Learning Query-Dependent Distance Metrics for Interactive Image Retrieval

Junwei Han, Stephen J. McKenna, and Ruixuan Wang

School of Computing, University of Dundee, Dundee DD1 4HN, UK
{*jeffhan,stephen,ruixuanwang*}*@computing.dundee.ac.uk*

Abstract

An approach to target-based image retrieval is described based on on-line rank-based learning. User feedback obtained via interaction with 2D image layouts provides qualitative constraints that are used to adapt distance metrics for retrieval. The user can change the query during a search session in order to speed up the retrieval process. An empirical comparison of online learning methods including ranking-SVM is reported using both simulated and real users.

1 Introduction

Critical components in any content-based image retrieval (CBIR) system are the methods used to compute the dissimilarity of images and to obtain feedback from users during search. Most early CBIR systems relied on pre-defined distance functions (e.g., Euclidean distance in feature space) for image dissimilarity measurement. Although these so-called computer-centric systems are relatively easy to implement, some inherent drawbacks limit their performance. Image understanding is highly subjective; each user will have different personal intentions and preferences when searching for images. These can vary from session to session even if identical queries are posed. Therefore, pre-defined distance metrics based on fixed combinations of features are inadequate.

Relevance feedback (RF) was proposed to help address this limitation [11]. It attempts to adapt to a user's preferences by performing on-line learning of a query-dependent distance function based on user feedback. Most RF techniques operate under the assumptions that users are looking for a category of images and start with a query example from that category. After each iteration of retrieval, the user provides feedback on the relevance of the retrieved images. Machine learning methods such as support vector machines [14] or

manifold learning [4] are then used to refine the distance function based on the feedback. The refined distance function is then applied to obtain new retrieval results.

In general, traditional RF techniques learn from user feedback that consists of labelling returned images as relevant or irrelevant, or perhaps assigning quantitative relevance scores to them. The first type of feedback ignores information on the degree of relevance. The second type of feedback can be difficult and time-consuming for users to provide. For example, a user might puzzle unnecessarily over whether a relevance score should be 0.80 or 0.85. A more appropriate form of feedback is based on relative comparisons or ranks, e.g. "the query image is more similar to image A than it is to image B". The study of how to learn from relative comparisons is attracting increasing attention. Joachims [7] proposed a ranking-SVM method which converted the learning task to a standard SVM classification task. It was applied to learning from 'clickthrough' data for Web search engines. Schultz and Joachims [12] extended this approach to learn distance metrics. Freund *et al.* [2] presented the RankBoost algorithm.

Rank-based distance learning has been used to solve vision problems. Frome [3] proposed a method to learn local image-to-image distance functions for classification by combining relative comparison information and image shape features. Hu *et al.* [5] explored a multiple-instance ranking approach based on ranking-SVM to order images within each category for retrieval. Lee *et al.* [8] employed a rank-based distance metric to retrieve images of tattoos. However, these three approaches [3,5,8] were evaluated under the scenario of *offline* learning.

This paper proposes a target-based interactive image retrieval approach that incorporates on-line rank-based learning. It makes the following contributions. (i) A novel user feedback mechanism is proposed. Instead of asking users to label training data as relevant or irrelevant, the users are able to offer relative, qualitative information to the system. (ii) Rank-based online learning is used to refine the distance metric based on constraints generated from user feedback. (iii) The user can change the query example during a session in order to speed up retrieval. (iv) An empirical comparison of methods including ranking-SVM is reported. This includes evaluations based on simulated users and a preliminary evaluation with real users.

2 Problem formulation

The scenario considered here is that of search for a specific target image in an image set I. Search is terminated when the target is found or the user decides

to give up the search. This contrasts with many CBIR systems in which it is assumed that users are searching for images that belong to some category of images. Target-based retrieval is very useful in applications such as logo, trademark, historical photograph, and painting search [1].

A search session consists of a series of iterations. At the t^{th} iteration, the user is presented with a 2D visualisation of a set of retrieved images, a subset of I. These appear as a 2D layout, L_{t-1}, and are arranged based on their content. The user selects a query image, q_t, from this layout. The selected query may or may not be the same image as the previous query, q_{t-1}. The user can also move the images in the layout to express judgements about relative similarity to the query. The user's similarity judgements will depend on the target and the context. This results in a set of inequality constraints, P_t. A learning algorithm then uses the selected query and the constraints to obtain a new distance metric, D_t. This metric is then used to retrieve the closest matches from the image set and a visualization algorithm produces a new 2D layout from these matches for the user. This sequence can be summarised as follows.

$$\{L_{t-1}\} \xrightarrow{\text{user}} \{q_t, P_t\} \tag{1}$$

$$\{I, q_t, P_t\} \xrightarrow{\text{learner}} \{D_t\} \tag{2}$$

$$\{I, q_t, D_t\} \xrightarrow{\text{matcher and visualizer}} \{L_t\} \tag{3}$$

Let $\mathbf{q}$ and $\mathbf{x}$ denote feature vectors of a query image and an image in the database, respectively. A parameterized (Mahalanobis) distance metric can be used: $D(\mathbf{q}, \mathbf{x}; \mathbf{W}) = \sqrt{(\mathbf{q}-\mathbf{x})^T \mathbf{W} (\mathbf{q}-\mathbf{x})}$, where the symmetric matrix $\mathbf{W}$ should be positive semi-definite (i.e., $\mathbf{W} \succeq 0$) to ensure that D is a valid metric. If $\mathbf{W}$ is a diagonal matrix, the distance metric becomes a weighted Euclidean distance, which is adopted here:

$$D(\mathbf{q}, \mathbf{x}; \mathbf{w}) = \sqrt{\sum_i W_{mm}(q_m - x_m)^2} = \sqrt{\langle \mathbf{w} \cdot ((\mathbf{q}-\mathbf{x}) * (\mathbf{q}-\mathbf{x})) \rangle} \tag{4}$$

where q_m and x_m denote the m^{th} elements of $\mathbf{q}$ and $\mathbf{x}$. The m^{th} diagonal element, W_{mm}, of $\mathbf{W}$ reflects the importance of the m^{th} feature. "$\langle \cdot \rangle$" denotes the inner product and "$*$" the element-wise product of two vectors. $\mathbf{w}$ is the vector consisting of diagonal elements of $\mathbf{W}$. Note that $\mathbf{w} \succeq 0$.

An initial layout L_{-1} can be generated using a representative (or randomly selected) subset of I. Alternatively, if the user has a query example q_0 already

to hand, the first two steps, (1) and (2), are omitted in the first iteration and $\mathbf{w}_0$ is taken to be a vector of ones so that D_0 is a Euclidean distance metric.

Key to the proposed target-based retrieval approach is to learn the parameter $\mathbf{w}_t$ based on the user-provided constraints, P_t. This learning component of the system will be described in Section 4. First, the user interaction and the visualization component are introduced.

3 User interaction based on 2D visualization

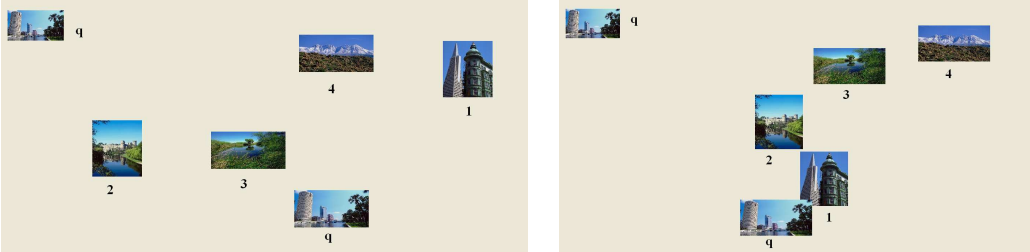
In a CBIR system, the user interface must present retrieval images to users and enable interaction for the purpose of providing feedback. A well-designed interface will make this interaction easy and quick for users, and enhance the efficiency of the system. Firstly, retrieval results are visualised as 2D layouts. Secondly, users can move images on the layout to convey their preferences. The relative locations of the images are then taken to provide ranking information from which the system can learn.

Most traditional CBIR systems show retrieval results as lists sorted in order of decreasing similarity to the query. Moghaddam *et al.* [9] argued that visualizing images in a 2D space can be superior, allowing mutual similarities to be reflected. Rodden [10] performed user studies which demonstrated that 2D layouts could enable users to find a target image or group of images more quickly.

The visualisation method of Wang *et al.* [15] was modified here in order to generate 2D layouts, L_t. The main idea of this method is to combine an unsupervised dimensionality reduction algorithm with a term that rewards layouts that have high entropy. It enables layouts to be generated that represent a trade-off between (i) preserving the relative distances between images and (ii) avoiding image overlaps and unoccupied layout regions. The distance between two images was measured using D_t. Whereas Wang *et al.* [15] used an ISOMAP term, a multi-dimensional scaling (MDS) term was used instead in this paper because of the relatively small number of images in each layout. In common with [15], Renyi quadratic entropy was used. Fig. 1 shows an example layout. The query is at the top-left of the interface. The 50 most similar images to this query are arranged automatically based on color correlogram features [6]. Visually similar images are grouped together which is helpful for users when making judgements and providing feedback on their preferences.

Traditional RF techniques assume that users are searching for a category of images and require users to label results as relevant or irrelevant thus indicating whether or not they are in the same category as the query. Such feedback

Fig. 1. An example layout showing a query (top left) and 50 closest matches.



(a) (b)

Fig. 2. An example of user interaction in which images are arranged relative to the query. (a) Before interaction. (b) After interaction.

is essentially a set of class labels. Cox *et al.* [1] argued that this burdens the user by forcing them to decide upon a useful categorization of images even if unfamiliar with the database. It is appropriate to category-based search rather than target-based search. In this paper, users are allowed to drag images in the 2D visualization space and the relative locations of images and query image convey their preferences. Fig. 2 shows an example in which only five retrieved images are used, for clarity of presentation. Fig. 2(a) shows the automatically generated layout. Fig. 2(b) shows the layout after the user has chosen to move the images to reflect their perceived relative similarity to the query (image 1). This user-defined layout yields an ordering of the images in terms of similarity to the query, in this case $1 \succ 2 \succ 3 \succ 4$, where $\succ$ denotes a ranking relationship. This ordering implies a set of inequalities on the distance measure being used by the user. If the user arranges N images relative to the query then there are $\frac{N(N-1)}{2}$ such inequalities. However, if we assume that the user's measure is a metric then most of these are redundant and only $N-1$ inequalities are needed. These are used to provide constraints for the learning algorithm to learn a new metric D_t. In the example shown in Fig. 2(b) the constraints would be $P_t = \{D_t(q_t, 1; \mathbf{w}) < D_t(q_t, 2; \mathbf{w}), D_t(q_t, 2; \mathbf{w}) < D_t(q_t, 3; \mathbf{w}), D_t(q_t, 3; \mathbf{w}) < D_t(q_t, 4; \mathbf{w})\}$.

Moghaddam *et al.* [9] also used 2D visualization to collect feedback but their method differs in two main respects. Since their purpose was to group images

for browsing, all relationships between images were used. Instead, this paper is concerned with target-based retrieval so only relationships between images and query are used. Secondly, [9] used absolute locations of images for learning. Instead, a more qualitative feedback is adopted here for reasons discussed earlier.

4 Rank-based distance metric learning

The objective of the learner is to infer the parameter $\mathbf{w}$ of the distance metric $D(.,.;\mathbf{w})$. Ideally, this metric should satisfy the constraints P. (For clarity, the subscript t is omitted in this section). A maximal-margin formulation with slack variables is adopted here to perform this learning task. The task is formulated as the following optimization problem which has the same form as in [3] and [12].

$$
\begin{aligned}
&\min_{\mathbf{w},\xi_{(q,i,j)}} \tfrac{1}{2} \parallel \mathbf{w} \parallel^2 + C \sum_{(q,i,j)} \xi_{(q,i,j)} \\
&s.t. \\
&\forall (D(q,i;\mathbf{w}) > D(q,j;\mathbf{w})) \in P : \; D^2(q,i;\mathbf{w}) - D^2(q,j;\mathbf{w}) \geq 1 - \xi_{(q,i,j)} \qquad (5) \\
&\forall (q,i,j) : \; \xi_{(q,i,j)} \geq 0 \\
&\mathbf{w} \succeq 0
\end{aligned}
$$

Here, $\parallel \mathbf{w} \parallel^2$ is a regularization term and indicates structural loss, $\xi_{(q,i,j)}$ are slack variables, and C is a trade-off parameter. Substituting (4) into the first set of constraints in (5) leads to

$$\langle \mathbf{w} \cdot (\mathbf{d}_{q,i} - \mathbf{d}_{q,j}) \rangle \geq 1 - \xi_{(q,i,j)} \qquad (6)$$

where $\mathbf{d}_{q,i} = (\mathbf{q} - \mathbf{x}_i) * (\mathbf{q} - \mathbf{x}_i)$, and $\mathbf{x}_i$ is the feature vector for the i^{th} image.

The constraint $\mathbf{w} \succeq 0$ is needed to ensure that the learned distance is a valid metric. Incorporating this constraint is non-trivial. Without this constraint, the setting of the optimization would be the same as that of ranking-SVM and standard quadratic programming solvers such as SVM-Light could be used [7]. The purpose of ranking-SVM is to learn a ranking function which is expected to correctly sort data. In ranking-SVM, elements of $\mathbf{w}$ can have negative values and the ranking values can be negative. Although image retrieval can be formulated as a ranking problem using such an approach [5], it is not suitable

for query-by-example. If ranking-SVM is used to perform query-by-example, the output for the query itself will be zero as desired. However, outputs for other images can be negative since elements of $\mathbf{w}$ can be negative. It leads to an undesirable situation in which other images can be deemed to be more similar to the query than the query itself. In Section 5, this point will be demonstrated empirically.

Frome [3] proposed a custom dual solver for the optimization problem which is adopted here. This approach can guarantee that $\mathbf{w}$ is non-negative. Moreover, its fast optimization speed and good performance make it suitable for online learning. It iteratively updates dual variables until convergence:

$$\mathbf{w}^{(t)} = \max\{\sum_{(q,i,j)} \alpha^{(t)}_{(q,i,j)}(\mathbf{d}_{q,i} - \mathbf{d}_{q,j}), 0\} \tag{7}$$

$$\alpha^{(t+1)}_{(q,i,j)} = \min\{\max\{\frac{1 - \langle \mathbf{w}^{(t)} \cdot (\mathbf{d}_{q,i} - \mathbf{d}_{q,j})\rangle}{\parallel \mathbf{d}_{q,i} - \mathbf{d}_{q,j} \parallel^2} + \alpha^{(t)}_{(q,i,j)}, 0\}, C\} \tag{8}$$

where $0 \leq \alpha_{(q,i,j)} \leq C$ are the dual variables and are initialized to zero. The reader is referred to [3] for implementation details.

5 Experiments

A set of $10,009$ images from the Corel dataset was used for experiments. These images have semantic category labels and there are 79 categories such as tiger, model, and castle. Each category contains at least 100 images. Category labels are not used in what follows. Three types of low-level feature were used: a 36-dimensional color histogram, an 18-dimensional texture feature based on a wavelet transformation [13], and a 144-dimensional color correlogram. In all experiments, the trade-off parameter C was set to 20. The computational speed mainly depends on the learning algorithm and the visualization algorithm. A Matlab implementation normally takes a few seconds to perform both learning and visualization on a 2.4GHz, 3.5GB PC which is adequate for on-line processing.

It is common to evaluate CBIR systems using simulated users since interactions from real users are expensive to collect. This is usually done using pre-defined, fixed category labels. Retrieved results are automatically marked as relevant if they share a category label with the query. Rank-based learning has been evaluated similarly [3,8,5]. The underlying assumption is that images within the same pre-defined category are always more similar to each other than images from other categories. In contrast, the purpose of the system in

this paper is to find a target image without the use of pre-defined categories. Therefore, a different evaluation method is proposed.

Two experiments were performed. In the first experiment, the user was simulated. The second experiment involved online evaluation with four real users.

5.1 Evaluation using a simulated user

Experiments using a simulated user were performed as follows. A fixed distance metric, $D_{user}(\mathbf{q}, \mathbf{x}; \mathbf{w}_{user})$ based on the image features was used by the simulated user. Each simulated search session was initiated by randomly selecting two images from the database, one as query and one as target. In the first iteration, the system retrieved images based on a pre-specified metric $D_0(\mathbf{q}, \mathbf{x}; \mathbf{w}_0)$ that differed from D_{user}. At each iteration, the simulated user used D_{user} to select the closest retrieved image to the target and rank ordered the retrieved images in terms of distance to the query. In this way, $N - 1$ inequality constraints were generated and used by the learning algorithm to update the distance metric to better approximate D_{user}. Search terminated when the target was retrieved or when a maximum number of iterations was reached. Once an image had been retrieved it was excluded from being retrieved in subsequent iterations.

More specifically, 36-dimensional color histograms and 18-dimensional texture features were concatenated to give 54-dimensional feature vectors to represent the images. A total of 100 search sessions was simulated with up to 50 iterations per session. The distance metric was initialised to only use the texture features. In other words, the weights in Eqn. (4) were set to equal, non-zero values for each of the 18 texture features, and to zero for each of the 36 colour features. In contrast, the simulated user used a distance metric in which colour features had equal, non-zero values and texture features had weights of 0.

The number of images retrieved at each iteration is a free parameter, N. Larger values of N result in more feedback information at each iteration and greater choice of query for the subsequent iteration. However, large N also results in increased time and effort from the user at any given iteration. Fig. 3 shows results obtained by using different values of $N \in \{5, 10, 20, 30, 40, 50\}$. Performance was measured as the fraction of trials in which the target was retrieved within a given number of iterations. The proposed method retrieved nearly all targets within a few iterations provided that N was large enough.

The method was compared to several alternatives. Ranking-SVM was used to learn $\mathbf{w}$ from the inequality constraints. Code from SVM Light [7] was used. Another method involved randomly selecting N images without any simulated user interaction. Another method used the initial metric (i.e. texture only) for

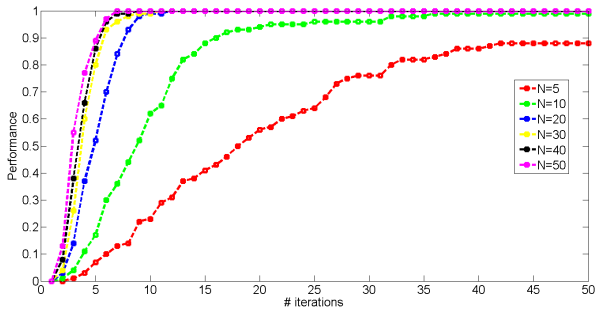


Fig. 3. Performance comparisons with different N.

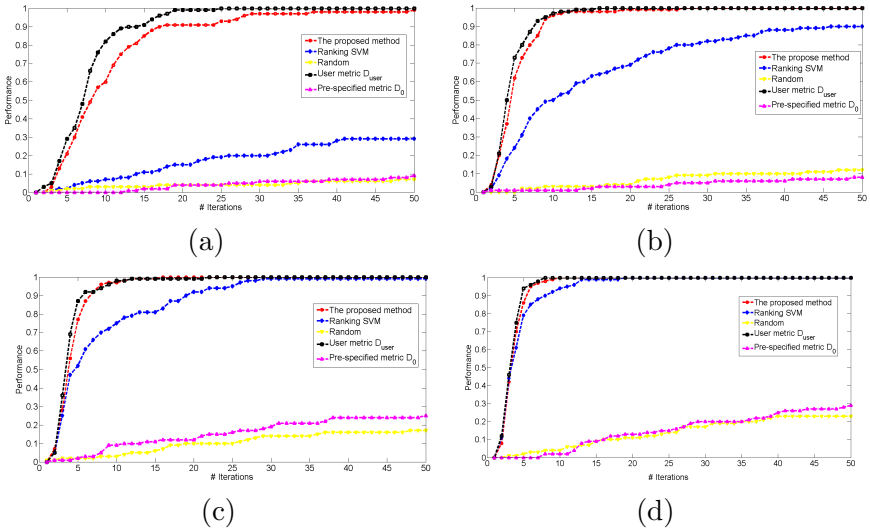


Fig. 4. Performance comparisons using different methods. (a) $N = 10$, (b) $N = 20$, (c) $N = 30$, and (d) $N = 40$.

matching. Finally, the methods were compared to retrieval using the ideal metric, D_{user} (i.e. colour only). Fig. 4 shows comparative results for various values of N. The results demonstrate that the proposed approach is better than ranking-SVM especially when the value of N is small. For example, for $N = 10$, the proposed method achieved the retrieval rate of 59% and ranking-SVM achieved the retrieval rate of 7% at the tenth iteration. For $N = 20$, the performance of the proposed method and ranking-SVM was 96% and 50% respectively at the tenth iteration. Retrieval rates obtained by the proposed method quickly approached those obtained using the ideal metric as N increased. When $N = 40$, the proposed method differed from the ideal metric by about 5% only between the second round and fifth iterations. This indicates that the method was able to capture preferences.

5.2 Interactive online experiment with users

Future work will be needed to fully evaluate the proposed approach with users. Here we report only a preliminary user experiment. Four subjects (2 male and 2 female) tested the system. Each subject performed ten search sessions. Target images were selected by users and came from 36 different categories. Before each session, the system displayed a layout of 100 images selected randomly from the image database. The user selected whichever of these images was most similar to the target as the initial query image unless the user did not consider any of these 100 images to be similar to the target. In the latter case, the system offered them another 100 randomly selected images from which they were forced to choose. Given the results of the simulation above, $N = 20$ was chosen as a reasonable trade-off. A 144-dimensional color correlogram feature vector was used to represent each image in this experiment.

Each iteration requires the user to select a query and move images to provide feedback on similarity to the query. This is more time-consuming than the CPU time for learning, matching and visualization. Query selection normally took less than $10s$ while arranging the images took $25 - 50s$. If a target was not found after 10 iterations, search was deemed to have failed. There were 40 search sessions in total and, of these, 5 failed, 3 found the target without any interaction other than initial query selection, 20 were successful within 5 iterations, and 12 others were successful using more than 5 iterations. Overall, successful sessions required an average of 5 iterations to retrieve the target.

6 Conclusions and Recommendations

An approach to adaptive target-based image retrieval was proposed. Maximal-margin learning based on constraints provided through user feedback was used to learn distance metrics. The experimental results suggest that the approach has potential for application to real-world interactive image retrieval.

The idea of RF is to bring users into the loop and so evaluations with the real users are essential. However, few previous papers report interactive online tests. The interactive online test lead to two useful observations. Firstly, some retrieved images were considered irrelevant to the query by users. As such, users were not interested in them and did not like or found it difficult to provide judgements about them. Future work should investigate improved feedback mechanisms that allow users to efficiently select images they are interested in from the retrieval layouts and only provide feedback on those images. A second observation is that selecting appropriate query images to start a search session plays an important role in yielding success. Most failed

searches were due to the initial query being very dissimilar to the target image. This problem would be reduced by selection of the initial query using an image browsing system that can present a global view of the whole database [16].

Acknowledgments

The authors thank A. Ward for valuable discussions. This research was supported by the UK Technology Strategy Board grant "FABRIC: Fashion and Apparel Browsing for Inspirational Content" in collaboration with Liberty Fabrics Ltd., System Simulation Ltd. and Calico Jack Ltd. The Technology Strategy Board is a business-led executive non-departmental public body, established by the government. Its mission is to promote and support research into, and development and exploitation of, technology and innovation for the benefit of UK business, in order to increase economic growth and improve the quality of life. It is sponsored by the Department for Innovation, Universities and Skills (DIUS). Please visit www.innovateuk.org for further information.

References

[1] Cox, I. J., Miller, M. L., Minka, T. P., Papathomas, T. V., and Yianilos, P. N.: The Bayesian image retrieval system, Pichunter: Theory, implementation, and psychophysical experiments. IEEE Trans. on Image Processing 9(1), 20–37 (2000).

[2] Freund, A., Iyer, R., Schapire, R. E., Lozano-Perez, T.: An efficient boosting algorithm for combining preferences. J. Machine Learning Research 4, 939–969 (2003).

[3] Frome, A.: Learning Distance Functions for Examplar-based Object Recognition. PhD thesis, UC Berkeley, (2007).

[4] He, X., Ma, W.-Y., and Zhang, H.-J.: Learning an image manifold for retrieval. In: Proceedings of the 12th Annual ACM international conference on Multimedia, New York, 17–23 (2004).

[5] Hu, Y., Li, M., and Yu, N.: Multiple-instance ranking: Learning to rank images for image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, USA, (2008).

[6] Huang, J., Ravi Kumar, S., Mitra, M., Zhu, W., and Zabih, R.: Spatial color indexing and applications. Int. J. of Computer Vision 35, 245–268 (1999).

[7] Joachims, T.: Optimizing search engines using clickthrough data. In: The Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Alberta, Canada, 133–142 (2002).

[8] Lee, J.-E., Jin, R., and Jain, A. K.: Rank-based distance metric learning: An application to image retrieval. In: Computer Vision and Pattern Recognition (CVPR), Anchorage, USA, (2008).

[9] Moghaddam, B., Tian, Q., Lesh, N., Shen, C., and Huang, T. S.: Visualization and user-modeling for browsing personal photo libraries. International Journal of Computer Vision 56, 109–130 (2004).

[10] Rodden, K.: Evaluating Similarity-Based Visualisations as Interfaces for Image Browsing. PhD thesis, University of Cambridge, (2001).

[11] Rui, Y., Huang, T. S., Ortega, M., and Mehrotra, S.: Relevance feedback: A power tool in interactive content-based image retrieval. IEEE Trans. on Circuits and Systems for Video Technology 8, 644–655 (1998).

[12] Schultz, M. and Joachims, T.: Learning a distance metric from relative comparisons. In: Neural Information Processing Systems (NIPS), Berlin, (2003).

[13] Smith, J. R. and Chang, S.-F.: Automated binary texture feature sets for image retrieval. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, Atlanta, USA, 2239–2242 (1996).

[14] Tong, S. and Chang, E.: Support vector machine active learning for image retrieval. In: ACM Conference on Multimedia, Ottawa, Canada, 107–118 (2001).

[15] Wang, R., McKenna, S. J., and Han, J.: High-entropy layouts for content-based browsing and retrieval. In: ACM International Conference on Image and Video Retrieval, Santorini, Greece, (2009).

[16] Ward, A. A., McKenna, S. J., Buruma, A., Taylor, P., and Han, J.: Merging technology and users: applying image browsing to the fashion industry for design inspiration. In: Content-based Multimedia Indexing, London, 288-295, (2008).