ACTIVE LEARNING BASED CLOTHING IMAGE RECOMMENDATION WITH IMPLICIT USER PREFERENCES

Chiao-Meng Huang[⋆] *Chia-Po Wei*[†] *Yu-Chiang Frank Wang*[†]

[⋆] Dept. Electrical Engineering, University of Southern California, Los Angeles, USA
[†] Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[⋆] chiaomenghuang@gmail.com [†] {cpwei, ycwang}@citi.sinica.edu.tw

ABSTRACT

We address the problem of user-specific clothing image recommendation in this paper. Different from prior retrieval approaches, we advance an active learning scheme during retrieval for inferring user preferences. With a recently developed sparse-coding based algorithm for content-based image retrieval, we utilize support vector regression (SVR) with a user-interaction training stage to observe user preferences based on the feedback of retrieval results. Therefore, there is no need to explicitly ask his/her preferences such as desirable colors or patterns of clothing images. A subjective evaluation on a commercial clothing image dataset confirms the effectiveness of our method, which is shown to produce more satisfactory recommendation results when comparing to state-of-the-art content-based image retrieval approaches.

Index Terms— Image retrieval, active learning, sparse representation

1. INTRODUCTION

When solving the tasks of image retrieval or recommendation, existing approaches can typically be divided into two categories: collaborative filtering (CF) [1] or content-based (CB) filtering [2]. Given a query input, collaborative filtering utilizes the behavior of prior users (e.g., purchase or browsing history) for suggesting items of potential (and common) interests. On the other hand, the goal of content-based filtering is to extract and analyze *features* for describing the query input, so that similar items from the database can be properly identified. It is not surprising that, if one needs to recommend items based on the query of a *new* product, collection of prior user histories is not possible, and thus collaborative filtering methods would encounter *cold-start* problems. While the above issue does not exist for content-based filtering approaches, how to select features which would be able to bridge the semantic gap between the input data and its high-level concept remains a challenging task. Moreover, neither of the above approaches are able to achieve *user-specific* retrieval or recommendation.

In this paper, we present an active learning based clothing image retrieval framework, which is able to implicitly learn the preferences from the user during retrieval processes, so that user-specific recommendation can be achieved. In the proposed framework, we utilize a sparse-coding based content-based image retrieval algorithm and employ a user-interaction stage to observe user feedback on initial retrieval results (in terms of ranking scores). This interaction stage can be considered as an active learning scheme for learning user-specific retrieval models, which will be able to produce personalized image recommendation. The flowchart of our clothing image recommendation method is shown in Fig. 1, in which blocks (a), (b), and (c) are explained in Sections 3.1, 3.2, and 3.3, respectively. Our experiments on a commercial clothing image dataset will confirm that our method improves user satisfaction and outperforms state-of-the-art content-based filtering approaches.

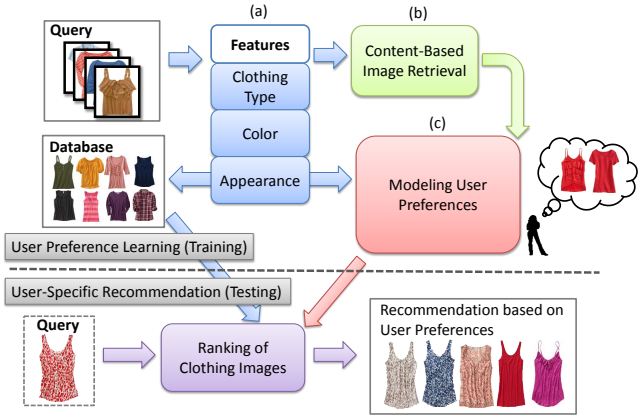

Fig. 1. Flowchart of our clothing image recommendation method.

2. RELATED WORKS

For image retrieval or recommendation, content-based filtering approaches focus on determining features or distance metrics for identifying images which are similar to the query input. For example, Chao *et al.* [3] considered the uses of HOG and LBP features as textural information to describe clothing images. Yang and Yu [4] applied HOG, SIFT, and DCT features to recognize different clothing images. Instead of choosing particular image features, Li *et al.* [5] performed independent component analysis (ICA) on query images, and

the extracted image components were considered as representative elements for retrieval. Chen *et al.* [6] studied the use of different types of features (i.e., shape, texture, or color) with the earth mover's distance (EMD) [7] as the distance metric, but they only evaluated individual performances of each feature. Recently, Wang and Zhang [8] proposed a retrieval framework which re-ranks color-based retrieval results by high-level feature attributes such as clothing type. Nevertheless, the above works did not take user preferences into consideration when addressing the retrieval task.

In order to extract high-level information for improved retrieval performance, Geng *et al.* [9] proposed to re-rank image search results by *image attractiveness*, but this property was determined by three pre-selected users and thus cannot be applied to user-specific recommendation. For solving this problem, we propose to advance support vector regression (SVR) [10] to infer user preferences. The use of SVR allows us to observe the relevance between the retrieved outputs and the feedback scores provided by each user, so that the learned SVR model is able to refine the retrieval results for improved user satisfaction. We note that, besides *pointwise* ranking [11] techniques like SVR, other learning-to-rank algorithms such as pairwise (e.g., RankSVM [12]) or listwise ranking (e.g., ListNet [13]) have been well studied in literatures. However, these methods require a complete ranking list from the entire dataset of interest, while pointwise ranking like SVR can be performed on a subset of retrieved outputs for inferring the rest of the data. Moreover, SVR is able to produce real-value ranking outputs when learning/predicting image retrieval results, which meets the goal of this work.

3. OUR PROPOSED METHOD

3.1. Heterogeneous Feature Extraction

As suggested in [3, 14], we consider three types of features to represent clothing images, including clothing type [14], color [6], and appearance; the corresponding feature dimensions are 13, 6250, and 1336, respectively. Once all three types of features are extracted, we normalize each and concatenate them into a *joint feature representation* $\mathbf{x}$, which will be applied for later retrieval/learning processes.

3.2. Sparse-Coding Based Image Retrieval

Sparse coding aims at reconstructing a signal $\mathbf{x} \in \mathbb{R}^{d\times 1}$ as a compact linear combination of $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_K] \in \mathbb{R}^{d\times K}$. It solves the following problem [15]:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1, \qquad (1)$$

where λ controls the sparsity of $\boldsymbol{\alpha}$. Since *data locality* has been observed to be a key issue in problems of clustering, dimension reduction [16], and data encoding/classification [17], Wang *et al.* [18] recently proposed a novel sparse coding scheme named locality-constrained linear coding (LLC) to ensure that similar $\mathbf{x}$ would obtain similar coding results

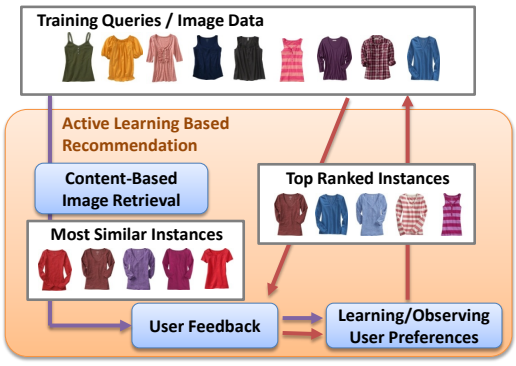

Fig. 2. Observing user preferences via active learning.

$\boldsymbol{\alpha}$. More specifically, the ℓ_1-norm regularization in (1) is replaced by a *locality adaptor* $\mathbf{p}$ as follows:

$$\min_{\boldsymbol{\alpha}} \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\mathbf{p} \odot \boldsymbol{\alpha}\|_2^2 \text{ s.t. } \mathbf{1}^\top\boldsymbol{\alpha} = 1. \qquad (2)$$

In our work, the ith entry of $\mathbf{p}$ in (2) calculates the Euclidean distance between $\mathbf{x}$ and $\mathbf{d}_i$, and the symbol $\odot$ denotes element-wise multiplication. We now rewrite (2) by the following Lagrange function:

$$L(\boldsymbol{\alpha}, \eta) = \|\mathbf{x} - \mathbf{D}\boldsymbol{\alpha}\|_2^2 + \lambda\|\mathbf{p} \odot \boldsymbol{\alpha}\|_2^2 + \eta(\mathbf{1}^\top\boldsymbol{\alpha} - 1).$$

Thus, the solution $\boldsymbol{\alpha}$ can be derived (see details in [14]) as

$$\boldsymbol{\alpha} = \tilde{\boldsymbol{\beta}}/(\mathbf{1}^\top\tilde{\boldsymbol{\beta}}), \quad \tilde{\boldsymbol{\beta}} = (\mathbf{C} + \lambda\,\mathrm{diag}(\mathbf{p})^2)^{-1}\mathbf{1}, \qquad (3)$$

where $\mathbf{C} = (\mathbf{x}\mathbf{1}^\top - \mathbf{D})^\top(\mathbf{x}\mathbf{1}^\top - \mathbf{D})$.

As discussed in Sect. 3.1, each clothing image $\mathbf{x}$ (and instances $\mathbf{d}_i$ to be retrieved) are in the form of a joint feature representation. To measure the distance between color features, we apply the chi-square distance since it is preferable for histogram-based features. As for clothing type and appearance features, the Euclidean distance is considered. From (2), it can be seen that this LLC-based retrieval algorithm would produce similar coding results $\boldsymbol{\alpha}$ for similar input $\mathbf{x}$, since it prefers and chooses dictionary atoms $\mathbf{d}_i$ which are close to the input. This property particularly favors image retrieval applications as shown in [14]. Thus, using this algorithm for content-based image retrieval, we apply (2) to calculate the coefficient $\boldsymbol{\alpha}$ for the query $\mathbf{x}$, and the entries of $\boldsymbol{\alpha}$ indicate the ranking/relevance of the corresponding clothing image $\mathbf{d}_i$ in the dataset to be retrieved.

3.3. Observing User Preferences via Active Learning

Although it has been shown in [14] that the use of (2) achieves promising retrieval results, such a content-based retrieval algorithm is not able to produce user-specific image recommendation. In order to address this problem, we apply active learning [19] for observing user preferences during the training stage of retrieval, so that the learned model will be able to achieve user-specific recommendation without the need to inquire desirable features of each user explicitly.

Active learning advances an interactive or iterative scheme for selecting training data, which provides new or updated information for the analysis of subsequent learning processes. Our proposed active-learning based framework for implicit learning of user preferences is shown in Fig. 2. Given a training set, we first randomly select 25 clothing images as queries, and apply (2) to identify the top five matches for each. For these $5 \times 25 = 125$ training query-output pairs, we have the user score each pair from 1 (lowest) to 5 (highest), which indicates the degree of satisfaction of himself/herself. Once this initial user feedback is complete, we advance support vector regression (SVR) [10] to model the relationship between these query-output pairs and their feedback scores. We take the differences of the joint feature representations between each query-output pair as inputs $\mathbf{x}_i^d$, and the associated feedback scores y_i as outputs for training SVR. Thus, the SVR in our active learning framework solves:

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\xi}^*} \; & \frac{1}{2}\mathbf{w}^\top \mathbf{w} + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \\ \text{s.t.} \; & y_i - (\mathbf{w}^\top \phi(\mathbf{x}_i^d) + b) \le \epsilon + \xi_i, \\ & (\mathbf{w}^\top \phi(\mathbf{x}_i^d) + b) - y_i \le \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \le 0, \; i = 1, 2, \ldots, n. \end{aligned} \tag{4}$$

In (4), n is the number of training instances, $\mathbf{w}$ represents the nonlinear SVR model, and C is the tradeoff between the generalization and the upper/lower training errors ξ_i/ξ_i^*, subject to a threshold ϵ. In our work, we consider Gaussian kernels for the SVRs, and their parameters are selected via cross-validation using training set data.

In the subsequent interactive stage of active leaning for observing user preferences, we randomly select another 25 query inputs from the training set and have the above SVR predict the user satisfaction scores on the resulting query-images pairs from the training data. Again, the top five ranked outputs for each query are sent to the user for obtaining their feedback scores. This interactive process allows us to *actively* select training query-output pairs for retraining/refining the SVR model. We note that, together with the previous round of query-output pairs and their feedback scores, the final SVR model will be specifically trained to observe user preferences with *high* satisfaction scores. This would meet the goal of user-specific recommendation.

Once this final learning process is complete, the SVR model will be applied to that particular user for retrieval tasks, as shown in the lower part of Fig. 1. In other words, for each query input from test data, this SVR will directly retrieve the most *relevant* and *satisfactory* output according to the SVR prediction scores. This is how we apply our proposed framework for solving user-specific recommendation.

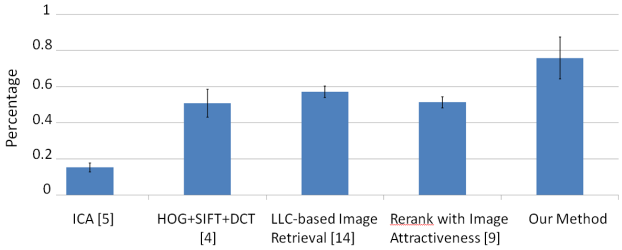

Fig. 3. Subjective evaluation of different methods.

4. EXPERIMENTS

4.1. Dataset and Experiment Settings

We collect 616 clothing images from Old Navy[1] as our clothing image dataset. We randomly divide this database into a training set with 100 images for implicit learning of user preferences, and a test set with the remaining 516 images for evaluating user-specific recommendation results.

During the training stage, we randomly select 25 out of the 100 training images as queries, and the top 5 matches for each query are extracted from the remaining 75 training images as the initial retrieved outputs. The user of interest will be asked to provide their feedback scores on these query-output pairs for SVR learning. As discussed in Sect. 3.3, the observed SVR will be applied to another 25 queries randomly chosen from the training set, and the retrieved results with top five ranking scores will be added to the training of the final SVR model. Thus, the total number of training query-output image pairs for SVR learning will be $25 \times 5 \times 2 = 250$. To evaluate the retrieval performance on the test set, we then randomly choose 60 out of the 516 test images as query inputs, and the remaining 466 images are the images to be retrieved (and to be evaluated by users).

4.2. Evaluations and Discussions

To compare our retrieval results with state-of-the-art content-based filtering approaches, we consider the ICA-based method [5], the approach of Yang and Yu [4] utilizing multiple types of features with ℓ_2-norm as the similarity measure, and a re-ranking method based on image attractiveness *et al.* [9]. We ask ten volunteers to perform a subjective evaluation on the retrieved outputs produced by different methods. For each volunteer, he or she needs to provide their feedback scores during the aforementioned training stage. When complete, we provide the query image from the test set, and the top five matches produced by different methods will be anonymously shown to that user (in a random order). Finally, each volunteer will rank the results provided by different approaches in terms of his/her satisfaction. Fig. 3 shows the performance comparisons on the averaged subjective evaluation results of all volunteers using different methods. Note that the vertical axis indicates the *percentage* of the retrieved outputs which are considered to be more satisfactory than those produced by

[1] http://www.oldnavy.com/

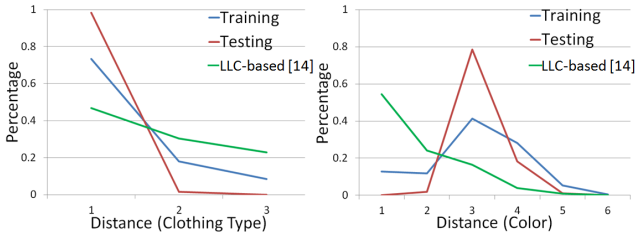

Fig. 4. An example of user preference analysis.

other approaches. It can be seen that our method achieved the highest satisfaction scores among all approaches. It is worth repeating that the method of [9] determined image attractiveness based on three pre-selected users, so it cannot be easily extended to user-specific recommendation.

To further discuss why our approach is able to achieve user-specific recommendation, we shown example retrieval results of a particular user in Fig. 4. In the left half of Fig. 4, the horizontal axis indicates the difference/distance between the query-output pair using clothing-type features, and the vertical axis denotes the percentage of the satisfaction scores which are higher than or equal to 4. Similarly, we analyze the relationship between the color features and the satisfaction scores of that user in the right half of Fig. 4. It can be seen that, although the LLC-based approach [14] is designed to retrieve images with similar features, the actual feedback given by the user for clothing type and color features were very different. More precisely, comparing the blue curves in the above two sub-figures, our method observed during the training stage that the user preferred clothing images with similar types but *not* colors. From the right half of Fig. 4, it can be observed that the retrieval results suggested by [14] with most similar colors actually achieved *poor* user satisfaction. Our empirical results on test query-retrieval pairs successfully preserved such preferences (see the red curves in Fig. 4), but state-of-the-art methods like [14] always suggest images with similar features. In other words, content-based filtering is only preferable when there is a strong correlation between the features of interest and the corresponding user preferences. From the above experiments and discussions, it can be verified that our proposed method is able to infer user preferences and achieves improved user-specific clothing image recommendation.

5. CONCLUSION

We proposed a clothing image retrieval method with the ability to infer user preferences implicitly. In contrast to prior collaborative or content-based filtering approaches, we advanced a user-interaction stage for observing user preferences without the need to explicitly ask their desirable clothing image features. As a result, besides avoiding cold-start problems, we will not necessarily retrieve images with most similar features as the query input does. Compared to state-of-the-art content-based image retrieval approaches, experiments on a commercial clothing image dataset confirmed the use of our proposed method for user-specific image recommendation with improved user satisfaction.

Acknowledgement This work is supported in part by National Science Council of Taiwan via NSC100-2221-E-001-018-MY2.

6. REFERENCES

[1] W. Deng, Q. Zheng, and L. Chen, "A fast and accurate collaborative filter," in *IEEE Int'l Conf. Granular Computing*, 2009.

[2] B. Logan, "Music recommendation from song sets," in *Proc. of the 5th Int'l Conf. on Music Information Retrieval*, 2004.

[3] X. Chao, M.J. Huiskes, T. Gritti, and C. Ciuhu, "A framework for robust feature selection for real-time fashion style recommendation," in *Proc. of the 1st ACM Int'l Workshop on Interactive Multimedia for Consumer Electronics*, 2009.

[4] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *IEEE ICIP*, 2011.

[5] X. Li, H. Yao, X. Sun, R. Ji, X. Liu, and P. Xu, "Sparse representation based visual element analysis," in *IEEE ICIP*, 2011.

[6] Z. Chen, L.Y. Duan, C. Wang, T. Huang, and W. Gao, "Generating vocabulary for global feature representation towards commerce image retrieval," in *IEEE ICIP*, 2011.

[7] Y. Rubner, C. Tomasi, and L.J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[8] X. Wang and T. Zhang, "Clothes search in consumer photos via color matching and attribute learning," in *ACM MM*, 2011.

[9] B. Geng, L. Yang, C. Xu, X.S. Hua, and S. Li, "The role of attractiveness in web image search," in *ACM MM*, 2011.

[10] A.J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, 2004.

[11] T.Y. Liu, "Learning to rank for information retrieval," *Foundations and Trends in Information Retrieval*, 2009.

[12] A.J. Smola, P.L. Bartlett, B. Schölkopf, and D. Schuurmans, *Advances in large margin classifiers*, vol. 1, MIT press Cambridge, MA, 2000.

[13] Z. Cao, T. Qin, T.Y. Liu, M.F. Tsai, and H. Li, "Learning to rank: from pairwise approach to listwise approach," in *Proc. Int. Conf. Machine Learning (ICML)*, 2007, pp. 129–136.

[14] C.-M. Huang, S. Chen, M. Cheng, and Y.-C. F. Wang, "A sparse-coding based approach to clothing image retrieval," in *IEEE Int'l Symposium on Intelligent Signal Processing & Communication Systems*, 2012.

[15] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *The Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[16] S.T. Roweis and L.K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, 2000.

[17] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *NIPS*, 2009.

[18] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *IEEE CVPR*, 2010.

[19] B. Settles, "Active learning literature survey," Tech. Rep., University of Wisconsin, Madison, 2009.