigure of Fig. 5(g), the width of the lower part is comparable to those of other parts of lowerwear, but is not necessary the widest part of the entire lowerwear area.

To characterize this style element, four features that capture the ratio of the width of the waist zone to four other parts of lowerwear below the waist zone are proposed. These four parts are located at $l_{\ell_1} = (2/9)l_\ell$, $l_{\ell_2} = (4/9)l_\ell$, $l_{\ell_3} = (2/3)l_\ell$, and $l_{\ell_4} = (8/9)l_\ell$ of the lowerwear length l_ℓ. The features to differentiate this style element are then defined as $f_{w_1} = (w_{\ell_1}/w_\omega)$, $f_{w_2} = (w_{\ell_2}/w_\omega)$, $f_{w_3} = (w_{\ell_3}/w_\omega)$, and $f_{w_4} = (w_{\ell_4}/w_\omega)$, where w_ω is width of the waist zone and w_{ℓ_1}, w_{ℓ_2}, w_{ℓ_3}, and w_{ℓ_4} are width of those four parts of lowerwear, respectively. A lowerwear is classified as expanded width type if it has a wider lower part, e.g., features with larger values of f_{w_4}. Otherwise, it has the comparable width type.

V. CLOTHING GENRE LEARNING AND CLASSIFICATION

The proposed system framework consists of two major phases: 1) an *offline training phase* for extracting clothing visual features and 2) an *online recognition phase* for identifying clothing genres. The first phase focuses on building

the upper- and lower-wear classifiers supervised by pairs of labeled clothing genres and corresponding feature vectors with components of the visual features described in the previous section, while the latter phase focuses on recognizing the upper- or lower-wear genres for a given full-body image with the learned classifiers.

In the offline training phase for extracting clothing features, the steps of the processing pipeline are described as follows.

1) *Data Collection:* Collect full-body shots, $I_i^T (i = 1, 2, \ldots, n)$, as described later in Section VI-A. The resultant dataset contains all full-body shots I_i^T along with the upperwear labels L_i^u and lowerwear labels L_i^l.
2) *Face Detection:* Roughly estimate, where the face located in $I_i^T (i = 1, 2, \ldots, n)$ by running [36]. To achieve consistency in extracting features for any size of image, we normalize the size of each image based on the size of the detected face. In this paper, each detected face consists of 45×45 pixels.
3) *Body Part Identification:* Locate body parts, including torso τ, arms $\mathcal{A}_l$ and $\mathcal{A}_r$, and legs $\mathcal{L}_l$ and $\mathcal{L}_r$, in each image using human body part detector of [34].
4) *Feature Extraction:* Extract clothing visual features from each image $I_i^T (i = 1, 2, \ldots, n)$. To avoid interference that might introduced by the hands during feature extraction, we define two kinds of lowerwear regions ℓ: ℓ with a part of the lowerwear area blocked by hands, ℓ_p, and ℓ with hand skin pixels filtered out, ℓ_{-p}. Features depending on the shrunk waist zone, i.e., f_l, f_o, f_{w_1}, f_{w_2}, f_{w_3}, and f_{w_4}, or features depending on the texture information, i.e., f_{g_1}, would be severely damaged by the blocked hands. Therefore, to extract features from images with hands blocked in front of lowerwear, we exploit ℓ_p to extract f_{g_1}, f_l, f_o, f_{w_1}, f_{w_2}, f_{w_3}, and f_{w_4}; while ℓ_{-p} is exploited to extract $f_{g_2}, f_{g_3}, f_{n_1}, f_{n_2}, f_t$, and f_r. Otherwise, we exploit complete lowerwear area ℓ to extract all features of lowerwear style elements. After that, we concatenate the extracted features into a unified feature vector. The feature vector describing upperwear style elements, called *upperwear vector*, is denoted by $V_i^{T_\upsilon} (i = 1, 2, \ldots, n)$; while the feature vector describing lowerwear style elements, called *lowerwear vector*, is denoted by $V_i^{T_\iota} (i = 1, 2, \ldots, n)$.
5) *Supervised Learning:* Build the upper- and lower-wear genre classifiers, respectively, based on normalized upperwear feature vectors $V_i^{T_\upsilon} (i = 1, 2, \ldots, n)$ and normalized lowerwear feature vectors $V_i^{T_\iota} (i = 1, 2, \ldots, n)$. Through a multiclass supervised learning algorithm [29], two upper- and lower-wear classification models can be learned to discriminate input feature vectors extracted from the upper- and lower-wear areas. By learning a discriminant function that measures the correctness of the mapping from an input feature vector to a class label, the classification framework can recognizing clothing genres with the class associated with the largest value of discriminant function.

In the online recognition phase, the upper- and lower-wear genres of an existing image I_i^T from the dataset or other new full-body shots I_i^N are recognized sequentially. Our proposed framework follows the same approaches to recognize face ($\mathcal{F}$), identify body parts (τ, $\mathcal{A}_l$ and $\mathcal{A}_r$, and $\mathcal{L}_l$ and $\mathcal{L}_r$), and extract visual features ($f_{g_1}, f_l, f_o, f_{w_1}, \ldots, f_{w_4}, f_{g_1}, \ldots, f_{g_3}, f_{n_1}, f_{n_2}, f_t$, and f_r). After that, the framework recognizes the upperwear (lowerwear) genre based on the normalized feature vectors of the input image, V^{I_υ} (V^{I_ι}), using upper- and lower-wear genre classifiers obtained from the training stage.

TABLE V
ALIAS(ES) AND THE NUMBER OF IMAGES FOR EACH UPPER- OR LOWER-WEAR GENRE IN OUR DATASET. THE "−" SYMBOL MEANS NO ALIAS AVAILABLE FOR THE GENRE

Upperwear genre	Alias	# of images
Formal shirt	Dress shirt	390
Henley shirt	-	360
Informal shirt	-	470
Long-sleeved T-shirt	Long sleeve T-shirt	400
Polo shirt	Tennis shirt, Golf shirt	440
Spaghetti shirt	Camisole, Cami, Strappy top	340
Tank top	Sleeveless shirt	370
T-shirt	-	480
Total	-	3,250
Lowerwear genre	**Alias**	**# of images**
A-line long skirt	-	420
A-line short skirt	-	340
Hot pants	-	350
Shorts	-	400
Skinny	Skin tight pants, Leggings	470
Straight long skirt	Pencil long skirt	320
Straight short skirt	Pencil short skirt	510
Trouser	-	440
Total	-	3,250

VI. EXPERIMENTS

In this section, we first describe how we prepared the dataset used for training and testing. We then evaluate the recognition performance of our proposed approach and the impact of different style elements and features. After that, we conduct investigations on frame processing time required for recognizing both upper- and lower-wear genres.

A. Dataset

We built a dataset including 3250 full-body shots downloaded from various popular E-commerce websites, i.e., Amazon,[2] Esprit,[3] H&M,[4] Oasis,[5] Rakuten,[6] Uniqlo,[7] and Zara.[8] The categories of upper- and lower-wear in each image scattered approximately equally among eight different upper- and lower-wear genres as shown in Table V.

To automatically crawl a massive number of full-body shots from those E-commerce websites, a Python program was written to retrieve output images by issuing textual queries with names of targeted upper- and lower-wear genres, e.g., "henley shirt" and "skinny." Because some upper- or lower-wear

[2]Amazon.com, http://www.amazon.com/b/ref=topnav_storetab_sl?ie=UTF8&node=71411123011.
[3]Esprit online-shop, http://www.esprit.co.uk/.
[4]H&M, http://www.hm.com/.
[5]Oasis clothing, http://www.oasis-stores.com/.
[6]Rakuten, Inc., http://global.rakuten.com/.
[7]UNIQLO TOP, http://www.uniqlo.com/.
[8]ZARA official website, http://www.zara.com.

genres might not be specified by a consistent name among these websites, our script had to issue multiple queries, each of which is the text of an alternative name used to search one upper- or lower-wear genre, to retrieve images associated with possible alias names in a website. Table V lists all aliases for each genre name aggregated from those websites. Some genre names, e.g., henley shirt or "A-line long skirt," are consistent among websites while others, e.g., "formal skirt," have alternative names, e.g., an alias "dress skirt" for specifying a formal skirt, and not consistent among websites. After we got the full-body shots for both upperwear and lowerwear genres from all websites, we manually selected image with both upperwear and lowerwear appearing in the image belonging to one of the targeted upperwear and lowerwear genres in our framework. To prepare training data for enabling supervised learning and verifying the recognition performance of our framework, we manually labeled each image with the corresponding upper- and lower-wear genres.

B. Experimental Setup

Our proposed framework was developed under Microsoft Windows 7 with MATLAB R2011b. The hardware used to run the experiment was 16 GB RAM, and Intel Core i7-4770 3.40 GHz processor-based PC.

To demonstrate the effectiveness of our proposed method, we compare the performance of our method with two deep learning approaches. We start this experiment from a baseline setting that makes use of the pretrained CaffeNet model. CaffeNet is the reference implementation of Alexnet [37] by Caffe framework [38], trained on 1000 ImageNet categories. The architecture consists of five convolutional layers and three fully connected layers. We use dropout in the fully connected layers, with a dropout probability of 0.5, to avoid overfitting. The last fully connected layer (also known as fc8 layer) has 1000 nodes, each node corresponding to one ImageNet category. A softmax loss layer is implemented as the final layer of network and is used for classification. We fine-tune the weights of the pretrained network on the upper- or lower-wear training set by replacing the last fully connected layer of the pretrained network with the upper- or lower-wear dataset.

To verify the effectiveness of our proposed method, we first compare the recognition performance of our proposed feature design with fine-tuned CNN features. The fine-tuned CNN features are obtained based on the procedures described in [37] and [39]: the training set is forward propagated by the network and the features extracted from the second-last fully connected layer (also known as fc7 layer) in the network are used as feature vectors. Similar to the classification stage of our proposed features, we also perform one-against-one method for multiclass support vector machine (SVM) [29] for the fine-tuned CNN features. The SVM is trained with the radial basis function-kernel, using the penalty parameter C and kernel parameter γ determined by grid search. For each value of C and γ, we conduct tenfold cross validation on the training set, and choose the parameters leading to the highest accuracy.

For the second comparison, we use end-to-end learning of classification through deeper architecture based on [38]. Specifically, the fine-tuned network forward propagates each test image and the output of the network is the final class distribution assigned to the image.

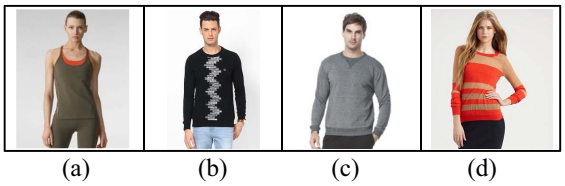

Fig. 8. Major cases confused the upperwear classifier. In these four shots, models might (a) mix and match clothes with different colors, (b) wear a long-sleeved T-shirt with printed style similar with patterns of full front buttons, (c) wear a long-sleeved T-shirt with printed style similar with patterns of half front buttons, and (d) wear a long-sleeved T-shirt with its color similar with skin color.

C. Results

Here, we report our classification performance and the impact of different style element features.

1) Recognition Performance of Upperwear Genres: We compare the upperwear recognition performance of our proposed method, end-to-end deep learning, and deep-learned features + SVM in Table VI. The results are obtained from tenfold cross-validation. In this matrix, each row represents a genre in an actual class and the columns represent the recognized class. Each entry is the total number of clothing genres of the actual class of that row that are recognized as the class of that column. Therefore, the entries on the main diagonal are the total of correct recognitions, and the sum of the entries in each row is the total number of images per category. The recognition results of using our proposed method, end-to-end deep learning, and deep-learned features + SVM are, respectively, shown in the first, second, and third order. On average, upperwear classification by employing our proposed features (end-to-end deep learning and deep-learned features + SVM) achieved overall precision of 88.76% (60.13%, 62.58%), recall of 88.53% (58.29%, 61.16%), and F-score of 88.64% (59.19%, 61.86%).

Fig. 8 selected four major failed cases to discuss why our upperwear classifier did not perform well when recognizing long-sleeved T-shirts and spaghetti shirts. For spaghetti shirt recognition, our system correctly recognized 87.06% of the actual spaghetti shirts but incorrectly recognized 9.71% of the actual spaghetti shirts as tank tops. On the contrary, our system incorrectly recognized 8.11% of the actual tank tops as spaghetti shirts. Because models might mix and match clothes with different colors as shown in Fig. 8(a), they would expose a variable amount of skin near the shoulders, causing confusion of the classifier. Further, for long-sleeved T-shirt recognition, our classifier can recognize long-sleeved T-shirts with an average recall of 87.25%. A part of long-sleeved T-shirts were incorrectly classified as formal shirts, henley shirts, informal shirts, or polo shirts due to print styles similar with patterns of full front buttons appearing on formal and informal shirts as shown in Fig. 8(b), print styles similar with patterns of half front buttons appearing on henley and polo shirts as shown in Fig. 8(c), or skin color appearing on tank top and T-shirts as shown in Fig. 8(d).

TABLE VI
CONFUSION MATRIX OF UPPERWEAR GENRE RECOGNITION. THE RECOGNITION RESULTS OF USING OUR PROPOSED METHOD, END-TO-END DEEP LEARNING, AND DEEP-LEARNED FEATURES + SVM ARE, RESPECTIVELY, SHOWN IN THE FIRST, SECOND, AND THIRD ORDER. THE HIGHEST RECOGNITION VALUE FOR EACH GENRE IS HIGHLIGHTED IN BOLD TYPE

	Formal shirt			Henley shirt			Informal shirt			Long-sleeved T-shirt			Polo shirt			Spaghetti skirt			Tank top			T-shirt		
Formal shirt	**355**	231	265	5	67	55	7	10	5	15	11	6	6	58	50	0	0	3	1	0	2	1	13	4
Henley shirt	1	31	27	**317**	183	202	6	5	5	8	44	35	9	1	2	0	0	0	0	8	11	19	88	78
Informal shirt	11	29	21	5	7	6	**420**	304	300	10	58	53	7	60	69	0	0	3	0	12	14	17	0	4
Long-sleeved T-shirt	2	57	44	13	27	31	8	0	4	**349**	204	233	8	6	6	0	0	0	4	0	0	16	106	82
Polo shirt	5	21	16	16	60	51	6	18	31	9	0	0	**385**	295	295	0	0	2	0	0	4	19	46	41
Spaghetti shirt	0	10	9	0	0	3	0	0	4	6	0	0	0	0	6	**296**	102	113	33	183	171	5	45	34
Tank top	0	6	4	1	3	5	0	0	0	7	0	3	0	0	4	30	110	91	**327**	205	221	5	46	42
T-shirt	1	0	2	10	0	1	7	0	4	12	34	38	10	21	30	0	2	0	10	0	6	**430**	423	399

TABLE VII
CONFUSION MATRIX OF LOWERWEAR GENRE RECOGNITION. THE RECOGNITION RESULTS OF USING OUR PROPOSED METHOD, END-TO-END DEEP LEARNING, AND DEEP-LEARNED FEATURES + SVM ARE, RESPECTIVELY, SHOWN IN THE FIRST, SECOND, AND THIRD ORDER. THE HIGHEST RECOGNITION VALUE FOR EACH GENRE IS HIGHLIGHTED IN BOLD TYPE

	A-line long skirt			A-line short skirt			Hot pants			Shorts			Skinny			Straight long skirt			Straight short skirt			Trousers		
A-line long skirt	**377**	298	305	10	40	40	0	11	19	3	13	4	0	4	12	21	27	14	0	2	6	9	25	20
A-line short skirt	9	37	30	**295**	120	135	7	8	2	9	25	25	0	0	3	0	0	3	19	140	136	1	10	6
Hot pants	0	12	12	0	0	3	**304**	156	158	14	27	30	6	81	88	0	11	11	26	46	42	0	17	6
Shorts	3	7	1	8	1	9	6	4	4	**348**	332	330	5	13	17	0	0	0	13	17	5	17	26	34
Skinny	0	1	13	0	0	14	5	0	4	6	37	32	**418**	395	375	19	0	0	3	12	7	19	25	25
Straight long skirt	23	51	57	0	31	27	0	3	6	1	19	15	6	5	5	**278**	97	106	7	87	87	5	27	17
Straight short skirt	0	40	34	18	0	5	18	6	16	14	12	6	3	3	3	7	13	2	**450**	436	444	0	0	0
Trousers	8	12	1	0	14	5	0	9	0	6	34	20	13	25	11	12	9	4	1	28	23	**400**	309	376

2) Recognition Performance of Lowerwear Genres: Table VII summarizes the performance on recognizing lowerwear genres using our proposed method, end-to-end deep learning, and deep-learned features + SVM. Similar to Table VI, each entry is the total number of clothing genres of the actual class of that row that are recognized as the class of that column. The recognition results of using our proposed method, end-to-end deep learning, and deep-learned features + SVM are, respectively, shown in the first, second, and third order. On average, the proposed lowerwear classification framework using our proposed features (end-to-end deep learning and deep-learned features + SVM) achieved overall precision of 88.21% (66.63%, 69.58%), recall of 88.17% (62.99%, 65.67%), and F-score of 88.19% (64.76%, 67.57%).

From Tables VI and VII, we see that our proposed method obviously outperforms both "end-to-end deep learning" and "deep-learned features + SVM" approaches. These experiments further confirm the contribution of the proposed method.

Fig. 9 selected four exemplar failed cases to understand why the lowerwear classifier performed not well when recognizing A-line short skirts, shorts, and straight long skirts. Because the bottom of an A-line short skirt might not obviously billow out, and hence, its bottom edges were slightly shorter than a typical prolonged skirt bottom of an A-line short skirt or became comparable to those of other tight-fitting lowerwears, e.g., straight short skirts as shown in Fig. 9(a) or hot pants shown in Fig. 9(b), in our training dataset. These tight-fitting, i.e., nonbillowed-out, A-line short skirts tended to make the classifier failed to recognize their correct genre. For shorts, because of the misdetection of skin color pixels in leg areas,

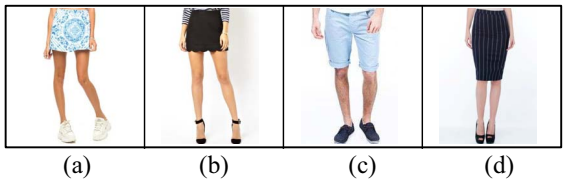

Fig. 9. Major cases confused the lowerwear classifier. In these four shots, a model might wear a tight-fitting A-line short skirt which was incorrectly classified as a (a) straight short skirt or (b) hot pant. (c) Model with the dark shade on his leg worn a short which was incorrectly classified as a trouser. (d) Model worn a straight long skirt with vertical textures which was incorrectly classified as an A-line long skirt.

some skin areas would be detected as belonging to pixels in lowerwear areas. These incorrectly detected lowerwear areas inevitably increased the size of the detected lowerwear area, and therefore confused the classifier to assign incorrect genres with a larger lowerwear area, e.g., trousers as shown in Fig. 9(c). Moreover, for straight long skirts, 13.12% of actual straight long skirts were failed to be recognized. Because of false pleats caused by vertical textures shown in Fig. 9(d), 7.19% of the actual straight long skirts was recognized as A-line long skirts. Further, false skin-color pixels contributed by specific printed patterns, resembling skin of legs, of straight long skirts shrunk the size of detected lowerwear areas, and made 2.19% of the actual straight long skirts miss-classified as straight short skirts.

3) Impact of Different Style Elements on Classification Performance: Tables VIII and IX report the performance of each style element recognition algorithm of our proposed

TABLE VIII
PERFORMANCE OF EACH STYLE ELEMENT OF UPPERWEAR

Style element	Avg. accuracy [%]			Avg. informative value [%]
	Precision	Recall	F-score	
Collar types	91.13	90.05	90.59	19.26
Print style types	95.06	94.94	95.00	13.89
Shoulder skin types	97.54	96.66	97.10	22.49
Front button types	94.96	94.89	94.93	34.40
Sleeve types	95.84	96.05	95.95	9.96

TABLE IX
PERFORMANCE OF EACH STYLE ELEMENT OF LOWERWEAR

Style element	Avg. accuracy [%]			Avg. informative value [%]
	Precision	Recall	F-score	
Leg gap types	97.14	96.85	96.99	62.49
Length types	93.65	91.76	92.69	3.48
Print style types	95.17	94.48	94.83	15.68
Side types	94.34	94.41	94.37	1.05
Pleat types	94.22	93.99	94.10	0.40
Wrinkle types	90.35	90.25	90.30	0.21
Width types	94.14	93.96	94.05	16.69

framework. In this evaluation, we concatenated all features associated with a style element at one time, repeated the experiment, and recorded the value of corresponding precision, recall, and F-score. From Table VIII, we can find that the performance of our shoulder skin types recognition algorithm achieved the highest recognition rates among all style elements of upperwear genres, while collar types recognition algorithm achieved the lowest recognition rates. On average, the classifier can recognize shoulder skin types with precision of 97.54%, recall of 96.66%, and F-score of 97.10%. For collar types recognition algorithm, the classifier achieved the precision of 91.13%, recall of 90.05%, and F-score of 90.59% on average. Moreover, from Table IX, we can find that our leg gap types recognition algorithm achieved the highest performance among all style elements of lowerwear genres with the average precision of 97.14%, recall of 96.85%, and F-score of 96.99%.

To gain knowledge of the importance of each style element in our classification task, we investigated the weight of each feature obtained through the SVM learning procedure [40]. In general, the features with a larger weight have more influence on the classification decision as compared to the features with a smaller weight. Having weight of the jth feature of style element i, $\mathbf{w}_i^j$, obtained from the SVM model, the informative value of style element i is estimated by

$$W(i) = \frac{\sum_j \left|\mathbf{w}_i^j\right|}{\sum_i \sum_j \left|\mathbf{w}_i^j\right|} \times 100\%. \tag{3}$$

The right column of Tables VIII and IX report the average informative value of each style element on, respectively, upperwear and lowerwear recognition. From Table VIII, the rank order, from high to low, of informative value of style elements on upperwear genre recognition is front button, shoulder skin, collar, print style, and sleeve types. While from Table IX, the rank order, from high to low, of informative value of style elements on lowerwear genre recognition is leg gap, width, print style, length, side, pleat, and wrinkle types.

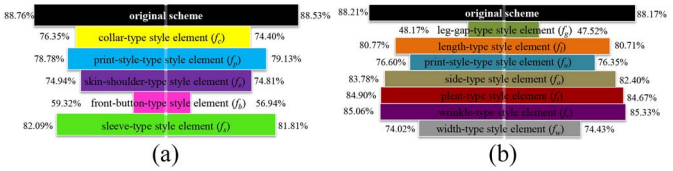

Fig. 10. Impact of style elements. Performance for recognizing (a) upperwear genres and (b) lowerwear genres.

To better understand the impact of our proposed style elements on overall performance, we perform ablation analysis of each style element. Therefore, we compare the proposed upperwear genre recognition algorithm against the original scheme that does not include the following style element: 1) collar-type; 2) print-style-type; 3) skin-shoulder-type; 4) front-button-type; and 5) sleeve-type. We also compare the proposed lowerwear genre recognition method against the original scheme that does not include the following style element: 1) leg-gap-type; 2) length-type; 3) print-style-type; 4) side-type; 5) pleat-type; 6) wrinkle-type; and 7) width-type. The performance comparison between different conditions is reported in Fig. 10(a) for upperwear genres and Fig. 10(b) for lowerwear genres. Precision and recall are shown on the left and right column, respectively. The results suggest that the use of all proposed style elements is useful for describing visual patterns of clothing genres.

Fig. 10(a) shows the performance degradations when removing features associated with each style element when performing upperwear genre recognition. Style element that has the strongest impact is f_b (front button types). As indicated in Fig. 10(a), removing features of this style element significantly degraded the precision by 29.44% to a lower value of 59.32% and the recall significantly degraded by 31.59% to a lower value of 56.94% in comparison with the original scheme. On the contrary, the weakest style element of upperwear is f_s (sleeve types). Removing features of this style element slightly degraded the precision and recall by 6.67% and 6.72% to comparable numbers of 82.09% and 81.81%, respectively.

Fig. 10(b) shows the performance degradations when removing features associated with each style element when performing lowerwear genre recognition. The style element with the strongest impact is f_g (leg gap types), while the style element with the weakest impact is f_r (wrinkle types). Removing f_g features significantly degrades the precision by 40.04% to a value of 48.17% and the recall by 40.65% to a value of 47.52% in comparison with the original scheme. On the contrary, removing f_t features slightly degrades the precision and recall by 3.15% and 2.84% to comparable values of 85.06% and 85.33% in comparison with original scheme.

Based on the results shown in Fig. 10(a) and (b), the performance degradations when removing features associated with a style element are in line with the informative values reported on the right column of Tables VIII and IX,

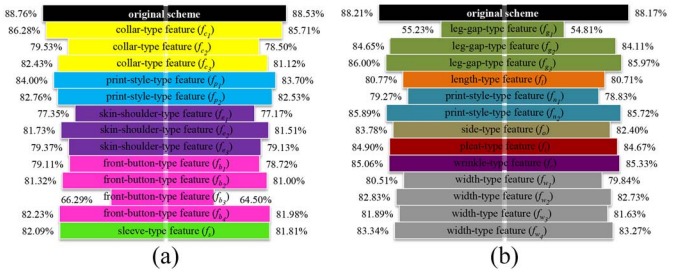

Fig. 11. Impact of features used to characterize a style element. Performance for recognizing (a) upperwear genres and (b) lowerwear genres.

respectively. In other words, the higher informative value of a style element has more influence for the overall performance.

4) Impact of Different Features on Classification Performance: Fig. 11(a) and (b) shows the performance degradation when removing each individual feature employed in our proposed framework, respectively. Similarly, we removed one specific feature at a time, repeated the experiment, and finally recorded the value of corresponding precision, recall, and F-score. The performance degradation was measured by calculating the percentage of degradation as compared to the original scheme that contains the complete features.

As shown in Fig. 11(a), f_{b_3} (front-button-type feature) is the strongest feature while f_{c_1} (collar-type feature) is the weakest feature to describe genres of upperwear. Removing f_{b_3} significantly degrades the precision and recall by 22.47% and 24.03% to lower values of 66.29% and 64.50%, respectively, while removing f_{c_1} slightly degrades the precision and recall by 2.48% and 2.82% to 86.28% and 85.71%, respectively.

Moreover, for lowerwear features as presented in Fig. 11(b), f_{g_1} (leg-gap-type feature) is the strongest feature and f_{g_3} is the weakest feature. Removing f_{g_1} significantly degraded the precision and recall by 32.98% and 33.36% to lower values of 55.23% and 54.81%, respectively, while removing f_{g_3} slightly degraded the precision and recall to a comparable values of 86.00% and 85.97%, respectively. In addition, among all features of the strongest style element f_g as discussed previously, Fig. 11(b) shows that f_{g_1} is the strongest feature with largest degradation.

5) Frame Processing Time: The average total time required for sequentially recognizing the genres of upperwear and lowerwear is 2.63 s. The system spent 16.33%, 37.98%, 29.67%, 15.43%, and 0.59% of the total time to locate the face region, estimate the body parts location, extract the upper- and lower-wear features, and recognize the genres, respectively.

VII. Conclusion

In this paper, we proposed a novel framework for fully automatically recognizing clothing genres, i.e., upper- and lower-wear genres, in unconstrained full-body shots. A set of visual features were defined to identify fundamental upper- and lower-wear style elements adopted from the theory of clothing design [3]. By preparing feature vectors composing of these visual features, our framework can learn models to describe upper- and lower-wear genres based on a supervised learning algorithm and then recognize upper- and lower-wear genres based on the learned classifiers. To validate the proposed framework, a dataset with 3250 full-body shots crawled from major E-commerce websites was prepared. Results show that our proposed algorithms achieved overall precision, recall, and F-score with promising average values of 88.76%, 88.53%, and 88.64% for recognizing upperwear genres, and 88.21%, 88.17%, and 88.19% for recognizing lowerwear genres, respectively. Further, a set of experiments involving different sets of style elements or features was conducted to demonstrate the effectiveness of each style element and its visual features.

We are planning to improve current design and implementation in several future directions. First, current design of feature extraction heavily relies on frontal face detection. By employing other robust face detection algorithm, we hope to extract discriminative features from faces with occlusions and arbitrary viewpoints so that our framework can effectively identify the face of a model who might touch his/her face with hands or face away from the camera in a fashionable pose. Second, we also plan to incorporate more advanced features for better characterizing clothing genres from fashion viewpoints, e.g., measuring the attractiveness of each genres [41], to boost the performance of clothing genre identification.

References

[1] S. C. Hidayati, W.-H. Cheng, and K.-L. Hua, "Clothing genre classification by exploiting the style elements," in *Proc. ACM Int. Conf. Multimedia*, Nara, Japan, 2012, pp. 1137–1140.

[2] I. Daniel, *E-Commerce Get It Right!: Essential Step-by-Step Guide for Selling and Marketing Products Online*. Manchester, U.K.: NeuroDigital, 2011.

[3] L. Svendsen, *Fashion: A Philosophy*. London, U.K.: Reaction Books, 2006.

[4] Z. Lu and H. H. S. Ip, "Spatial Markov kernels for image categorization and annotation," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 4, pp. 976–989, Aug. 2011.

[5] W. Zheng *et al.*, "Strip features for fast object detection," *IEEE Trans. Cybern.*, vol. 43, no. 6, pp. 1898–1912, Dec. 2013.

[6] R. Hong *et al.*, "Image annotation by multiple-instance learning with discriminative feature mapping and selection," *IEEE Trans. Cybern.*, vol. 44, no. 5, pp. 669–680, May 2014.

[7] J. Yu, Y. Rui, Y. Y. Tang, and D. Tao, "High-order distance-based multiview stochastic learning in image classification," *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2431–2442, Dec. 2014.

[8] S. Yan, X. Xu, D. Xu, S. Lin, and X. Li, "Image classification with densely sampled image windows and generalized adaptive multiple kernel learning," *IEEE Trans. Cybern.*, vol. 45, no. 3, pp. 381–390, Mar. 2015.

[9] X. Wang, T. Zhang, D. R. Tretter, and Q. Lin, "Personal clothing retrieval on photo collections by color and attributes," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2035–2045, Dec. 2013.

[10] Y. Kalantidis, L. Kennedy, and L.-J. Li, "Getting the look: Clothing recognition and segmentation for automatic product suggestions in everyday photos," in *Proc. ACM Int. Conf. Multimedia Retrieval*, Dallas, TX, USA, 2013, pp. 105–112.

[11] S. Liu *et al.*, "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 3330–3337.

[12] M. Weber, M. Bauml, and R. Stiefelhagen, "Part-based clothing segmentation for person retrieval," in *Proc. IEEE Int. Conf. Adv. Video Signal Based Surveillance*, Klagenfurt, Austria, 2011, pp. 361–366.

[13] N. Wang and H. Ai, "Who blocks who: Simultaneous clothing segmentation for grouping images," in *Proc. Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 1535–1542.

[14] X. Wang and T. Zhang, "Clothes search in consumer photos via color matching and attribute learning," in *Proc. ACM Int. Conf. Multimedia*, Scottsdale, AZ, USA, 2011, pp. 1353–1356.
[15] M. Mizuochi, A. Kanezaki, and T. Harada, "Clothing retrieval based on local similarity with multiple images," in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 1165–1168.
[16] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 3343–3351.
[17] S. Vittayakorn, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Runway to realway: Visual analysis of fashion," in *Proc. IEEE Win. Conf. Appl. Comput. Vis.*, 2015, pp. 951–958.
[18] W. Zhang, T. Matsumoto, J. Liu, M. Chu, and B. Begole, "An intelligent fitting room using multi-camera perception," in *Proc. Intell. User Interfaces*, 2008, pp. 60–69.
[19] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *Proc. IEEE Int. Conf. Image Process.*, Brussels, Belgium, 2011, pp. 2937–2940.
[20] S. C. Hidayati, K.-L. Hua, W.-H. Cheng, and S.-W. Sun, "What are the fashion trends in New York?" in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 197–200.
[21] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving similar styles to parse clothing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 5, pp. 1028–1040, May 2015.
[22] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, and R. Urtasun, "A high performance CRF model for clothes parsing," in *Proc. Asian Conf. Comp. Vis.*, Singapore, 2014, pp. 64–81.
[23] X. Liang *et al.*, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2402–2414, Dec. 2015.
[24] X. Liang *et al.*, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, 2015, pp. 1386–1394.
[25] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, and Y. Taniguchi, "Mix and match: Joint model for clothing and attribute recognition," in *Proc. Brit. Mach. Vis. Conf.*, 2015, pp. 51.1–51.12.
[26] H. Chen, A. Gallagher, and B. Girod, "Describing clothing by semantic attributes," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, 2012, pp. 609–623.
[27] R. Scapp and B. Seitz, *Fashion Statements: On Style, Appearance, and Reality*. Basingstoke, U.K.: Palgrave Macmillan, 2010.
[28] X.-C. He and N. H. C. Yung, "Corner detector based on global and local curvature properties," *Opt. Eng.*, vol. 47, no. 5, pp. 1–12, 2008.
[29] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.
[30] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, Nov. 1986.
[31] T. Kashiwagi and S. Oe, "Introduction of frequency image and applications," in *Proc. SICE*, Takamatsu, Japan, 2007, pp. 584–591.
[32] J. Kovac, P. Peer, and F. Solina, "Human skin color clustering for face detection," in *Proc. EUROCON*, Ljubljana, Slovenia, 2003, pp. 144–148.
[33] A. Cheddad, D. Mohamad, and A. Manaf, "Exploiting Voronoi diagram properties in face segmentation and feature extraction," *Pattern Recognit.*, vol. 41, no. 12, pp. 3842–3859, 2012.
[34] Y. Yang and D. Ramanan, "Articulated pose estimation with flexible mixtures-of-parts," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, 2011, pp. 1385–1392.
[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, San Diego, CA, USA, 2005, pp. 886–893.
[36] P. A. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.
[37] A. Krizhevsky, I. Sutskever, and G. E. Hintonl, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[38] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM Int. Conf. Multimedia*, Orlando, FL, USA, 2014, pp. 675–678.
[39] P. Sermanet *et al.*, "Overfeat: Integrated recognition, localization and detection using convolutional networks," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–16.
[40] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Mach. Learn.*, vol. 46, no. 1, pp. 389–422, 2002.
[41] J. Fan, W. Yu, and L. Hunter, *Clothing Appearance and Fit: Science and Technology*. Cambridge, U.K.: Woodhead, 2004.

Shintami C. Hidayati received the B.S. degree in informatics from Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia, in 2009, and the M.S. degree in computer science and information engineering from the National Taiwan University of Science and Technology, Taipei, Taiwan, in 2012, where she is currently pursuing the Ph.D. degree.

She was with Institut Teknologi Sepuluh Nopember as a Research Staff Member. Her current research interests include machine learning, data mining and their applications to multimedia analysis, information retrieval, and computer vision.

Chuang-Wen You received the Ph.D. degree in computer science and information engineering from National Taiwan University (NTU), Taipei, Taiwan.

He is an Assistant Research Fellow of the NTU IoX Center, National Taiwan University. His current research interests include mobile computing, human–computer interaction, and ubiquitous computing.

Wen-Huang Cheng received the B.S. and M.S. degrees in computer science and information engineering from National Taiwan University, Taipei, Taiwan, in 2002 and 2004, respectively, and the Ph.D. (Hons.) degree from the Graduate Institute of Networking and Multimedia, National Taiwan University, in 2008.

He was a Principal Researcher with MagicLaboratory, HTC Corporation, Taoyuan, Taiwan, from 2009 to 2010. He is an Associate Research Fellow with the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan, where he is the Founding Leader with the Multimedia Computing Laboratory, CITI, and an Assistant Research Fellow with a joint appointment with the Institute of Information Science. His current research interests include multimedia content analysis, computer vision, mobile multimedia computing, and human–computer interaction.

Dr. Cheng was a recipient of the Numerous Research Awards, including the Prize Award of Multimedia Grand Challenge from the 2014 ACM Multimedia Conference, the K. T. Li Young Researcher Award from the ACM Taipei/Taiwan Chapter in 2014, the Outstanding Young Scholar Awards from the Ministry of Science and Technology in 2014 and 2012, the Outstanding Social Youth of Taipei Municipal in 2014, the Best Reviewer Award from the 2013 Pacific-Rim Conference on Multimedia, and the Best Poster Paper Award from the 2012 International Conference on 3-D Systems and Applications. He supervised his Post-Doctoral Fellows to Award the Academia Sinica Post-Doctoral Fellowship in 2011 and 2013.

Kai-Lung Hua received the B.S. degree in electrical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2000, the M.S. degree in communication engineering from National Chiao Tung University, Hsinchu, in 2002, and the Ph.D. degree from the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, USA, in 2010.

Since 2010, he has been with the National Taiwan University of Science and Technology, Taipei, Taiwan, where he is currently an Associate Professor with the Department of Computer Science and Information Engineering. His current research interests include digital image and video processing, computer vision, and multimedia networking.

Dr. Hua was a recipient of several research awards, including the Second Award of the 2014 ACM Multimedia Grand Challenge, the Best Paper Award of the 2013 IEEE International Symposium on Consumer Electronics, the Best Poster Paper Award of the 2012 International Conference on 3-D Systems and Applications, and the MediaTek Doctoral Fellowship. He is a member of Eta Kappa Nu and Phi Tau Phi.