Pattern Recognition Letters 112 (2018) 212–218

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Which shirt for my first date? Towards a flexible attribute-based fashion query system

Kenan E. Ak[a,b,*], Joo Hwee Lim[b], Jo Yew Tham[c], Ashraf A. Kassim[a]

[a] Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Block E4, Singapore 117583, Singapore
[b] Institute for Infocomm Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, Singapore 138632, Singapore
[c] ESP xMedia Pte. Ltd., 75 Ayer Rajah Crescent, #02-15, Singapore 139953, Singapore

ARTICLE INFO

Article history:
Received 13 October 2017
Available online 17 July 2018

MSC:
41A05
41A10
65D05
65D17

ABSTRACT

In this paper, we present a query & retrieval system specially designed for upper-body clothing products such as Shirts, Jackets etc. To facilitate precise learning, each clothing product image is divided into its constituent parts (e.g. torso and sleeves) to enable the learning of the respective attributes of the parts. The regions associated with the constituent clothing parts are first identified by an algorithm which exploits the structure of the upper-body clothing items. Our proposed "Part Guided" Convolutional Neural Network (CNN) is then "steered" towards different regions according to the attributes and expedite the training. In order to extract compact color information and enable flexible query formulation in a richer space, a Markov Random Field based color encoding method is adopted. A detailed dataset called "Shirt Attributes Dataset" was constructed to evaluate the performance of the proposed system. The proposed system allows more specific query formulations and achieves better results compared to different baselines in terms of classification and retrieval accuracy.

1. Introduction

In recent years, there have been many new innovative developments aimed at enhancing the online shopping and e-commerce experience for customers. However, the ever increasing number of clothing and fashion products makes it more demanding to find preferred products with ease. For consumers, typing something in the search box and scrolling through numerous clothing products (many unrelated) can be an extremely painful and frustrating process. In this work, we focus on this very important issue for the fashion retail business by aiming to improve user search experience through enhanced guidance to find the desired products.

While there have been some works carried out on clothing attribute recognition [1–4] and clothing image retrieval [5–8], only a few recent studies enabled clothing searches with attribute manipulation [9–11]. Many clothing recognition and retrieval studies are not able to describe clothing items well, thus limiting user involvement. Also, there have been no reported works that focus on recognition/retrieval in a rich color space. To address these issues, our proposed system uses constituent parts of clothing products to steer a Convolutional Neural Network (CNN) which enable it to conduct learning in a detailed space. Also, a Markov Random Field based color encoding method is adapted to generate specific color descriptions. Our interactive attribute-based search system allows more specific query formulations and greatly improves the user experience.

The ultimate goal of this work is to develop *an image retrieval system with attribute manipulation capability which is based on the ability to classify each attribute of the constituent parts of upper-body clothing products.* As shown in Fig. 1, the system is able to differentiate different pattern attributes of the torso and sleeve parts for the given query image of a shirt. An image search conducted after the color attribute of the torso is changed from "black" to "white" while maintaining that of the sleeve and collar (at "yellow"), returns the most relevant image with the changed attribute. Instead of using a small number of color attributes, a color-book with rich color descriptors is constructed to summarize the color attributes of each "clothing product" image which can be manipulated to any color as requested.

Inspired by signal processing we use a simple unsupervised jump detection algorithm which exploits the structure of input images to extract the constituent parts. Parts provide useful information about the attributes to facilitate image classification and retrieval by our proposed "Part Guided CNN". Although CNNs have been proven to be quite effective for attribute classification, they may not be able to capture the color information effectively. For

* Corresponding author at: Electrical and Computer Engineering, National University of Singapore, 4 Engineering Drive 3, Block E4, Singapore 117583, Singapore.
E-mail addresses: emir.ak@u.nus.edu, e0013609@u.nus.edu (K.E. Ak).

https://doi.org/10.1016/j.patrec.2018.07.019
0167-8655/

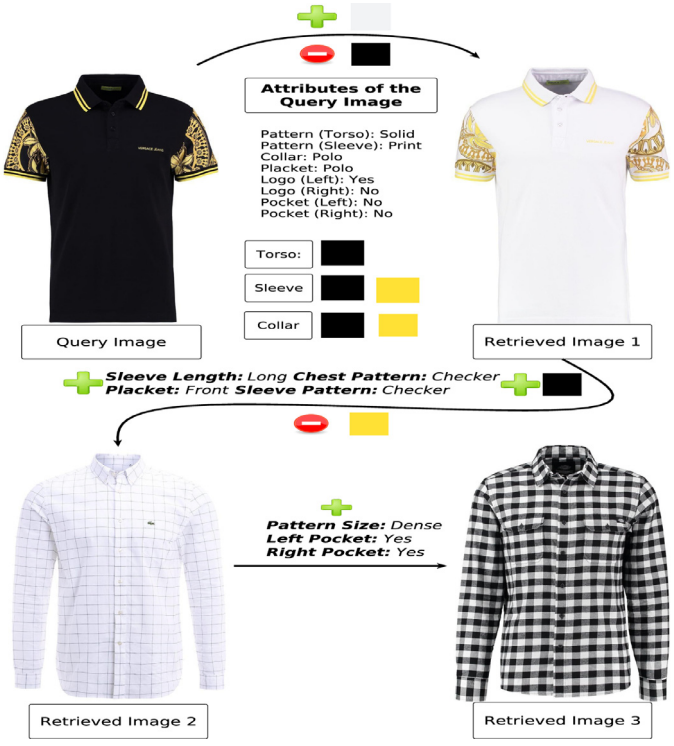

Fig. 1. Overview of the proposed work. Given a query image and a set of attributes are recognized to assist the user. After applying demonstrated attribute manipulations (denoted as "+", "-" for "inclusion" and "exclusion" of attributes respectively) the most similar image is retrieved repeatedly.

example, the color attribute may be classified as "blue" while the true label is "open blue" as it is not possible for all color labels to appear in the ground truth of the dataset. The user interaction is thus restricted by the fixed structure of the labeling scheme. To address this issue, we use MRF-based color encoding to extract more distinct color information which enables classification/retrieval in specific domains. Another advantage of the proposed color encoding method is that wanted/unwanted colors can be adjusted directly which provides greater flexibility to the user.

The main contributions of this paper are summarized as follows:

- "Unsupervised part extraction algorithm" which identifies the constituent parts of input clothing images.
- A new "Part guided CNN" which includes constituent part information and able to produce more discriminative feature for image classification and retrieval.
- A MRF-based color encoding method to learn more specific color information.

2. Related work

2.1. Clothing attribute recognition

Earlier works on clothing attribute recognition [1,2,4,12] are based on handcrafted features such as SIFT [13], HoG [14] and color features rather than learning more abstract representations. The advent of deep learning algorithms led to the emergence of superior methods. Mix and Match [15] uses features extracted from CNN and combines them with a CRF which exploits the fashion rules. Vittayakorn et al. [16] proposed an automatic approach to discover and analyze visual attributes from a noisy collection with the help of attributes and neural activations. Most recently, FashionNet [17] and MTCT [3] networks are proposed which empowered by large data.

2.2. Clothing parsing

Our color encoding was inspired by the clothing parsing approach which was first successfully applied in by Yamaguchi [18]. In [19], a retrieval based approach was applied to the clothing parsing problem and Liu et al. [20] proposed the use of weak color-category labels while Simo et al. proposed a high-performance Conditional Random Field model for clothing parsing. Recently, deep learning based methods have also been applied to fashion parsing [21,22].

2.3. Clothing retrieval

Since 2012, a number of fashion recommendation [7,23] and fashion retrieval [5,6,8,24–26] methods have been proposed. Most of these methods deal with classical fashion search problems while a few have been proposed for image retrieval through manipulating certain attributes [9–11]. AMNet [9] which does "fashion search with attribute manipulation" is able to manipulate clothing products using a memory gate. Compared to AMNet [9], our proposed system places more emphasis on the color attributes as well as other attributes of the constituent parts. While part/region specific learning methods have also been recently proposed [10,11], our method is able to learn attributes of different constituent parts along with more specific color descriptions and uses a much smaller dataset.

Many of the above mentioned methods take advantage of the recent detection and pose-estimation algorithms to segment/partition clothing items. In any case, low picture quality could impact the partitioning of clothing products possibly leaving out some crucial useful information that can affect the learning capability. In this work, we use a smaller dataset comprising real-world commercial clothing products than those used in current deep learning based algorithms. Triplet similarity learning [6] is not used for similarity learning as the number of positive images for each triplet is quite low.

3. Construction of shirt attributes dataset

Among the many available fashion datasets [1,2,5–7,10,17,19,27], only Shopping100k dataset [10] has images with people posing with the clothing items. A drawback of the Shopping100k dataset [10] is the lack of part information as only an overall label for the pattern attribute is given while in reality, the pattern attribute of one constituent part (e.g. torso) may be different from that of the other parts (e.g. sleeves). Given these considerations, we constructed a new dataset with more precise attributes and kept the dataset to a moderate size to make the problem even more challenging.

We started the construction of our dataset with about 20,000 high-resolution images with a clean background under the "men's shirts" category from various websites. Following this, a CNN is trained to select images with a frontal view of the shirts facing the camera with no humans. After eliminating the undesirable images, we were left with around 5000 images. The labeling procedure processes information about the selected images from their respective associated shopping site's meta-data by taking into consideration certain defined key attributes including collar, patterns etc to generate specific attribute label values such as: "collar: spread, pattern: print". To reduce noisy labels, similar labels are merged and labels with a small number of samples are eliminated. In total, there were 9 attribute categories and 39 attribute values. For

Table 1
The number of samples for each attribute in the training data.

Pattern (Torso,Sleeve)	Solid (1188,1200), Grindle (433,435), Horizontal (51, 53), Vertical (149, 150), Plaid (481, 478), Print (282, 272)
Color (Dominant, Major Color for Torso, Collar, Sleeve, Placket, Sleeve end)	Beige (32, 76, 75, 75, 73, 75), White (368, 899, 880, 898, 870, 914), Red (91, 196, 181, 196, 76, 193), Wine (59, 67, 66, 68, 64, 54), Cyan (55, 75, 72, 74, 71, 66), Blue (447, 618, 593, 620, 590, 602), Navy(345, 589, 564, 588, 555, 575) , Green (117, 197, 183, 196, 179, 186), Yellow (51, 75, 66, 71, 69, 71), Orange (57, 87, 80, 85, 80, 80), Pink (126, 147, 145, 146, 143, 139), Purple (66, 68, 68, 67, 66), Gray (132, 199, 199, 198, 185, 203), Brown (64, 86, 77, 85, 78, 77), Black (133, 238, 228, 236, 233, 245)
Collar	Classic (293), Spread (878), Button-Down (1078), Mandarin (116), Polo (213)
Placket	Front (1772), French (435), Fly (73), 1/2 (62), Polo (228)
Sleeve Length	Long (2050), Short (538)
Pocket Presence (Right, Left)	Yes (258, 1242), No (2332,1348)
Logo Presence (Right, Left)	Yes (5, 726), No (2585, 1864)

Fig. 2. Attribute examples of the shirt attributes dataset.

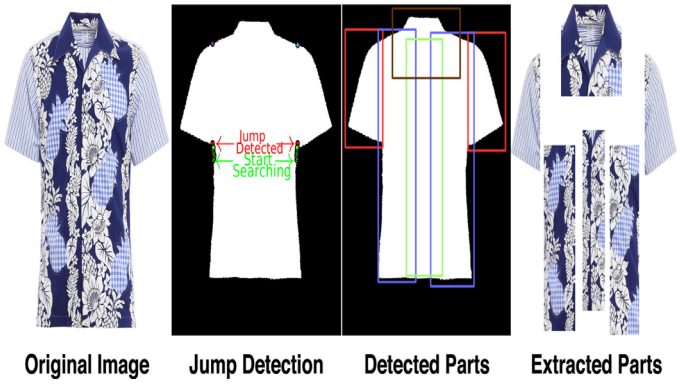

Fig. 3. Unsupervised part extraction algorithm. After the input image is binarized, the discontinuity points between sleeve/torso parts are detected and from these key points, several parts can be extracted.

example, there are 6 possible attribute values for the pattern attribute.

In Fig. 2 some attribute examples are presented and those who are not shown are sleeve length, pocket/logo presence (right-left) and pattern size (sub-attribute of vertical/horizontal/print for torso and sleeve). The value proposition of our dataset is that it provides different attribute values for different parts e.g., major/dominant color attribute labels are specific to each part. Table 1 shows the number of samples for each attribute in the training data. For the training data, we chose 2590 images while the rest of were used for the recognition/retrieval experiments.

4. Recognition, retrieval and manipulation

4.1. Part extraction

The images must first be cropped to the input aspect ratio of the CNN which would inevitably lead to resolution loss. To perform the classification/retrieval tasks in a high-resolution and detailed manner, some part information must be learned by feeding the extracted parts to the network. As it is costly to generate ground-truth of bounding boxes, we exploit the structure of the upper-body clothing items to identify distinctive parts with a straightforward strategy. To begin with, all pictures in the dataset contain a simple background which is detected and removed with a color-based segmentation algorithm. The resulting binary mask (B) is then enhanced by applying dilation and erosion which eliminates the small holes.

After extracting the mask which highlights the clothing item, two different key points are discovered. Using a step detection method which detects discontinuities between torso and sleeve regions. The point where the first discontinuity occurs is assigned to be "Jump Location" and as illustrated in Fig. 3. After finding the left and right jump locations, the sleeve parts are extracted. From jump locations, upper jump locations are also estimated and B is approximately into 5 different pieces with the help of the binary mask. The extracted parts are as follows: Left/Right Torso, Left/Right Sleeves, and Collar. The placket part is also approximately estimated and extracted by using the center of the torso region. The illustration of extracted parts can be seen in Fig. 3.

While for the case given in Fig. 3 there is a visible gap between the sleeves and torso, this gap may not appear in some images. Special requirements are needed to ensure that the distances between 2 jump points and leftmost- rightmost points are sufficiently large. To satisfy this, the detection algorithm uses different starting points until each requirement is satisfied.

4.2. Attribute recognition with part guided CNN

Convolutional Neural Networks turned out to be exceptionally valuable in both attribute recognition and image retrieval. In order to handle multiple attributes, we modify the well-known CNN model AlexNet [28] by removing the last fully connected layer and add a new fully connected layer for each attribute. In order to combine extracted parts (p) with the CNN, following modified softmax function is used where the network considers parts as the prior information. We call this Part Guided CNN as parts guide the network where to look for which set of attributes (a).

$$p(g_{ipa}|x_{ipa}) = \begin{cases} \dfrac{e^{x_{ipa}^T w_a}}{\sum_{k=1}^{K} e^{x_{ipa}^T w_k}} & \text{if } part_p \in att_a \\ 1, & \text{otherwise} \end{cases} \quad (1)$$

In Eq. (1), g_{ipa} denotes the ground truth of the ath attribute of ith image with pth part. Posterior probability is estimated as

shown in Eq. (1) using the modified softmax function. x_{ipa} represents the feature vector, K is the number of corresponding all possible values for g_{ipa}. The predefined prior part knowledge is used in the computation of Eq. (1). If the part is not an element of an attribute (att_a), the probability is set to "1" and loss is "0". The concurrences of parts and attributes are given in the dataset.

$$L = -\sum_{i=1}^{N}\sum_{p=1}^{P+1}\sum_{a=1}^{A} log(p(g_{ipa}|x_{ipa})) \tag{2}$$

The loss function is given in Eq. (2) where N denotes the number of training examples, $(P+1)$ is the number of parts + original image and A denotes the number of attributes. The pre-trained AlexNet model [28] which is trained for the ImageNet challenge is used after removing the last fully connected layer. The network is trained using L where losses of different attribute branches are accumulated. There are some attributes which have several corresponding parts, in order to handle this, probabilities which come from different parts and the original image are averaged. Sleeve length attribute can be an example of this case where the output label is generated by averaging the softmax output of the left/right sleeve parts and the image. The purpose of using the original image along with parts is to correct possible errors from the part extraction method.

4.3. MRF-based Color encoding

While the proposed CNN can solve the image classification problem in a detailed fashion, we need a method which can give a precise color summarization of the input images. In order to output major color labels, the softmax function can be replaced with a sigmoid function or new fully connected layers can be added for each class of color attribute. These solutions are costly and could be error-prone as they might assign "blue" colored image to both similar color attributes such as "blue" and "navy". Unless input images have a plain pattern or a single color, CNN may not be the best option, even then the learned color label might not be the "true" color label. For these reasons, the only dominant color attribute is included in the CNN to learn better feature descriptions. We propose to use MRF to classify major/dominant color attributes and color encoding method to tackle the retrieval task. Different than [20], which proposed to use the color segmentation method for cloth parsing, we aim to use color parsing method to improve image retrieval performance and give more flexibility to the users.

4.3.1. Color recognition with parsing

The first step of is to parse each image into the color attributes as follows:

1) Employ SLIC superpixel algorithm [29] to divide images into non-overlapping patches. Following the usual practice of [20], each patch is assumed to contain a single color.

2) Optimize the following MRF loss function:

$$C^*_{ipj} = \underset{C_{ipj}}{\operatorname{argmin}}\left(\sum_j \left[f_d(\eta, z_{ipj}) + \lambda \sum_{y\in N(j)} f_p(f_d(j), f_d(y))\right]\right) \tag{3}$$

C^*_{ipj} is defined as the color label of the patch j and λ is the weighting factor. The data term (f_d) comes from the response of the multi-class SVM which is trained in a weakly supervised manner where images with the single major color attribute label are chosen. z_{ipj} denotes features calculated from the combination of LAB/RGB mean + LAB/RGB BoW features. The pairwise term (f_p) is defined for the boundaries of neighboring superpixels and η denotes classifier parameters.

After parsing colors, Eqs. (4) and (5) are used to find the dominant and major color attributes. T is the threshold value to accept a specific color attribute as a major color attribute (to get rid of noisy regions). $Color_{ip}$ is a vector denoting if a major color attribute value (n) is available or not, Dom_{ip} is the dominant color and NN denotes the total number of color attributes (=15).

Fig. 4. Given an image, SLIC algorithm [29] is applied first to extract superpixels. The color map is created by using MRF based color parsing algorithm and 3 major color attributes are found (Beige, Black, Gray). The color encoding method is used to find more precise color value/name of each color attribute.

$$\sum_{n=1}^{NN} Color_{ip}(n) = \begin{cases} 1, & \text{if } (\sum_j (C_{ipj} == n)) > T \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$Dom_{ip} = \underset{n}{\operatorname{argmax}}\left(\sum_j (C_{ipj} == n)\right) \tag{5}$$

Given Eqs. (3)–(5), there are some parameters to be decided: λ, T, number of superpixels. Deciding all parameters using a validation set and keeping them constant for all inputs may not be the best way. This is because some images are more complicated than the others. For example, a shirt with a solid pattern is computationally much easier to parse into color components than a striped shirt. In order to make the color parsing method adaptive for each image, parameters of the MRF are decided using the features extracted from the *fc*7th layer (the layer before the last fully connected layer) of CNN. We use a validation set to go through which set of parameters minimize the classification error (Eq. (3)) and (the run-time. In the test time, the most suitable parameters are chosen by comparing *fc*7 features. Moreover, different parts of an image generally share the same color descriptors. However, this may not be true for all images. We fix this issue by training an SVM to decide if different parts of an image are similar to each other using *fc*7 features of parts and color attributes.

4.3.2. Color encoding (CE)

After applying color parsing method, the color information (Color Map) is captured as illustrated in Fig. 4. However, the extracted color components are still not the "true color". Additionally, the following encoding method is proposed in order to encode the image into a more specific color name/value using a color-book with 675 different color labels.[1] Summary of the algorithm is given as follows: (**a**) Employ MRF and calculate RGB/LAB median values from the same color labeled clusters. (**b**) Find the closest candidate from the color book. (**c**) Save the candidate with coverage information (number of superpixels). (**d**) Repeat the same steps for the other color attributes. We assume that the same color labeled clusters correspond to a single color descriptor in the color-book.

[1] http://www.colorhexa.com/color-names.

4.4. Image retrieval and attribute manipulation

Features extracted from the layers of CNNs proved to be quite useful in the similarity search [9,30–32] as they provide decent representations for images. For the retrieval task, we propose to use *fc*7 features along with attribute recognition and color encoding to retrieve the most similar images.

We denote *fc*7 features extracted from the part (p) of the query image (q) as F_{qp}. Attributes are represented as positives and negatives (*pos, neg*) and belong to the query image, (pos^*, neg^*) corresponds to manipulated attributes and come from the user interaction, each candidate image from the retrieval gallery denoted as r.

$$S_{1,q,r}(F_{qp}, F_{rp}, pos, neg, pos^*, neg^*) = \sum_{p=1}^{P+1}\left(\lambda\Big(\sum_{p\in pos} d(F_{qp}, F_{rp})/2 + \sum_{p\in neg} d(F_{qp}, F_{rp})\Big) + (1-\lambda)\Big(\sum_{p\in pos^*} d(F_{qp}, F_{rp})/2 + \sum_{p\in neg^*} d(F_{qp}, F_{rp})\Big)\right) \quad (6)$$

The idea behind Eq. (6) is to compare each part and image with retrieval gallery and give more cost to the mismatched attributes. All attributes are included except major color attributes. Note that, all distances are normalized into the range of (0,1]. We set λ to 0.05 to give more importance to attribute manipulations and if there are no attribute manipulations (direct retrieval), (pos^*, neg^*) becomes "none".

In terms of color similarity, encoded colors of the query are compared with the images in the retrieval gallery by following 3 steps:

1) In Eq. (7), V is RGB/Lab color value and comparison is made between color attributes (xx) of q and color attributes (yy) of r.

$$\{xx^*, yy^*\} = \underset{\{xx,yy\}}{\operatorname{argmin}} ||V_q(xx) - V_r(yy)|| \quad (7)$$

2) Color matching scores are calculated using Eq. (8) where H is the coverage percentage (Percentage of the number of superpixels) which is normalized into the range of (0-0.5].

$$S_{2,q,r} = S_{2,q,r} + (H_q(yy^*) + H_r(xx^*))|V_q(xx^*) - V_r(yy^*)| \quad (8)$$

3) Matched color attributes are removed from the computation and the same process is repeated until no unmatched color attribute is left. If for example, the number of color attributes of r is larger than q, remaining color attributes matched with already matched color attributes of q. In this case, another normalization is applied to $S_{2,\,q,\,r}$.

After q is compared with every image in the retrieval gallery, both S_1 and S_2 scores are normalized and combined as shown in Eq. (9). Sorting S in ascending order gives most similar images with respect to the q.

$$S_q = (S_{1,q} + S_{2,q}) \quad (9)$$

5. Experiments

Implementation details: We finetune AlexNet [28] by removing the last fully connected layer and training it with the given loss function using a learning rate of 0.01. Note that any other model CNN model can be chosen rather than AlexNet [28] and parts can be integrated to become Part Guided CNN. The main reason for AlexNet [28] choice is its accessibility and our aim is here not to use the best CNN model but to show the positive effects of parts. MRF is implemented by using the MATLAB wrapper for graph cuts multi-label energy minimization [33] and the calculation of different parameters such as λ, T, number of superpixels are provided in Section 4.3.

Baselines: As Shirt Attributes Dataset is a new dataset and there are not many methods to handle part specific recognition and retrieval we developed several competing methods for attribute recognition and retrieval as follows:

a) Attribute recognition: (1) A hand-crafted model is developed using HoG and SIFT features and training an SVM for each attribute. (2) The same model as the first one but without using part extraction method. (3) We implement a regular CNN (without parts) and in order to handle attributes of parts, the number of fully connected layers is increased to fit possible attribute-part combinations. For example, in Part Guided CNN, only 1 fully connected layer is used for pattern attribute classification but now 2 fully connected layers are being used for torso and sleeve parts.

b) Image retrieval: 3 different methods are prepared for the ablation studies: (1) The proposed method without color encoding. (2) The proposed method without parts. (3) The proposed method without color encoding and parts. For the experiments without CE, the dominant color attribute is included into Eq. (6) and S_1 is directly used to retrieve images. For the image retrieval after attribute manipulation experiments, we do not have a method without attributes as they are required to change attributes.

Evaluation metrics: For the attribute recognition experiments, we report the classification accuracy for each attribute. For direct retrieval and retrieval after manipulating an attribute, top-k retrieval accuracy is reported. Top-k retrieval considers the scenario of the algorithm finding an item with exactly the same attributes as demanded by the query. If a correct item is found it is assigned to a hit (1) otherwise it is a miss (0).

The retrieval experiments are more challenging than attribute prediction as for retrieval to be successful all attributes should match with the query and refinement of the user. Therefore, it is expected that each attribute prediction experiments would have higher accuracies than image retrieval.

Dataset preparation: For recognition experiments, 2072 images are used for training and testing is conducted on the remaining 518 images. Same 518 images are used as the queries to test image retrieval/manipulation performances while retrieval gallery consists of 3085 images.

5.1. Attribute recognition

As can be seen from the Table 2, when the attribute is easier to learn (Pattern, Sleeve length) classic CNN performs similarly to the proposed system which can be learned from small data. The advantage of the Part Guided CNN classification method shows that it can learn more specific attributes (Collar, Placket) by guiding the network exact location of each attribute. Table 2 also shows the superiority of CNN based method vs handcrafted features + SVM. Neural networks are extremely flexible in the types of data they support and can learn the important features directly from the data structure. Although SVM results are comparable in for some attribute (Pattern, Pocket), it fails on troublesome attributes (Collar, Placket). Another observation is that parts improve both CNN and SVM based methods by a quite large margin.

In terms of color classification, the main problem with the baseline methods come from the confusion of 2 similar colors. For example, when the image only possesses only "red" color, it is likely that SVM might output "red" and "wine" labels which is not the case for MRF. This is the main reason why they perform worse in major color attribute classification and the proposed method is able to deal with this issue.

5.2. Direct retrieval

For the direct retrieval experiments, top-k retrieval accuracy results are given in Fig. 5(a). The best performance is obtained by

Table 2
Attribute classification accuracy comparison of different methods. T: Torso, Sl: Sleeve. The proposed method gives marginal improvements and although in some attributes it is not the best, the difference is quite small.

Attribute	SVM (w/o Parts)	SVM (w/ Parts)	CNN (w/o Parts)	Proposed
Pattern (T, Sl)	85.55, 82.27	87.04, 82.44	87.71, 86.88	**90.54, 89.00**
Collar	54.48	72.64	78.54	**88.42**
Placket	56.78	71.62	83.54	**89.96**
Logo	80.30	80.80	92.29	**93.71**
Pocket	83.12	87.57	86.25	**89.77**
Sleeve Length	88.92	98.38	**99.69**	99.42
Pattern Size (T, Sl)	83.72, 79.19	89.16, 86.42	**93.13, 91.29**	92.86, 90.93
Dominant Color	79.02	79.02	76.46	**80.01**
Major Color	75.38	79.81	X	**83.79**

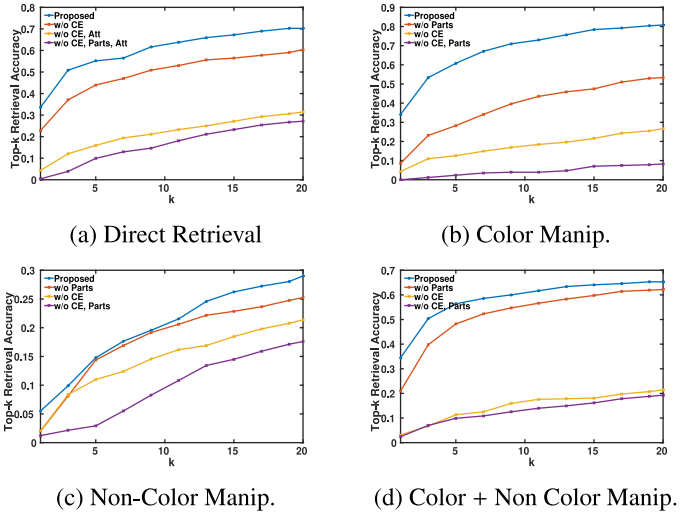

Fig. 5. Top-k Retrieval Accuracies of Retrieval (a) and Retrieval After Manipulation (b,c,d) Tasks.

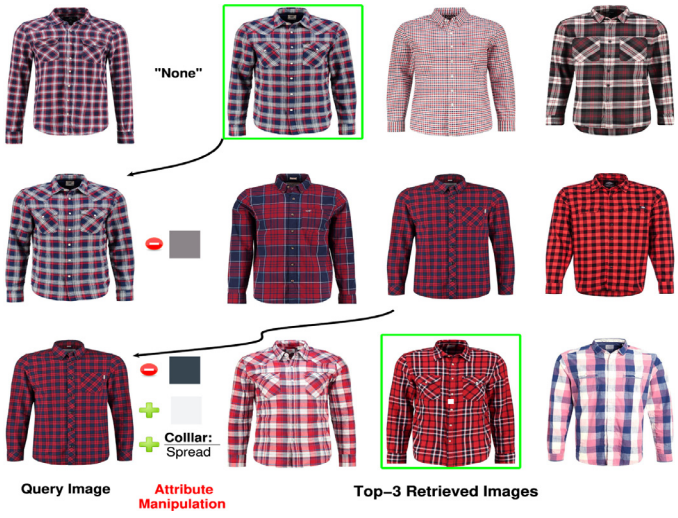

Fig. 6. Given a query image and attribute manipulations (denoted as "+", "-" for "inclusion" and "exclusion" of attributes respectively), top-3 retrieved images of the proposed system are shown. The first column corresponds to query image and second column represents manipulation operations. Each row conducts different searches according to manipulation operations and an image is chosen from top-3 retrieved examples to conduct the search again.

the proposed method with 63.73% top-20 retrieval accuracy. It is directly observed that the main improvement comes from the usage of attributes. Although results without the CE method looks close to the proposed method, visually more feasible results are obtained. The reason for similar accuracies is that it is not quite possible to find an exact match between the query image and the retrieval gallery with a specific color name (using the color-book). For this experiment, the worst performance comes from "w/o CE, Parts, Att" as it directly corresponds to *fc*7 features.

5.3. Retrieval with attribute manipulation

We prepare attribute manipulation experiments by changing a single attribute but in reality, it is possible to make more than one attribute manipulation. 3 query formulations are constructed to prove the robustness of the system, where each formulation investigates a different case. For each query image, an attribute is manipulated such that there is at least one target image in the retrieval gallery which matches all desired attributes. This subsection also shows detail level of our query formulations.

(**1**) The first experiment examines single color attribute manipulation (dominant color) without changing any other attribute and the results are reported in Fig. 5(b). The improvement of the CE method is more marginal than the direct retrieval experiment. Different than the direct retrieval experiment, the target image's color descriptor is annotated by an annotator using the color-book. Color encoding and part information provide more discriminative features for the search and boost the performance. When the MRF-based color encoding method is disabled, a huge performance drop is observed. This is because the exact color information cannot be applied to the query image and target color is assigned to only 15 labels while in the proposed method there are 675 different color labels.

(**2**) The second experiment is similar to direct retrieval case but with a single non-color attribute manipulation. Results are reported in Fig. 5(c). Note that different attribute variations are used to see robustness of the system and average top-k accuracy is given. Again the proposed method performs the best. The reason overall accuracies are lower than the other experiments is some attributes classes of Collar and Placket are quite similar to each other and system have a hard time finding them while keeping desired attributes.

(**3**) In the last experiment, we consider the case of non-color attribute + color attribute manipulation which generally happens when the user wants the same image with different pattern/collar/placket and does not like a specific color. Results of this section can be found in Fig. 5(d). Again it is shown that each method improves overall performance and the proposed system achieves a decent result with 65.30% top-20 retrieval accuracy. The method without parts performs similarly to the proposed method which shows part extraction is not as effective as the color encoding method to handle non-color attribute manipulations.

We report a recurring visual example in Fig. 6 for the image retrieval case. The image with the green boundary means every attribute is matched with the query image and note that there is only 1 exact match in the retrieval gallery for the first and third

query image. The first row corresponds to the direct retrieval case and the system finds the desired image on the first try. In the second row, the user does not like the “gray” color attribute and conducts the search again. For this example, no image is an exact match in the retrieval gallery but still good results are produced. Lastly, the effect of the non-color attribute + color attribute manipulation is observed. Note that the algorithm now puts more importance on the “spread collar attribute” which is available in all 3 retrieved images. The perfect match is captured on the 2nd try. With this example, we prove that even with complicated query inputs, the proposed algorithm can still retrieve images with desired attributes.

6. Conclusion

In this paper, we proposed a flexible attribute-based query & retrieval system that can be applied to upper-body clothing products. Unlike previous approaches, our method considers different constituent parts to enable more precise learning and expedite the training. Moreover, the proposed color encoding method is able to handle specific color attribute manipulations. Our experiments show that the proposed system performs well in both attribute recognition and image retrieval. Going forward, we intend to extend this work to different clothing items such as dresses, pants etc. and extend the number of attributes.

Author declaration

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

References

[1] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, L. Van Gool, Apparel classification with style, in: Asian Conference on Computer Vision, ACCV, 2012, pp. 321–335.
[2] H. Chen, A. Gallagher, B. Girod, Describing clothing by semantic attributes, in: European Conference on Computer Vision, ECCV, 2012, pp. 609–623.
[3] Q. Dong, S. Gong, X. Zhu, Multi-task curriculum transfer deep learning of clothing attributes, in: IEEE Winter Conference on Applications of Computer Vision, WACV, 2017, pp. 520–529.
[4] M. Yang, K. Yu, Real-time clothing recognition in surveillance videos, in: IEEE International Conference on Image Processing, ICIP, 2011, pp. 2937–2940.
[5] M. Hadi Kiapour, X. Han, S. Lazebnik, A.C. Berg, T.L. Berg, Where to buy it: matching street clothing photos in online shops, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2015, pp. 3343–3351.
[6] J. Huang, R.S. Feris, Q. Chen, S. Yan, Cross-domain image retrieval with a dual attribute-aware ranking network, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2015, pp. 1062–1070.
[7] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, S. Yan, Hi, magic closet, tell me what to wear!, in: Proceedings of the 20th ACM International Conference on Multimedia, 2012, pp. 619–628.
[8] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, S. Yan, Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set, in: IEEE Conference Computer Vision Pattern Recognition, CVPR, 2012, pp. 3330–3337.
[9] B. Zhao, J. Feng, X. Wu, S. Yan, Memory-augmented attribute manipulation networks for interactive fashion search, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6156–6164.
[10] K.E. Ak, J.H. Lim, J.Y. Tham, A.A. Kassim, Efficient multi-attribute similarity learning towards attribute-based fashion search, in: IEEE Winter Conference on Applications of Computer Vision, WACV, 2018, pp. 1671–1679.
[11] K.E. Ak, J.H. Lim, J.Y. Tham, A.A. Kassim, Learning attribute representations with localization for flexible fashion search, in: IEEE Conference Computer Vision Pattern Recognition, CVPR, 2018, pp. 7708–7717.
[12] H. Chen, Z.J. Xu, Z.Q. Liu, S.C. Zhu, Composite templates for cloth modeling and sketching, in: IEEE Conference Computer Vision Pattern Recognition, CVPR, 2006, pp. 943–950.
[13] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vision 60 (2) (2004) 91–110.
[14] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conference Computer Vision Pattern Recognition, CVPR, 2005, pp. 886–893.
[15] K. Yamaguchi, T. Okatani, K. Sudo, K. Murasaki, Y. Taniguchi, Mix and match: joint model for clothing and attribute recognition., in: The British Machine Vision Conference, BMVC, 2015. 51–1
[16] S. Vittayakorn, T. Umeda, K. Murasaki, K. Sudo, T. Okatani, K. Yamaguchi, Automatic attribute discovery with neural activations, in: European Conference on Computer Vision, ECCV, 2016.
[17] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: powering robust clothes recognition and retrieval with rich annotations, in: IEEE Conference Computer Vision Pattern Recognition, CVPR, 2016, pp. 1096–1104.
[18] K. Yamaguchi, M.H. Kiapour, L.E. Ortiz, T.L. Berg, Parsing clothing in fashion photographs, in: IEEE Conference Computer Vision Pattern Recognition, CVPR, 2012, pp. 3570–3577.
[19] K. Yamaguchi, M. Hadi Kiapour, T.L. Berg, Paper doll parsing: retrieving similar styles to parse clothing items, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2013, pp. 3519–3526.
[20] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, S. Yan, Fashion parsing with weak color-category labels, IEEE Trans. Multimedia 16 (1) (2014) 253–265.
[21] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, S. Yan, Human parsing with contextualized convolutional neural network, in: Proceedings of the IEEE International Conference on Computer Vision, ICCV, 2015, pp. 1386–1394.
[22] P. Tangseng, Z. Wu, K. Yamaguchi, Looking at outfit to parse clothing, (2017). arXiv:1703.01386.
[23] Y. Kalantidis, L. Kennedy, L.-J. Li, Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos, in: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, 2013, pp. 105–112.
[24] Y. Song, Y. Li, B. Wu, C.Y. Chen, X. Zhang, H. Adam, Learning unified embedding for apparel recognition, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 2243–2246, doi:10.1109/ICCVW.2017.262.
[25] M. Mizuochi, A. Kanezaki, T. Harada, Clothing retrieval based on local similarity with multiple images, in: Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 1165–1168.
[26] E. Simo-Serra, H. Ishikawa, Fashion style in 128 floats: joint ranking and classification using weak data for feature extraction, in: IEEE Conference Computer Vision Pattern Recognition, CVPR, 2016, pp. 298–307.
[27] M.H. Kiapour, K. Yamaguchi, A.C. Berg, T.L. Berg, Hipster wars: discovering elements of fashion styles, in: European Conference on Computer Vision, ECCV, 2014, pp. 472–488.
[28] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[29] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, S. Süsstrunk, Slic superpixels compared to state-of-the-art superpixel methods, IEEE Trans. Pattern Anal. Mach. Intell. 34 (11) (2012) 2274–2282.
[30] A. Babenko, A. Slesarev, A. Chigorin, V. Lempitsky, Neural codes for image retrieval, in: European Conference on Computer Vision, ECCV, 2014, pp. 584–599.
[31] J. Wan, D. Wang, S.C.H. Hoi, P. Wu, J. Zhu, Y. Zhang, J. Li, Deep learning for content-based image retrieval: acomprehensive study, in: Proceedings of the 22nd ACM International Conference on Multimedia, 2014, pp. 157–166.
[32] H. Wang, Y. Cai, Y. Zhang, H. Pan, W. Lv, H. Han, Deep learning for image retrieval: what works and what doesn't, in: IEEE International Conference on Data Mining Workshop, ICDMW, 2015, pp. 1576–1583.
[33] B. Fulkerson, A. Vedaldi, S. Soatto, Class segmentation and object localization with superpixel neighborhoods, in: International Conference on Computer Vision, ICCV, 2009, pp. 670–677.