MVC: A Dataset for View-Invariant Clothing Retrieval and Attribute Prediction

Kuan-Hsien Liu
Academia Sinica
128, Sec. 2, Academia Rd.
Taipei, Taiwan
liukh@iis.sinica.edu.tw

Ting-Yen Chen
Academia Sinica
128, Sec. 2, Academia Rd.
Taipei, Taiwan
timh20022002
@iis.sinica.edu.tw

Chu-Song Chen
Academia Sinica
128, Sec. 2, Academia Rd.
Taipei, Taiwan
song@iis.sinica.edu.tw

ABSTRACT

Clothing retrieval and clothing style recognition are important and practical problems. They have drawn a lot of attention in recent years. However, the clothing photos collected in existing datasets are mostly of front- or near-front view. There are no datasets designed to study the influences of different viewing angles on clothing retrieval performance. To address view-invariant clothing retrieval problem properly, we construct a challenge clothing dataset, called Multi-View Clothing dataset. This dataset not only has four different views for each clothing item, but also provides 264 attributes for describing clothing appearance. We adopt a state-of-the-art deep learning method to present baseline results for the attribute prediction and clothing retrieval performance. We also evaluate the method on a more difficult setting, cross-view exact clothing item retrieval. Our dataset will be made publicly available for further studies towards view-invariant clothing retrieval.

CCS Concepts

•Information systems → Retrieval models and ranking;

Keywords

Clothing retrieval, multi-view clothing, view-invariant clothing retrieval, cross-view clothing retrieval.

1. INTRODUCTION

View invariant object recognition and cross-view object retrieval are important topics in computer vision and content-based image search. However, due to the lack of large-scale datasets, progress toward this direction is still limited. In this paper, we tackle the problem in the clothing retrieval domain. We introduce a dataset containing multi-view clothing images for each item collected in the dataset. This dataset, termed as the Multi-View Clothing (MVC), is organized with a regular setting and abounding attributes. Each item contains four views in the MVC dataset, and the attributes are arranged in a coarse-to-fine hierarchy such as *Women -> Shirts & Tops -> T Shirts -> Colors/Patterns*. With these attributes, one can verify the performance of a multi-view retrieval algorithm, finding the key attributes appropriate for view-invariant recognition, or facilitating the attribute search of clothing images in arbitrary views.

Figure 1: Some sample multiview clothing images from MVC dataset.

Compared to the existing datasets, MVC has a unique feature as it provides multiple views for a clothing item as shown in Figure 1. Besides, the amount of annotated images (161,260) is also large. Most publicly available datasets (eg., Clothing Attribute dataset[2], Colorful-Fashion dataset[12]) have only thousands of images. Apparel Style dataset [1] has more ($> 80,000$), but the image resolution is low. The MVC dataset that contains more images and a well-annotated hierarchy of attributes could be more suitable for learning fundamental feature representations for clothing retrieval.

We evaluate the performance of a recent successful deep CNN model, VGG [19], on the proposed MVC dataset with two different criteria. The first is attributed-based, where the cross-view image retrieval is considered having higher precision score if more attributes are matched. The second is item-based, which deals with a more difficult setting that the retrieval is treated correct if exactly the same item is found. The experimental results serve as baseline performance for further study and comparison.

This paper is organized as follows. In Section 2, we review the related work. In Section 3, the construction of the Multi-View Clothing (MVC) Dataset is introduced. We conduct experiments and provide discussion in Section 4. Finally, the concluding remarks are given in Section 5.

2. RELATED WORK

We first give a brief review on clothing retrieval/recognition approaches. Then, we review publicly available datasets for clothing retrieval and related research.

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912058

2.1 Clothing Retrieval

Clothing related studies can be roughly divided into the following categories: clothing modeling [3, 16], segmentation [7, 20], recognition [21], recommendation [13], people describing [4], classification [1], and retrieval [9, 2, 14, 6, 12]. They are summarized according to the feature types employed as follows.

Formula-Based: One of the formula-based studies is proposed by Chen *et al.* [3], where a context-based And-Or graph is introduced to deal with the wide variability of clothes configurations. Hasan *et al.* [7] built a prior shape model by a Markov Random Field (MRF) formulation for clothing segmentation.

Traditional Features Learning: These methods usually use hand-craft features combined with machine learning methods. Wang *et al.* [20] use clothing shapes to build a Bayesian model to segment clothes. Another similar work [21] adopts gender, age, skin, color and texture to train a linear SVM for clothing recognition. Recently, Chen *et al.* [2] use low-level features such as HOG [5] and SIFT [15] to learn attribute classifiers. Liu *et al.* [14] extract HOG, LBP [17], color moment, color histogram, and skin descriptor features from human parts with a nearest neighbor search to solve a cross-scenario clothing retrieval problem. Di *et al.* [6] also use 5 different low-level features to learn linear SVMs for clothing style recognition and retrieval.

Deep Features Learning: In past years, some deep learning based approaches have been introduced to deal with clothing retrieval. Chen *et al.* [4] built an Online Shopping dataset containing 341,021 images for the problem of describing people based on fine-grained clothing attributes. They designed a double-path deep convolutional neural network for the problem. Lin *et al.* [11] use the AlexNet [10] with an additional latent layer to learn effective hash bits, and perform a hierarchical deep search on a large dataset containing 161,234 images for clothing retrieval.

Although the above two datasets are large, they are not publicly available and thus the results cannot be further refined or compared by others. In following section, we review some publicly available datasets.

2.2 Datasets Publicly Available

Most datasets focus on the problem of clothing category or attribute classification and clothing retrieval.

Clothing Attribute dataset [2] constructed by Chen *et al.* [2] is designed for the attribute classification. This dataset contains only 1,856 images and 26 attributes in total. The image size for most clothes is around 260×400 to 500×750 pixels, where the resolution is somewhat insufficient for analyzing detailed information.

Colorful-Fashion dataset [12] is a good source for fashion data analysis. It consists of 2,682 images in total, and all the pixels in the images are annotated with color and category labels, where 13 colors and 23 categories are involved. The image resolution of this dataset is 400×600 pixels.

Apparel Style dataset [1] is another publicly available dataset used in two tasks: clothing type classification and attribute detection. For type classification, the authors collected a dataset defining 15 clothing types and consisting of over 80,000 images. For atribute detection, another dataset with 25,002 images is constructed with 78 attributes. The image resolution for most clothes is lower than 200×300 pixels. Although the size of this dataset is much larger than previous two datasets, its resolution is the lowest. We summarized some information of these three dataset in Table 1.

Table 1: Comparison of clothing datasets

Datasets	images	attributes	resolution
Clothing Attribute [2]	1,856	26	$\leq 500 \times 750$
Apparel Style [1]	25,002	78	$\leq 200 \times 300$
Colorful-Fashion [12]	2,682	36	400×600
MVC [ours]	161,638	264	1920×2240

In this paper, we introduce a dataset having more images and attributes of a higher resolution, and it supports view-invariant retrieval that is overlooked by previous studies.

3. MULTI-VIEW CLOTHING DATASET

In this section, we introduce the MVC dataset, which is made publicly available[1]. Existing clothing datasets do not take different viewing directions of clothes into account; they often have front views only, which are unsuitable for evaluating cross-view clothing classification and retrieval tasks. We introduce a new dataset containing four viewing angles for each clothing item.

3.1 Data Collection

We collect the MVC dataset by crawling images from several online shopping websites, such as Amazon.com, Zappos.com or Shopbop.com. The challenge in constructing the dataset is to gather complete four different views (front, back, left, and right views) for each clothing item, as there may be only two or three views available for some clothes.

In current stage, the MVC dataset consists of 37,499 items and 161,638 clothing images, where most items have at least four views. Most of the image resolutions are 1920×2240 pixels, which thus offers sufficient details for various clothing related studies such as clothing attribute localization and type classification. Some sample images with four different viewing directions are shown in Figure 1.

3.2 Clothing Attributes and Categories

To measure the clothing retrieval accuracy, we propose to use clothing attributes to establish the relevance between two images. We collect the ground truth attributes from the websites and manually select 264 attributes for similarity evaluation.

These 264 attributes are organized into a three-layer hierarchy. The first layer enforces the gender of clothes, which contains two attributes, Men and Women. There are eight categories for Men's clothes and nine categories for Women's clothes, where most of them overlap, as shown in Figure 2. The third layer contains more detailed attributes. Eg., in the branch "Women->Shirts & Tops", there are type, color, pattern and style attributes. Type includes Blouses, Button Up Shirts, T Shirts and Tank Tops. Color involves blue, brown, green, ... etc. Horizontal stripes, floral print are some attributes belonged to Pattern. Style contains attributes like short-sleeve, round-neckline, long-dress.

The attributes in the first two layers are disjoint, i.e, each image belongs to a single attribute in the first and second layers, respectively. The third layer is non-disjoint, where an image could have multiple attributes in this layer. Combining the three layers forms a multi-label dataset. Some basic statistics of MVC are summarized in Table 1.

[1]http://mvc-datasets.github.io/MVC/

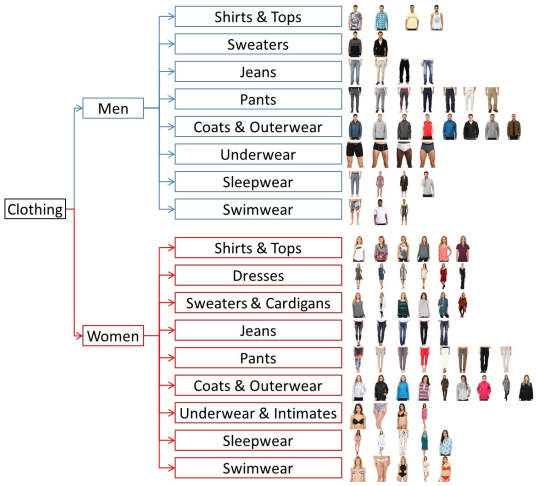


Figure 2: The clothing categories of the first two layers of the MVC dataset with some sample images.

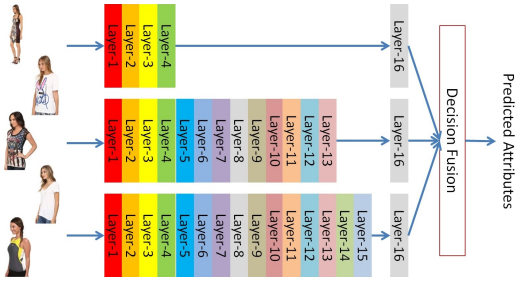


Figure 3: The proposed cross-view clothing retrieval approach. The outputs of three layers of VGG model are late-fused for attribute prediction.

4. EXPERIMENTS

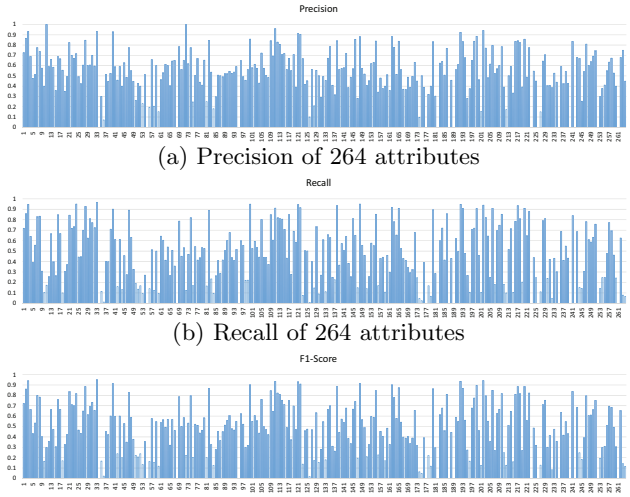
In this section, we first describe our method and experimental settings, and then show experimental results on attribute prediction, clothing retrieval and exact match.

4.1 Method

Considering the recent success of deep learning approaches, we choose a popular deep CNN model, VGG [19], for clothing retrieval. In our experiment, the 16-layer VGG model is selected because it offers similar performance as the 19-layer VGG model but has fewer parameters.

The 16-layer VGG model pre-trained on the ImageNet is fine-tuned on the MVC datasets by Caffe [8] for better fitting the multi-view clothing domain. The soft-max loss function (adopted for single-label classification) in the original VGG model is modified to the sigmoid cross-entropy loss for multi-label training on our problem. The final layer of the VGG model is replaced with 264 outputs.

A deep model consists of multiple feature extractors in different layers. Generally, lower layers extract features preserving local characteristics, and higher layers capture more global features with semantic information. Our approach fuses different layers to predict the clothes attributes. As it is time consuming to evaluate the performance of all layers, we simply choose three layers of VGG-16: 4-th (conv3-128), 13-th (conv3-512), and 15-th layers (FC-4096) in this work. To avoid the high dimensionality of the feature space, we choose not to fuse the extracted features directly, but late-fuse the attributes predicted by the three networks, as shown in Figure 3. The final prediction is then determined by a decision-fusion rule that is introduced below.



(a) Precision of 264 attributes

(b) Recall of 264 attributes

(c) F_1-score of 264 attributes

Figure 4: The precision, recall, and F_1-score of 264 attributes for the fusion system.

4.2 Experimental setting

We conduct several experiments on MVC dataset with the following setting. First, we randomly select one image for each clothing item to form the training set. We then randomly select another image of a different view-angle from the rest images of each clothing item to built the validation set. Finally, all the remaining images serve as the test set.

The validation set is used in deep CNN training to choose a stopping time on the Caffe package [8]. It is also used for building the decision-fusion rule. First, we get three attribute-prediction results via the three deep networks, respectively, based on the validation set; three F_1-scores are accordingly computed per attribute. Then, for each attribute, the network with the highest F_1-score on the validation set is chosen as the winner for this attribute on the test set. Eg., if attribute 1 has F_1-scores 0.55, 0.68, 0.47 with L4, L13, L15 networks, respectively on the validation set, then the L13 prediction results are chosen for attribute 1 in the test stage.

4.3 Results

Attribute recognition. We use *precision*, *recall*, and F_1-*score* to evaluate the attribute recognition performance. Because each image is predicted as being with/without each of the 264 attribute labels, the prediction accuracy per attribute can be computed by comparing the predicted labels with ground truth labels, as shown in Figure 4.

Among them, 35 attributes are predicted over 80% accuracy (in F_1-score), such as Bras, Snow Pants, Short Sleeves. They can be termed as the clothing attributes easier to recognize; 93 attributes are within the accuracy between 50% and 80%, including V-neck, Fringe, Scalloped, and so on. These attributes are termed with middle-level difficulty. 136 attributes are recognized below the accuracy 50%, such as Ultra Low Rise, Unlined, Wide Leg, referred to as the difficult attributes in MVC dataset. The results may be explained as follows. If the attributes (e.g., Short Sleeves) clearly show their distinction on clothing images for human eyes, they would be recognized better than those (e.g., Unlined) that are uneasy to be examined by human eyes.

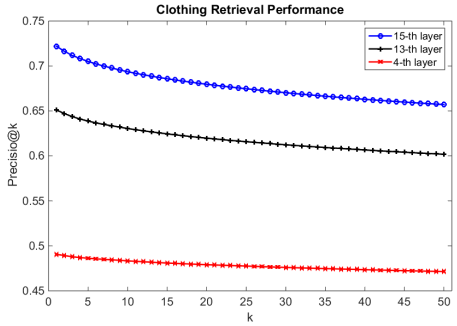


Figure 5: *precision@k* **curve of clothing retrieval.**

Table 2: Accuracy for top-1, top-5, top-10 and top-20 exact clothing match

Method	top-1	top-5	top-10	top-20
Layer-4	0.32 %	1.28 %	2.20 %	3.57 %
Layer-13	2.96 %	8.72 %	12.92 %	18.46 %
Layer-15	5.28 %	13.95 %	19.70 %	26.78 %
Fusion	5.31 %	14.01 %	19.74 %	26.81 %

Clothing retrieval. We treat each network's outputs as a feature vector of 264 dimension, and perform retrieval by using Euclidean distance in the feature space. We investigate the impact of different layers of the network in this experiment as well.

The retrieval performance is measured by the multi-attribute precision calculation suggested in [18]. Given a test image t, the retrieval procedure will assign a rank to all images in the dataset (here, the training set). A top-k retrieval precision with respect to a test image t is,

$$Precision(k) = \frac{\sum_{i=1}^{k} R(i)}{N}, \quad (1)$$

where $R(i)$ is the relevance between t and the i^{th} ranked image, and N is a constant making the precision score 1. Details can be found in [18]. Figure 5 shows the retrieval results with different networks. It can be observed that higher layers perform more favorably than lower layers for all k.

These results provide baseline performance for future studies and comparisons on the cross-view clothes retrieval problem introduced in this paper.

Exact match. The exact clothing match considers a retrieval result correct if one of the item in the top-k retrieved images is the same as that of the query image. We list top-1, top-5, top-10 and top-20 exact match results in Table 2. As can be seen, higher layers perform better than lower layers in this setting too, and the fusion system offers the highest matching accuracy. The results also reveal that finding exactly matched clothing item is more difficult than retrieving similar items.

5. CONCLUSIONS

We introduce a new dataset *MVC* with abundant attributes, and investigate a practically challenging problem, view-invariant clothing retrieval/match. *MVC* has 37,499 items and 161,638 clothing images with 264 attributes. We present a modified VGG-16 deep CNN network for multi-labels prediction to evaluate the performance. Our solution serves as a baseline in attribute recognition and cross-view clothing retrieval. In the future, we plan to enrich the dataset regarding *attribute*, *size*, *view* and *application*, and offer more results on the problem.

6. REFERENCES

[1] L. Bossard, M. Dantone, C. Leistner, C. Wengert, T. Quack, and L. Van Gool. Apparel classification with style. In *Computer Vision–ACCV 2012*, pages 321–335. Springer, 2013.

[2] H. Chen, A. Gallagher, and B. Girod. Describing clothing by semantic attributes. In *Computer Vision–ECCV 2012*, pages 609–623. Springer, 2012.

[3] H. Chen, Z. J. Xu, Z. Q. Liu, and S. C. Zhu. Composite templates for cloth modeling and sketching. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 943–950. IEEE, 2006.

[4] Q. Chen, J. Huang, R. Feris, L. M. Brown, J. Dong, and S. Yan. Deep domain adaptation for describing people based on fine-grained clothing attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5315–5324, 2015.

[5] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[6] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, and N. Sundaresan. Style finder: Fine-grained clothing style detection and retrieval. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 8–13. IEEE, 2013.

[7] B. Hasan and D. Hogg. Segmentation using deformable spatial priors with application to clothing. In *BMVC*, pages 1–11, 2010.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[9] I. King and T. K. Lau. A feature-based image retrieval database for the fashion, textile, and clothing industry in hong kong. In *Proc. of International Symposium Multi-Technology Information Processing*, volume 96, pages 233–240, 1996.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[11] K. Lin, H.-F. Yang, K.-H. Liu, J.-H. Hsiao, and C.-S. Chen. Rapid clothing retrieval via deep learning of binary codes and hierarchical search. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 499–502. ACM, 2015.

[12] S. Liu, J. Feng, C. Domokos, H. Xu, J. Huang, Z. Hu, and S. Yan. Fashion parsing with weak color-category labels. *Multimedia, IEEE Transactions on*, 16(1):253–265, 2014.

[13] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, and S. Yan. Hi, magic closet, tell me what to wear! In *Proceedings of the 20th ACM international conference on Multimedia*, pages 619–628. ACM, 2012.

[14] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, and S. Yan. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3330–3337. IEEE, 2012.

[15] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[16] H. N. Ng and R. L. Grimsdale. Computer graphics techniques for modeling cloth. *Computer Graphics and Applications, IEEE*, 16(5):28–41, 1996.

[17] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7):971–987, 2002.

[18] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 801–808. IEEE, 2011.

[19] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] N. Wang and H. Ai. Who blocks who: Simultaneous clothing segmentation for grouping images. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1535–1542. IEEE, 2011.

[21] M. Yang and K. Yu. Real-time clothing recognition in surveillance videos. In *Image Processing (ICIP), 18th IEEE International Conference on*, pages 2937–2940. IEEE, 2011.