MEAN SHIFT FEATURE SPACE WARPING FOR RELEVANCE FEEDBACK

Yao-Jen Chang, Keisuke Kamataki

Carnegie Mellon University
kevinchang@cmu.edu, keisuke@cs.cmu.edu

Tsuhan Chen

Cornell University
tsuhan@ece.cornell.edu

ABSTRACT

Relevance feedback has been taken as an essential tool to enhance content-based information retrieval systems by keeping the user in the retrieval loop. Among the fundamental relevance feedback approaches, feature space warping has been proposed as an effective approach for bridging the gap between high-level semantics and the low-level features. By examining the fundamental behavior of the feature space warping, we propose a new approach to harness its strength and resolve its weakness under various data distributions. Experiments on both synthetic data and real data reveal significant improvement from the proposed method.

Index Terms— Relevance feedback, content-based information retrieval, feature space warping

1. INTRODUCTION

With the prevalence in high-speed networking and high-volume networked storage, vast amounts of publicly accessible images and videos have become a very useful resource in our daily life. One fundamental question in utilizing this huge resource is how to search for the target media that matches user's intention. Content-based information retrieval (CBIR) systems are designed to deal with weakly annotated data by similarity matching [9]. And relevance feedback is one of the essential tools in reducing the semantic gap between the low-level features and the richness of human semantics [4, 12].

Relevance feedback approaches can be roughly divided into three categories: (i) Moving the query point: query point movement (QPM); (ii) Manipulating the feature space or the metric space: adjust the weights of each feature component, move every sample by feature space warping (FSW) [1], or modify the similarity measure based on user's feedback [5]; (iii) Learning classifier online: train a support vector machine (SVM) or an Adaboost classifier to separate relevant samples from irrelevant samples [10, 11].

In this work, we mainly focus on the FSW algorithm due to its capability in reducing the semantic gap by altering the original feature space. In the original proposed FSW algorithm [1], query point serves as a warping center such that every sample except the query point in the archive changes its position according to a warping function. On the contrary, QPM algorithm only moves the query point without changing other samples in the archive. We'll show that these two extreme algorithms are well-matched to each other, such that better performance can be achieved by integrating them together in the relevance feedback loop.

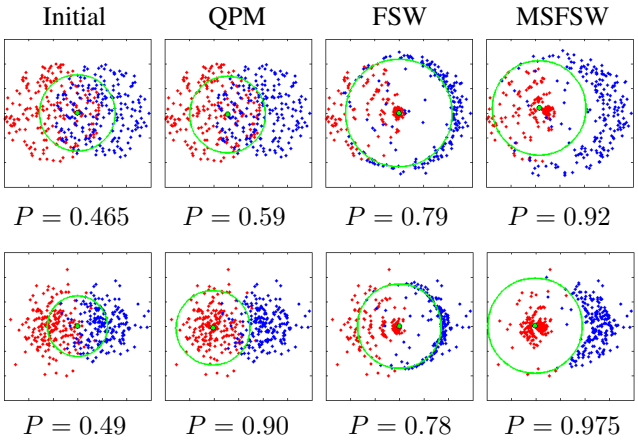


Fig. 1. Relevance feedback results in the toy example with uniform distributions (top), and Gaussian distributions (bottom). Two clusters (red, blue) are generated in each dataset. The green point represents the location of query point (or warping center), and the green circle shows the range containing the top 200 nearest neighbors to the query point. Precision evaluated within the green circle is shown under each graph.

The rest of this paper is organized as follows. In the next section, a toy example is given to demonstrate the strength and weakness of the feature space warping algorithm. Section 3 details the proposed mean shift feature space warping algorithm. The experimental results on real data are described in Section 4. And finally, Section 5 concludes the paper and addresses future works under investigation.

2. A TOY EXAMPLE

In the following example, two synthetic datasets are created to illustrate the behavior of different relevance feedback algorithms under different data distributions. As shown in Fig.1,

two partially overlapped data clusters are generated in each dataset. In the first dataset, each cluster is generated with 200 points from uniform distribution within a fixed distance to each cluster center while each cluster in the second dataset is generated from Gaussian distribution. The query point lies in the origin of feature space assumed to belong to the left cluster. Relevance feedbacks are given to the 200 samples retrieved by nearest neighbor search in each iteration. Five iterations of relevance feedback are conducted to examine performance of three methods: QPM, FSW, and the proposed mean shift feature space warping method (MSFSW).

Several observations from Fig.1 can be summarized as follows:

- QPM favors Gaussian distributions and gradually pushes the query point to its corresponding cluster center.
- QPM performs poorly under uniform distributions. The query point stops moving when nearby relevant feedbacks are uniformly distributed.
- FSW performs poorly under Gaussian distributions when the query point is far away from the cluster center. Closer relevant samples fast moving toward the query point make far away samples more difficult to move toward the query center.
- MSFSW outperforms QPM and FSW with satisfactory performance under both data distributions.

Based on the above observations, we found QPM and FSW behave quite differently under different data distributions. Even though FSW shows good performance in the literature [1], the inherent weakness could hinder a broader use. In the next section, the proposed MSFSW algorithm is presented which shows promising performance improvement on the toy example.

3. MEAN SHIFT FEATURE SPACE WARPING

With similar notations used in [1], we address our MSFSW algorithm as follows. Given a query point $\boldsymbol{q}$ in the feature vector space, k samples are retrieved by nearest neighbor search. Within these k samples, user specifies relevance feedbacks to M samples ($M \leq k$), forming a relevant set $\{\boldsymbol{f_p}\}$ and irrelevant set $\{\boldsymbol{f_n}\}$. These two sets of points form a force field to guide all data samples $\{\boldsymbol{p}\}$ in the whole feature space to move toward or away from the warping center $\boldsymbol{w}$. More formally, for each $\boldsymbol{p} \in \{\boldsymbol{p}\}$, its warped point $\boldsymbol{p'}$ is updated as

$$\boldsymbol{p'} = \boldsymbol{p} + \lambda \sum_{j=1}^{M} u_j \exp(-c|\boldsymbol{p} - \boldsymbol{f_j}|)(\boldsymbol{w} - \boldsymbol{p}), \qquad (1)$$

where the scalar value u_j can be simply set to $+1$ if $\boldsymbol{f_j} \in \{\boldsymbol{f_p}\}$, and -1 if $\boldsymbol{f_j} \in \{\boldsymbol{f_n}\}$, global coefficients c and λ are used to control the inference for each feedback to each sample

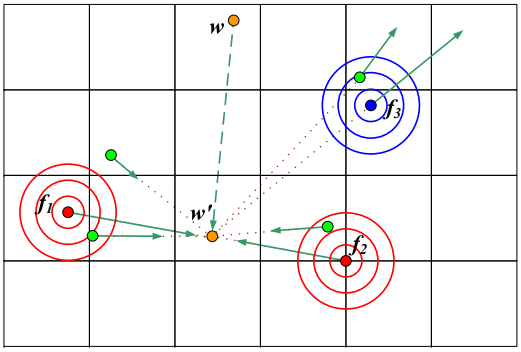


Fig. 2. Mean shift feature space warping: given two relevant samples $\boldsymbol{f_1}$ and $\boldsymbol{f_2}$, and an irrelevant sample $\boldsymbol{f_3}$ with user's feedback, the warping center $\boldsymbol{w}$ is shifted to the mean of relevant samples followed by the warping operation for all samples in the archive w.r.t. the updated warping center $\boldsymbol{w'}$.

and maximum moving factor from any point $\boldsymbol{p}$ toward the warping center $\boldsymbol{w}$.

According to the original feature space warping algorithm, the warping center $\boldsymbol{w}$ is equal to $\boldsymbol{q}$. Thus, the query point will always stay in the original position. Other points will move toward or far away from $\boldsymbol{q}$ based on its proximity to relevant and irrelevant sets. However, this will cause problems as illustrated in the toy example on the Gaussian distribution case. Therefore, we propose to move the warping center instead of staying at $\boldsymbol{q}$. A good strategy is to adopt the Rocchio's query point movement formula [9]:

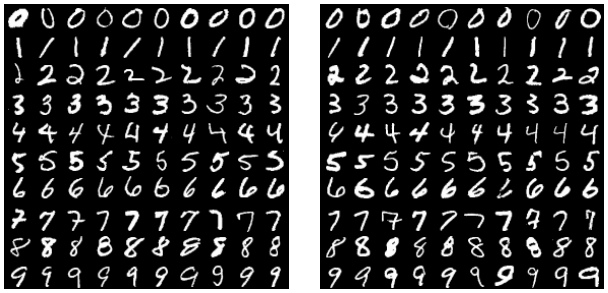
$$\boldsymbol{w'} = \alpha \boldsymbol{w} + \beta \overline{\boldsymbol{f_p}} - \gamma \overline{\boldsymbol{f_n}}, \qquad (2)$$

where $\boldsymbol{w}$ is the warping center initially set to $\boldsymbol{q}$, $\overline{\boldsymbol{f_p}}$ is the mean of relevant set $\{\boldsymbol{f_p}\}$, and $\overline{\boldsymbol{f_n}}$ is the mean of the irrelevant set $\{\boldsymbol{f_n}\}$. Parameters α, β, and γ can be tuned to optimize the performance. A natural choice without exhaustive search on the parameter space is to choose $\beta = 1, \alpha = \gamma = 0$, i.e., using the *mean* of relevant feedbacks to *shift* the warping center such that all points in the archive will move toward or far away from the adapted warping center as depicted in Fig. 2.

With the above formulations, the MSFSW algorithm provides a flexible parameterization for switching between the two extreme algorithms: QPM by setting $\alpha = \gamma = \lambda = 0$, $\beta = 1$, and FSW by setting $\alpha = 1$, $\beta = \gamma = 0$. Beyond simple switching, the integration of full parameter settings takes advantage from both algorithms and results in a more powerful feature space warping algorithm for relevance feedback.

4. EXPERIMENTAL RESULTS

In Section 2, dramatic improvement of the proposed method is shown with synthetic datasets created from Gaussian and uniform distributions. In this section, two real datasets are utilized to validate the performance of the proposed method.



(a) (b)

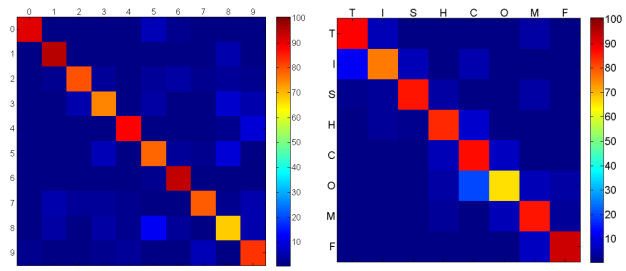
Fig. 3. Sample images from the MNIST database of handwritten digits: (a) the training set, and (b) the testing set.

Fig. 4. Sample images from the outdoor scene dataset: the training set (top two rows), and the testing set (bottom two rows).

4.1. Handwritten Digits Dataset

The MNIST database of handwritten digits [3], contains a training set of 60000 samples, and a test set of 10000 samples of ten digits written by approximately 250 writers. Each image is size-normalized to 28×28 pixels. Sample images are shown in Fig. 3. In this work, the gray-level image texture is concatenated to form a 784-dimensional feature vector. A randomly-selected subset of 1000 images (100 images/digit) from the training set is utilized to train a low-dimensional space by using linear discriminant analysis (LDA) [2]. Another set of 1000 images extracted from the testing set are projected to the LDA space to form the data samples in our evaluation. The confusion matrix of the LDA features shown in Fig. 5(a) has high diagonal components, meaning that each cluster is quite compact and discriminative.

Performance comparison is conducted with four iterations of relevance feedback based on the average precision evaluated from top 100 nearest neighbors by taking each individual sample as a query. The result for the three methods with different numbers of feedbacks is shown in Fig. 6. The abbreviations 'Q', 'F', 'M' in the legend of Fig. 6 stand for the three methods 'QPM', 'FSW', and 'MSFSW', respectively. The number (25, 50, 75, and 100) after each abbreviation



(a) (b)

Fig. 5. Confusion matrices on the test set in the LDA feature space: (a) the handwritten digits dataset with digits from 0 to 9, and (b) the outdoor scene dataset with 8 scene categories. Colors in each block (i, j) indicate the percentage of samples in class i being closest to the cluster center of class j.

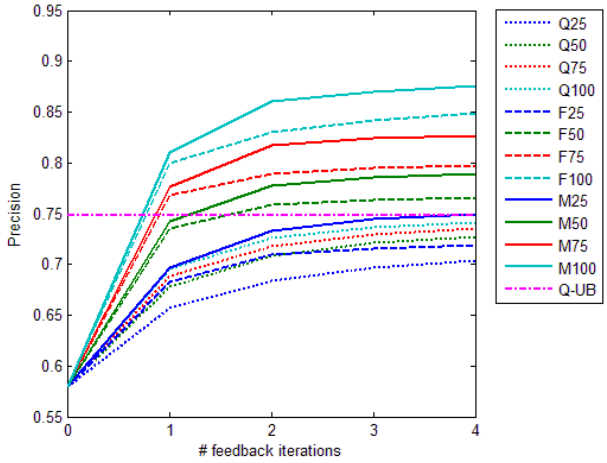


Fig. 6. Performance evaluation on the MNIST database of handwritten digits.

stands for the number of feedbacks given in each iteration. 'Q-UB' represents the theoretical performance upper bound of QPM estimated by moving the query to the cluster center of its corresponding class. With the compact cluster distribution implied by the confusion matrix, the upper bound of QPM is almost approachable by QPM with large amount of feedbacks in each feedback iteration as shown in the Q100 curve in Fig.6.

The proposed MSFSW method significantly outperforms the other two methods, while the FSW shows less improvement given fewer feedbacks. Note that both MSFSW and FSW methods can easily break the theoretical performance upper bound of QPM. This is because MSFSW and FSW have the ability to move potentially irrelevant samples away from the query center and attracts far away relevant samples toward the query center, while QPM can only stay with nearby irrelevant samples within the region of relevant samples.

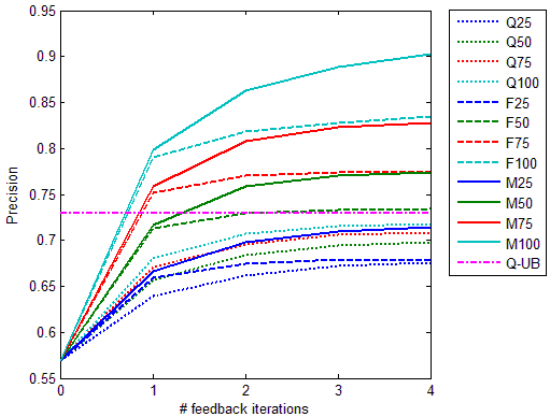


Fig. 7. Performance evaluation on the outdoor scene dataset.

4.2. Outdoor Scene Dataset

The outdoor scene dataset is collected by Oliva and Torralba [6]. This dataset contains 2688 images with eight outdoor scene categories: tall building, inside city, street, highway, coast, open country, mountain, and forest. Sample image are shown in Fig. 4. Different from the handwritten digits, the contents of the scene database contain large variations within each category. To provide a more consistent feature vector, a 512-dimensional Gist descriptor [7] is utilized to represent the whole scene within each image.

Similar to the procedure stated in the previous subsection, 100 images per category are selected to train a low-dimensional space by LDA. Another set of 800 images extracted to form the testing set are projected to the LDA space to form the data samples in our evaluation. Performance evaluation is done with the same way as the handwritten digits database. The result for the three methods with different numbers of feedbacks is shown in Fig. 7. As shown in the confusion matrix in Fig. 5(b), the feature space of this dataset is less discriminative than the handwritten digits dataset, implying the performance upper bound of QPM is more difficult to reach for QPM. Again, the FSW and MSFSW have no difficulty in breaking this performance upper bound with 50 or more feedbacks.

The overall performance trend is quite similar to the result of handwritten digits database. However, the FSW shows insignificant improvement to QPM with 25 feedbacks at the 4th iteration. Its performance improvement also becomes rather flat after the first iteration. On the contrary, MSFSW continues improvement with more and more iterations of relevance feedback. Consistent performance superiority justifies the proposed method a better solution for relevance feedback.

5. CONCLUSIONS

In this work, we presented the mean shift feature space warping algorithm for relevance feedback. By examining the strength and weakness of two extreme relevance feedback approaches, the proposed method takes advantage from both approaches to provide better enhancement in bridging the gap between low-level features and high-level semantics.

Although preliminary experimental results indicate promising performance improvement on the proposed method, several issues should be taken into consideration. The first problem is how to apply the proposed algorithm to a huge dataset. Exhaustive warping each sample in the database is highly expensive. A pre-filtering mechanism would be necessary to select a reasonable subset for use in the relevance feedback loop. Another interesting problem is how to apply the proposed approach to features represented by large and sparse visual words, such as the 1M visual words used in the task of object retrieval from over one million images [8]. Last but not least, the study on how to apply the feature space warping with online classifier learning could potentially bring another significant performance boost to relevance feedback.

6. REFERENCES

[1] H. Y. Bang and T. Chen. Feature space warping: an approach to relevance feedback. *IEEE Int. Conf. Image Processing*, 1:968–971, 2002.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):711–720, 1997.

[3] Y. LeCun and C. Cortes. The MNIST database of handwritten digits. http://yann.lecun.com/exdb/mnist/.

[4] Y. Liu, D. Zhang, G. Lu, and W. Y. Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40:262–282, 2007.

[5] G. P. Nguyen, M. Worring, and A. W. M. Smeulders. Interactive search by direct manipulation of dissimilarity space. *IEEE Trans. Multimedia*, 9(7):1404–1415, 2007.

[6] A. Oliva and A. Torralba. The outdoor scene dataset. http://people.csail.mit.edu/torralba/code/spatialenvelope/.

[7] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *Int. J. Computer Vision*, 42(3):145–175, 2001.

[8] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. *IEEE Conf. Computer Vision and Pattern Recognition*, 2007.

[9] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in MARS. *IEEE Int. Conf. Image Processing*, 2:815–818, 1997.

[10] D. Tao, X. Tang, X. Li, and X. Wu. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(7):1088–1099, 2006.

[11] K. Tieu and P. Viola. Boosting image retrieval. *Int. J. Computer Vision*, 56(1/2):17–36, 2004.

[12] X. S. Zhou and T. S. Huang. Relevance feedback in image retrieval: a comprehensive review. *Multimedia Systems*, 8:536–544, 2003.