Interactive Clothing Retrieval System

Jia-Lin Chen, Wan-Yu Chen, I-Kuei Chen, Chung-Yu Chi, and Liang-Gee Chen, *Fellow*, IEEE
DSP/IC Design Lab., Graduate Institute of Electrical Engineering, National Taiwan University, Taipei, Taiwan

***Abstract*-- We present an interactive clothing retrieval system, which supports query by a real-world image with target clothing and returns real-world images with similar clothing. The novel clothing shape feature is proposed to describe the shape of clothing in human-oriented coordinate system. We also propose a supervised method for learning a weighting matrix to minimize the intra-class distance while maximize the inter-class distance. The clothing retrieval results are reported quantitatively and the experiment results show that our system supports an accuracy of 61% over the dataset consist of practical various clothing types.**

I. Introduction

Content-based image retrieval (CBIR) is a research discipline of computer science that aims at searching for relevant images in a large database based on the actual contents rather than the metadata. Clothing retrieval, one of CBIR applications, is an increasingly popular research topic in recent years. Besides, researchers are enthusiastic on building an online content-based clothing retrieval system which could be as popular as their text-based counterparts [1][2].

Unlike general image retrieval system, users usually want to query by a region of interest (target clothing) rather than by the whole image in a clothing retrieval system. The input images downloaded from a website, captured on the street or reproduced from a magazine are usually photos of people with cluttered background. Many previous works [3][4][5][6] formulate the problem as cross-scenario retrieval, which means the query is a real-world image, while the returned images are captured in a clean environment. In this paper, we propose an interactive clothing retrieval system, which supports query by a real-world image with target clothing and returns real-world images with similar clothing.

The system integrates an interactive user interface, which is necessary not only to solve the issue of cluttered background, but also to assist the user in pointing out the target clothing. Interactive segmentation is the most flexible way to get what the user wants from an image. Instead of drawing a box containing target clothing adopted by [1][2], we use the geodesic based method proposed by Gulshan et al [7], which is fast and allows the user to do incremental refinement of the region selected. Thus the method is suitable for interactive retrieval system.

When people search for clothing similar to the query, shape is a discriminative feature to describe the type of clothing. Since the region of target clothing is obtained by interactive user interface, we propose a novel shape feature as fusion of several image moments with respect to body joints. Because each dimension of the shape feature is not equally important, we propose a supervised method to learn a weighting matrix which can minimize the distance between features in the same class while maximize the distance between features in the different classes.

The main contributions of this work can be summarized as follows. (1) we propose an interactive clothing retrieval system, which supports query by a real-world image with target clothing and returns real-world images with similar clothing; (2) we propose a novel clothing shape feature in human-oriented coordinate system; (3) we propose a supervised method for learning a weighting matrix to minimize the intra-class distance while maximize the inter-class distance.

The remainder of this paper is organized as follows. In the following section, we introduce the proposed clothing retrieval system. Section III shows our experiment results and analysis. We conclude with discussion in Section IV.

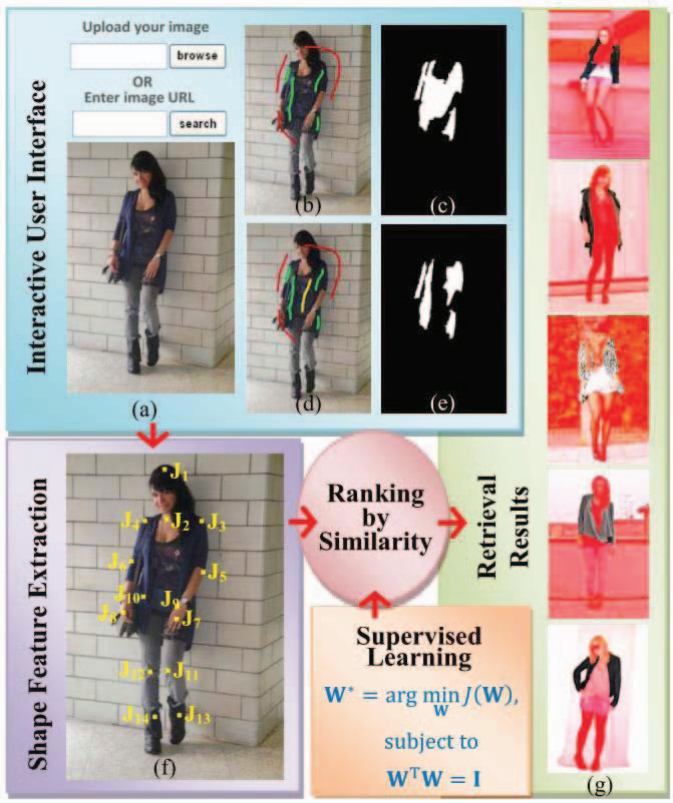

Fig. 1 System block diagram

II. Proposed System

The block diagram of the interactive clothing retrieval system is presented in Fig. 1. Firstly, an interactive user interface is applied to segmentation of target clothing. Secondly, the shape feature of clothing region is extracted. A supervised method is adopted to learn a weighting matrix offline. Finally, the clothing in the dataset is ranked by the weighted distance and the retrieval results are obtained. The details of each stage are described below.

A. Interactive User Interface

An interactive user interface is integrated to assist users in segmentation of the target clothing from an image uploaded from local or fetched on the pasted URL as shown in Fig. 1(a). In this paper, we adopt the interactive image segmentation method based on geodesic star convexity [7]. In Fig. 1(b), the user draws green strokes for target clothing, purple coat, and draws red stokes for others unwanted. The segmentation result is shown in Fig. 1(c), and we can notice that it is a challenging case because of similar color of coat and top. The sequential mode allows users to add input stokes (marked in yellow) sequentially for refining the segmentation as shown in Fig. 1(d). The result of segmentation is denoted as Eq. 1, and the binary segmentation map is shown in Fig. 1(e).

$$ROI(x,y) = \begin{cases} 1, & target\ clothing \\ 0, & otherwise \end{cases} \quad (1)$$

978-1-4799-1291-9/14/$31.00 ©2014 IEEE

B. Clothing Shape Feature Extraction

In the clothing retrieval system, human pose estimation is usually used as a preprocessing step to express the input image in human-oriented coordinate system. We adopted the pose estimation using flexible mixtures of parts model presented in [8]. Fig. 1(f) shows an example result of pose estimation and 14 joint locations are marked. Image moments are useful to describe objects after segmentation [9]. Therefore we extend the concept and use image moments with respect to human joint locations to describe the shape of segmented clothing region. The normalized n-th order image moment with respect to the reference point located at (x_R, y_R) is defined as

$$\mu_{pq} = \frac{1}{h^{p+q}} \sum_x \sum_y (x - x_R)^p (y - y_R)^q ROI(x, y) \tag{2}$$

where $n = p + q$ and h in the normalized factor is the height of the person in the image estimated by the difference between head location and ankle location. The reference points are chosen from the centroid of interest region and 14 joint locations. Finally we define a novel clothing shape feature $f \in \mathbb{R}^d$ as a fusion of first order and second order image moments with respect to each reference point.

C. Similarity Measurement

As mentioned above, we concatenate image moments as the clothing shape feature. Since each dimension is not equally important, we want to learn a weighting matrix $\mathbf{W} \in \mathbb{R}^{d \times d}$ that minimize the intra-class distance while maximize the inter-class distance. The weighted distance between two features f_i and f_j is denoted as

$$D_{ij} = [\mathbf{W}^T(f_i - f_j)]^T [\mathbf{W}^T(f_i - f_j)] \tag{3}$$

Motivated by [10], we define a matrix $\mathbf{S} \in \mathbb{R}^{l \times l}$ for supervising, which is incorporating the l pairwise labeled information as

$$\mathbf{S} = \begin{cases} 1 : (f_i, f_j) \in SC \\ -1 : (f_i, f_j) \in DC \\ 0 : otherwise \end{cases} \tag{4}$$

where $(f_i, f_j) \in SC$ is denoted as f_i and f_j share the same class label. Similarly, $(f_i, f_j) \in DC$ is denoted as f_i and f_j have different class labels.

We define the objective function measuring the empirical distance on the labeled data as Eq. 5. The compact matrix form is denoted as Eq. 6.

$$J(\mathbf{W}) = \sum_k \left\{ \sum_{(f_i, f_j) \in SC} {D_{ij}}^T D_{ij} - \sum_{(f_i, f_j) \in DC} {D_{ij}}^T D_{ij} \right\} \tag{5}$$

$$J(\mathbf{W}) = \frac{1}{2} \mathrm{tr}\{\mathbf{W}^\mathrm{T} \mathbf{D} \mathbf{S} \mathbf{D}^\mathrm{T} \mathbf{W}\} \tag{6}$$

$\mathbf{D} \in \mathbb{R}^{d \times l}$ joins the l pairwise unweighted distances. We intend to learn optimal weighting matrix $\mathbf{W}$ by minimizing the modified objective function with constraints as Eq. 7, which is a typical eigen-problem.

$$\mathbf{W}^* = \arg\min_{\mathbf{W}} J(\mathbf{W}), \quad \text{subject to } \mathbf{W}^\mathrm{T} \mathbf{W} = \mathbf{I} \tag{7}$$

After doing an eigenvalue decomposition on matrix $\mathbf{D}\mathbf{S}\mathbf{D}^\mathrm{T}$, $\mathbf{W}^*$ is the combination of the eigenvectors. Finally, all the clothing in the dataset can be ranked by similarity, which is defined as of the reciprocal of weighted distance.

III. Experiment Results and Analysis

We use the Fashionista dataset which is presented in [11] and publicly available. It contains 700 real-world photos with good visibility of the full body of the human model and covering a variety of clothing items. Each photo is segmented into superpixels annotated with 53 clothing labels. In our experiment, superpixels in an image with the same clothing labels are gathered into a clothing data. After ignoring the clothing labels which contains few data (*e.g.* cape and suit) and combining some similar labels (*e.g.* pants and jeans or jackets and coats), there are total 13 clothing labels concerned.

All clothing data are regarded as the results of segmentation and used to evaluate the discrimination of clothing shape feature and weighted similarity measurement. We divide all data into two sets, one for training and another for testing. The training set is created by randomly sampling 10 clothing data in each class, which is used to learn weighting matrix. The testing set consists of the remainder data of each class and there are total 3005 clothing data in testing set.

We follow the evaluation criterion of [5] using a ranking based criteria for evaluation. Given a query clothing q, all the clothing in the testing dataset can be assigned a rank by the weighted distance measurement. Let $Rel(i)$ be a binary value indicating the ground truth relevance between q and the i-th ranked clothing, relevant with value 1 or irrelevant with value 0. We can evaluate a ranking of top k retrieved clothing with respect to a query q by a precision defined as

$$Precision@k = \frac{\sum_i^k Rel(i)}{k} \tag{8}$$

Leave-one-out cross-validation is adopted and the precision@10 of each clothing label is shown in Table.1. Our system achieves 61% precision in average over the dataset consist of practical various clothing types. An example of retrieval results are reported qualitatively in Fig. 1(g). It is noted that our system appears promising with relevant clothing of a similar shape retrieved.

TABLE I FASHIONISTA DATASET PRECISION

Precision (%)	Hat	Glasses	Blouse	Top	Coat	Jeans	Socks	Skirt	Shorts	Shoes	Belt	Bag	Dress	All
Proposed	57	76	62	35	65	94	41	36	34	98	71	76	53	**61**

IV. Conclusion

In this work, we propose an interactive clothing retrieval system. Our system integrates a friendly user interface to assist users with pointing out the target clothing. The novel clothing shape feature is proposed to describe the shape of clothing in human-oriented coordinate system. A supervised method for learning a weighting matrix is also proposed to measure the distance between two fusion features. The experiment shows that our system appears promising with relevant clothing of a similar shape retrieved from real-world images and 61% precision over the dataset consist of practical various clothing types is achieved.

References

[1] PixCoo Information Technologies Company Limitid. Internet: http://www.pixcoo.com/, 2009.

[2] Hangzhou Taotaosou Science & Technology Corporation Limited. Internet: http://www.taotaosou.com/, 2010.

[3] S. Liu, et al. "Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set." CVPR, pp. 3330-3337, 2012.

[4] Y. Kalantidis, et al. "Getting the look: clothing recognition and segmentation for automatic product suggestions in everyday photos." ICMR, pp. 105-112, 2013.

[5] J. Fu, et al. "Efficient clothing retrieval with semantic-preserving visual phrases." ACCV, pp. 420-431, 2013.

[6] G. Cushen, et al. "Mobile visual clothing search." IMV, 2013.

[7] V. Gulshan, et al. "Geodesic star convexity for interactive image segmentation." CVPR, pp. 3129-3136, 2010.

[8] Y. Yang, et al. "Articulated Pose Estimation using Flexible Mixtures of Parts." CVPR, pp. 1385-1392, 2011.

[9] M. K. Hu. "Visual pattern recognition by moment invariants." Information Theory, IRE Transactions on 8.2, pp. 179-187, 1962.

[10] J. Wang, et al. "Semi-supervised hashing for scalable image retrieval." CVPR, pp. 3424-3431, 2010.

[11] K. Yamaguchi, et al. "Parsing clothing in fashion photographs." CVPR, pp. 3570-3577, 2012.