An Interactive Approach for Filtering out Junk Images from Keyword Based Google Search Results

Yuli Gao, Jinye Peng, Hangzai Luo, Daniel A. Keim, Jianping Fan,

***Abstract*—Keyword-based Google Images search engine is now becoming very popular for online image search. Unfortunately, only the text terms that are explicitly or implicitly linked with the images are used for image indexing and the associated text terms may not have exact correspondence with the underlying image semantics, thus the keyword-based Google Images search engine may return large amounts of junk images which are irrelevant to the given keyword-based queries. Based on this observation, we have developed an interactive approach to filter out the junk images from keyword-based Google Images search results and our approach consists of the following major components: (a) A kernel-based image clustering technique is developed to partition the returned images into multiple clusters and outliers. (b) Hyperbolic visualization is incorporated to display large amounts of returned images according to their nonlinear visual similarity contexts, so that users can assess the relevance between the returned images and their real query intentions interactively and select one or multiple images to express their query intentions and personal preferences precisely. (c) An incremental kernel learning algorithm is developed to translate the users' query intentions and personal preferences for updating the mixture-of-kernels and generating better hypotheses to achieve more accurate clustering of the returned images and filter out the junk images more effectively. Experiments on diverse keyword-based queries from Google Images search engine have obtained very positive results. Our junk image filtering system is released for public evaluation at: *http://www.cs.uncc.edu/∼jfan/google_demo/*.**

***Index Terms*—Junk image filtering, mixture-of-kernels, incremental kernel learning, hyperbolic image visualization, user-system interaction.**

I. Introduction

As online image sharing and personal journalism become more and more popular, there is an urgent need to develop more effective image search engines, so that users can successfully access large-scale image collections that are available on the Internet. Keyword-based Google Images search engine has achieved great success on exploiting the associated text terms for automatic indexing of large-scale online image collections. Unfortunately, Google Images search engine is still unsatisfactory because of the relatively low precision rate and the appearance of large amounts of junk images [1-5]. One major reason for this phenomena is due to the fact that Google Images search engine simplifies the image search problem as a purely text-based search problem, and the underlying assumption is that the image semantics are directly related to the associated text terms (which can be extracted automatically from the associated text documents, the file names or the URLs). However, such oversimplified online image indexing approach has ignored that the associated text terms may not have exact correspondence with the underlying image semantics. This is the major reason why Google Images search engine may return large amounts of junk images which are irrelevant to the given keyword-based queries. In addition, a lot of real world settings, such as photo-sharing web sites, may only be able to provide biased and noisy text terms for image annotation which may further mislead the keyword-based Google Images search engine. Therefore, there is an urgent need to develop new algorithms for filtering out the junk images from keyword-based Google Images search results [1-5].

The visual properties of the returned images and the visual similarity contexts between the returned images are very important for users to assess the relevance between the returned images and their real query intentions. Unfortunately, Google Images search engine has completely ignored such important characteristics of the images and the keywords for image indexing may not be expressive enough for describing the rich details of the visual content of the images, thus it is very hard for Google Images search engine to assist users on looking for some particular images according to their visual properties. The huge number of returned images and the appearance of the junk images may bring huge burden on the users to look for some particular images via page-by-page browsing. Even the low-level visual features may not be able to carry the image semantics directly, they can definitely be exploited to filter out the junk images and enhance the users' abilities on finding some particular images according to their visual properties.

With the increasing computational power of modern computers, it is possible to incorporate image analysis techniques into Google Images search engine without degrading its response speed significantly. Recent advance in computer vision and multimedia computing can also allow us to take advantages of the rich visual content (embedded in the images) for image semantics interpretation.

Another shortcoming for Google Images search engine is that the underlying techniques for query result display (i.e.,

Yuli Gao was with the University of North Carolina, Charlotte, NC 28223, USA. He is now in HP Labs. e-mail: yuli.gao@hp.com

Jinye Peng is with School of Electronics and Information, Northwestern Polytechnical University, Xi'an, CHINA. e-mail: jinyepeng@hotmail.com

Hangzai Luo was with the University of North Carolina, Charlotte, NC 28223, USA. He is now in East China Normal University. e-mail: hluo@sei.ecnu.edu.cn

Daniel A. Keim is with Computer Science Institute, University of Konstanz, Konstanz, Germany. email: keim@inf.uni-konstanz.de

Jianping Fan is with the Department of Computer Science, University of North Carolina, Charlotte, NC 28223, USA. e-mail: jfan@uncc.edu

Manuscript received November 26, 2008; revised Febrary 18, 2009.

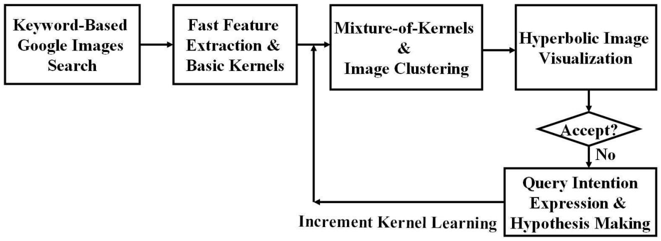

Keyword-Based Google Images Search
Fast Feature Extraction & Basic Kernels
Mixture-of-Kernels & Image Clustering
Hyperbolic Image Visualization
Accept?
No
Query Intention Expression & Hypothesis Making
Increment Kernel Learning

Fig. 1. The flowchart for our interactive junk image filtering system.

page-by-page ranked list of the returned images) cannot allow users to assess the relevance between the returned images and their real query intentions effectively. Many pages are needed for displaying large amounts of returned images, thus it is very tedious for users to look for some particular images of interest through page-by-page browsing. Things may become worse when the ambiguous keywords with many potential word senses are used for query formulation. Because the visual properties of the returned images are completely ignored for image ranking, the returned images with similar visual properties may be separated into different pages. Ideally, users would like to have a good global overview of the returned images in a way that can reflect the principal visual properties of the returned images effectively and allow them to navigate large amounts of returned images interactively according to their nonlinear visual similarity contexts, so that they can assess the relevance between the returned images and their real query intentions interactively.

By integrating multi-modal information (visual similarity, associated text terms, and users' feedbacks), we have developed an interactive approach to filter out the junk images from keyword-based Google Images search results. Our interactive junk image filtering scheme takes the following major steps as shown in Fig. 1: (a) Keyword-based Google Images search engine is first performed to obtain large amounts of returned images for a given keyword-based query. (b) Fast feature extraction is then performed on the returned images to extract both the global visual features and the local visual features for image content representation. (c) The diverse visual similarities between the returned images are characterized more accurately by combining multiple kernels (i.e., mixture-of-kernels) and a kernel-based clustering technique is used to partition the returned images into multiple clusters and outliers. (d) A hyperbolic visualization algorithm is integrated to display large amounts of returned images according to their nonlinear visual similarity contexts for supporting more understandable relevance assessment. (e) If necessary, users can select one or multiple relevant images to express their query intentions and personal preferences precisely and generate better hypotheses for junk image filtering. An incremental kernel learning algorithm is developed to translate the users' query intentions and personal preferences for updating the mixture-of-kernels to achieve more accurate characterization of the diverse visual similarities between the images and learn more accurate SVM classifier for filtering out the junk images more effectively. (f) The updated mixture-of-kernels is further used to achieve more accurate clustering of the returned images and create more precise visualization of the returned images for next hypothesis making loop.

The paper is organized as follows. Section 2 briefly reviews some related works on junk image filtering; Section 3 introduces our work on fast feature extraction for image content representation; Section 4 introduces our mixture-of-kernels algorithm; Section 5 describes our incremental kernel learning algorithm to achieve more accurate image clustering and generate better hypotheses for junk image filtering; Section 6 summarizes our work on algorithm and system evaluation; We conclude in Section 7.

II. Related Works

Some pioneer works have been done to improve the performance of keyword-based Google Images search engine [1-5]. To filter out the junk images from keyword-based Google Images search results, Fergus et al. have applied constellation model to re-rank the returned images according to the appearances of the image objects and some promising results have been achieved [1-2], where both the appearance models for the distinct object parts and the geometry model for all the possible locations of the object parts are incorporated to learn the object models explicitly from a set of training images. Unfortunately, large amounts of high-quality training images are needed to learn such complex object models reliably. However, image search results (returned by Google Images search engine) are very noisy and cannot directly be used as the reliable image set for training such complex object models. Because large amounts of training images are needed to achieve reliable learning of the object models, such process for object model learning could be computation-sensitive and thus it is very hard to achieve junk image filtering in real time or nearly in real time. In addition, the image semantics could be interpreted in multiple levels: the underlying object classes and the semantics of entire images at different concept levels [27].

The research team from Microsoft Research Asia have developed several approaches to achieve more effective clustering of online image search results by using visual, textual and linkage information [3-5]. Instead of treating the associated text terms as the single information source for online image indexing and retrieval, they have incorporated multi-modal information sources to exploit the mutual reinforcement between the images and their associated text terms. In addition, a triparties graph is generated to model the linkage relationships among the low-level visual features, images and their associated text terms. Thus automatic image clustering is achieved by supporting triparties graph partition. Incorporating multi-modal information sources for image indexing may improve the performance of online image search engines significantly, but it may be very hard to extend such approach for achieving online junk image filtering because the triparties linkage graph could be very complex and triparties graph partition could be a computation-sensitive process.

On the other hand, both the interpretation of image semantics and the assessment of image relevance are user-dependent (i.e., the user's background knowledge plays an important role in image semantic interpretation and relevance assessment), it is very important to incorporate human expertises and their powerful capabilities on pattern recognition for enhancing online image search. Thus one potential solution for junk image filtering is to involve users in the loop of image retrieval via relevance feedback, and many relevance feedback techniques have been proposed in the past [6-13]. Unfortunately, all these existing relevance feedback techniques require users to label a reasonable number of returned images into the relevant class and the irrelevant class for learning a reliable model to predict the user's query intentions, thus they may bring huge burden on the users [34]. When large-scale image collections come into view, a limited number of labeled images may not be representative for large amounts of unseen images and thus a limited number of labeled images may not be sufficient for learning an accurate model to predict the user's query intentions precisely. 2D data page is used for image display and the nonlinear visual similarity contexts between the images are completely ignored for image ranking, thus the returned images with similar visual properties may be separated into different pages. Such page-by-page image display approach cannot allow users to see a good global overview of large amounts of returned images (i.e., image clusters and their similarity structures) at the first glance, thus users cannot assess the relevance between the returned images and their real query intentions effectively and provide their feedbacks precisely for junk image filtering. In addition, all these existing relevance feedback techniques have not provided a good solution for hypotheses visualization and assessment (i.e., visualizing the margin between the relevant images and the junk images to enable better hypothesis assessment).

New visualization tools are strongly expected to support user-dependent and goal-dependent choices about what to display and how to provide feedback. Image seekers often express a desire for a user interface that can organize the search results into meaningful groups, in order to help them make sense of the search results, and to help them decide what to do next. Some pioneer works have been done on supporting similarity-based image visualization [36-43], but most existing techniques for image projection and visualization may perform well when the images belong to one single cluster, and fail to project the images nicely when they are spread among multiple clusters with diverse visual properties.

To capture the users' query intentions precisely for generating better hypotheses for junk image filtering, three key issues should be addressed jointly: (a) incremental kernel learning should be supported for reducing the computational cost [44-48], so that users can interactively change the underlying hypotheses for filtering out the junk images in real time or nearly in real time; (b) the convergence of the underlying techniques for incremental kernel learning should be guaranteed; (c) an interactive interface should be developed to enable similarity-based visualization of large amounts of returned images [36-43], generate more understandable assessment of the hypotheses for junk image filtering (i.e., make the margin between the relevant images and the junk images to be more visible and more assessable), and allow users to express their query intentions more precisely for generating better hypotheses and learning more accurate SVM classifier for junk image filtering.

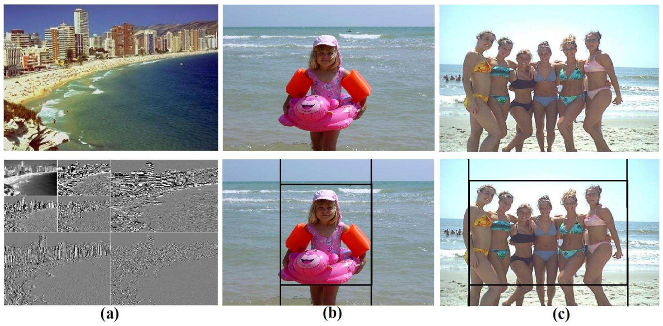


Fig. 2. Wavelet transformation for texture feature extraction and 10 simple image partition patterns for local color histograms extraction.

III. Fast Feature Extraction

It is very important to integrate the visual properties of the images for improving keyword-based image search, and there are three widely accepted approaches for image content representation and feature extraction: (1) *image-based* approach that extracts the visual features from entire image without performing image segmentation [17-19]. (2) *region-based* approach that extracts the visual features from homogeneous image regions by performing image segmentation [14-16]. (3) *object-based* approach that extracts the visual features from salient image objects [20-22].

The major advantage for the image-based approach is that no segmentation is performed, thus it can support fast feature extraction. However, the visual features that are extracted from entire images may not be able to characterize the intermediate image semantics effectively at the object level. The major problem with the region-based approach is that the homogeneous image regions or image grids may not correspond to the underlying salient image components, thus it cannot characterize the intermediate image semantics accurately at the object level. On the other hand, the object-based approach can characterize the intermediate image semantics effectively at the object level and the image contexts (i.e., pictorial structures or spatial relationships between the image objects) can further be extracted to achieve more accurate image semantics interpretation at the concept level. Unfortunately, automatic image object detection is still an open problem for computer vision community [32].

For online junk image filtering, the underlying approach for image content representation and feature extraction should be able to: (a) characterize both the global visual properties and the local visual properties of the images effectively and efficiently; (b) reduce the computational cost significantly for feature extraction and image similarity determination because such junk image filtering process should be achieved in real time or nearly in real time.

Based on these understandings, we have developed an alternative approach for fast feature extraction to achieve a good trade-off between the effectiveness for image content representation and the reduction of the computational cost for feature extraction and image similarity determination. To achieve more accurate representation of the diverse visual properties of the images, both the global visual features and the local visual features are extracted for image content representation and similarity characterization. To reduce the computational cost for feature extraction, we use the thumbnails from Google Images instead of the original-size images for feature extraction.

The global visual features such as global color histogram and wavelet texture features (shown in Fig. 2(a)) can provide the global perceptual properties of entire images, but they may not be able to capture the object information within the images accurately [23-24]. Even SIFT (scale invariant feature transform) features can allow object recognition against the cluttered background [25-26], they may be too computation-sensitive for supporting online junk image filtering. On the other hand, our local color histograms can provide the principal visual properties of the image objects at certain accuracy level and reduce the computational cost significantly. In our current implementations, the global visual features consist of 32-bin global color histogram and 62-dimensional texture features from Gabor filter banks. The local visual features consist of 10 32-bin local color histograms and they are extracted from 10 simple image partition patterns as shown in Fig. 2(b) and Fig. 2(c), so that the principal visual properties of the image objects will not be weakened by the visual properties of the background which may cover the most space of the picture. When people take the photos, they may normally put the attended objects in the centers of the images. Thus we assume that the image objects of attention normally locate at the centers of the images in two ways as shown in Fig. 2(b) and Fig. 2(c), and such assumption is correct in general for our experiments.

One major advantage of our fast feature extraction approach is that it can achieve a good trade-off between the effectiveness for image content representation (i.e., characterizing both the global visual properties of the entire images and the local visual properties of the image objects) and the significant reduction of the computational cost for feature extraction, thus it can be performed in real time. It is also important to note that our local color histograms focus on extracting the local visual properties of the image objects for achieving more accurate image clustering by reducing the misleading effects of the background on image similarity characterization at the object level, such local color histograms are not required to be discriminative enough to achieve automatic object detection and recognition because users will be involved in the loop of image retrieval.

IV. Image Similarity Characterization

To filter out the junk images from Google Images search results, the first question is to define more suitable similarity functions to characterize the diverse visual similarity contexts between the returned images accurately. Recently, the use of kernel functions for data similarity characterization plays an important role in the statistical learning framework [29-33], where the kernel functions may satisfy some mathematical requirements and possibly capture some domain knowledge.

To achieve more accurate approximation of the diverse visual similarity contexts between the images, different kernels should be designed for different feature subsets because their statistical properties of the images are very different. Unfortunately, most existing machine learning tools use one single kernel for diverse image similarity characterization and completely ignore the heterogeneity of the statistical properties of the images in the high-dimensional multi-modal feature space [31]. Thus three basic image kernels (global color histogram kernel, wavelet filter bank kernel, local color histogram kernel) are first constructed to characterize the diverse visual similarity contexts between the images, and a linear combination of these three basic image kernels (i.e., mixture-of-kernels) can further form a family of mixture-of-kernels for characterizing the diverse visual similarity contexts between the images more accurately [32]. Because multiple kernels are seamlessly integrated to characterize the heterogeneous statistical properties of the images in the high-dimensional multi-modal feature space, our mixture-of-kernels algorithm can achieve more accurate image clustering and can also provide a natural way to add new feature subsets and their basic kernels incrementally.

In this paper, we have incorporated three basic descriptors to characterize various visual properties of the images: (a) global color histogram; (b) texture histograms for wavelet filter banks; (c) local color histograms. The first two descriptors are computed from every pixel of the whole image; while the third descriptor is computed from 10 simple image partition patterns as shown in Fig. 2(b) and Fig. 2(c).

The global color histogram kernel $K_1(x, y)$, which is used to characterize the visual similarity between the global color histograms u and v for two images x and y, is defined as:

$$K_1(x,y) = e^{-\chi^2(u,v)/\delta} = \prod_{i=1}^{32} e^{-\chi_i^2(u(i),v(i))/\delta_i} \qquad (1)$$

where $\delta = [\delta_1, \cdots, \delta_{32}]$ is set to be the mean value of the χ^2 distances between all the images in our experiments, $u(i)$ and $v(i)$ are the ith component for two color histograms u and v. We quantify the HSV color space into 32 bins (i.e., 16 bins for H, 8 bins for S and 8 bins for V).

The wavelet texture kernel $K_2(x, y)$ can be decomposed as a product of component kernels for different wavelet filter banks $e^{-\chi_i^2(h_i(x),h_i(y))/\sigma_i}$:

$$K_2(x,y) = \prod_{i=1}^{n} e^{-\chi_i^2(h_i(x),h_i(y))/\sigma_i} \qquad (2)$$

where the component kernel $e^{-\chi_i^2(h_i(x),h_i(y))/\sigma_i}$ is used to characterize the similarity between two images x and y according to the ith wavelet filter bank, $h_i(x)$ and $h_i(y)$ are the histograms of the ith wavelet filter bank for two images x and y.

The local color histogram kernel $K_3(x, y)$, which is used to characterize the similarity between two sets of local color histograms Υ and Ψ for two images x and y, is defined as:

$$K_3(x,y) = e^{-\chi^2(\Upsilon,\Psi)/\Theta} = \prod_{j=1}^{10}\prod_{i=1}^{32} e^{-\chi_i^2(\Upsilon_j(i),\Psi_j(i))/\theta_i} \qquad (3)$$

where $\Theta = [\theta_1, \cdots, \theta_{32}]$ is set as the mean value of the χ^2 distances of all the images in our experiments. In our experiments, we have extracted 10 local color histograms for each image according to 10 simple image partition patterns as shown in Fig. 2(b) and Fig. 2(c). The local color histograms are used to characterize the appearances of the image objects and their local visual properties with certain accuracy level, thus the local color histogram kernel should be able to determine the image similarity at the object level with certain accuracy level.

The diverse visual similarity contexts between the returned images are characterized more accurately by using a linear combination of these three basic image kernels (i.e., mixture-of-kernels) [31-32]:

$$\kappa(x,y) = \sum_{i=1}^{3} \beta_i K_i(x,y), \qquad \sum_{i=1}^{3} \beta_i = 1 \tag{4}$$

where $\beta_i \geq 0$ is the importance factor for the ith basic image kernel $K_i(x,y)$ for image similarity characterization.

The rules for kernel combination (i.e., mixture-of-kernels construction by selecting the optimal values for these three importance factors β) depend on two key issues: (a) The relative importance of various feature subsets for diverse image similarity characterization; (b) The users' query intentions and personal preferences (which may not be known without user's input). Based on this observation, we have developed an incremental approach to achieve optimal kernel combination, which treats the junk image filtering process as an incremental process for SVM classifier training by taking the users' query intentions and personal preferences into consideration.

Because the keywords for query formulation may not be able to capture the user's query intentions effectively and efficiently, the keyword-based Google Images search engine may not know which image cluster is relevant to the user's intentions or which image cluster is irrelevant at the beginning. Thus it is very hard to define suitable criteria to achieve automatic junk image filtering. One promising solution for these difficulties is to allow the user to interactively provide additional information (i.e., his/her query intentions and personal preferences) for generating better hypotheses for junk image filtering. Unfortunately, most existing relevance feedback techniques require users to label a reasonable number of returned images into the relevant class and the irrelevant class, which may bring huge labeling burden on the users and may further stop them to use such tools for junk image filtering.

In order to allow users to express their query intentions more precisely and assess effectiveness of the underlying hypotheses for junk image filtering, it is very important to support similarity-based visualization of large amounts of returned images [36-43], so that users can see the margins between the relevant images and the junk images at the first glance. To generate more precise visualization of large amounts of returned images, it is very important to create a good partition of large amounts of returned images according to their nonlinear visual similarity contexts.

V. Junk Image Filtering

In this paper, we have developed an incremental kernel learning algorithm to determine the optimal values of the importance factors for kernel combination (i.e., mixture-of-kernels construction) by taking the users' query intentions and personal preferences into consideration: (1) The mixture-of-kernels for diverse image similarity characterization is initialized by maximizing the margins between the majority group of the returned images (i.e., the returned images which dominate the visual properties) and the outliers (i.e., obvious junk images which their visual properties are significantly different from the dominant visual properties). An interactive interface is designed to visualize the returned images according to their nonlinear visual similarity contexts, which can allow users to assess the relevance between the returned images and their query intentions more effectively and express their query intentions more precisely. (2) An incremental kernel learning algorithm is developed to translate and incorporate the users' query intentions and personal preferences for updating the mixture-of-kernels and learning the SVM classifier incrementally to filter out the junk images more effectively. (3) The updated mixture-of-kernels is used to create more accurate clustering of the returned images and achieve more precise visualization of the returned images for next hypothesis making loop.

A. Image Clustering

Our online junk image filtering system is implicitly connected with the keyword-based Google Images search engine, thus users are allowed to type in keywords to start their goal of image search and the returned images for a given keyword-based query are automatically obtained by the Google Images search engine. For the given keyword-based query, our online junk image filtering system can download 200 returned images automatically from Google Images. Obviously, our system can also allow users to define the number of returned images they want to look for according to their personal preferences. To reduce the computational cost for feature extraction, we use the thumbnails from Google Images instead of the original-size images for feature extraction.

The user-independent hypothesis (i.e., hidden hypothesis) for junk image filtering is that the returned images for a given keyword-based query can be partitioned into two groups: (a) *majority group of the returned images* which dominate the visual properties of the returned images and locate inside a cluster sphere; (b) *outliers* which their visual properties are significantly different from the dominant visual properties of the returned images and locate outside the cluster sphere.

One-class SVM algorithm [30] is used to model such hidden hypothesis by determining a smallest enclosing sphere of radius R to cover the majority group of the returned images for the given keyword-based query:

$$\forall_{j=1}^{N}: \quad \|\phi(x_j) - \mu^{\phi}\|^2 \leq R^2 \tag{5}$$

where N is the total number of images returned by the given keyword-based image query, μ^{ϕ} is defined as the center of the majority group of the returned images,

$$\mu^{\phi} = \sum_{j=1}^{N} \phi(x_j) \tag{6}$$

To enhance its robustness to the outliers, soft constraints are incorporated by adding slack variables:

$$\forall_{j=1}^{N}: \quad \|\phi(x_j) - \mu^{\phi}\|^2 \leq R^2 + \xi_j, \qquad \xi_j \geq 0 \tag{7}$$

Thus the problem for incorporating one-class SVM algorithm for image clustering can be defined as:

$$min \left\{ R^2 + \frac{C}{N} \sum_{j=1}^{N} \xi_j \right\} \tag{8}$$

subject to:

$$\forall_{j=1}^{N}: \quad \|\phi(x_j) - \mu^{\phi}\|^2 \leq R^2 + \xi_j, \qquad \xi_j \geq 0$$

where C is a constant and $\frac{C}{N}\sum_{j=1}^{N} \xi_j$ is a penalty term.

We can solve this optimization problem with Lagrangian multipliers:

$$L = R^2 - \sum_j (R^2 + \xi_j - \|\phi(x_j) - \mu^{\phi}\|^2)\alpha_j - \sum \xi_j \lambda_j + \frac{C}{N} \sum \xi_j \tag{9}$$

where $\alpha_j \geq 0$ and $\lambda_j \geq 0$ are the Lagrangian multipliers.

The Lagrangian multipliers problem can be solved by:

$$\frac{\partial L}{\partial R} = \frac{\partial L}{\partial \xi_j} = \frac{\partial L}{\partial \mu^{\phi}} = 0 \tag{10}$$

which can lead to:

$$\sum_{j}^{N} \alpha_j = 1, \qquad \alpha_j = \frac{C}{N} - \lambda_j \tag{11}$$

$$\xi_j \lambda_j = 0, \qquad (R^2 + \xi_j - \|\phi(x_j) - \mu^{\phi}\|^2)\alpha_j = 0 \tag{12}$$

The dual form for the Lagrangian optimization problem can be re-written as:

$$max \left\{ \sum_{j}^{N} \alpha_j \kappa(x_j, x_j) - \sum_{i,j}^{N} \alpha_i \alpha_j \kappa(x_i, x_j) \right\} \tag{13}$$

subject to:

$$\forall_{j=1}^{N}: \quad 0 \leq \alpha_j \leq \frac{C}{N}, \qquad \sum_{j=1}^{N} \alpha_j = 1$$

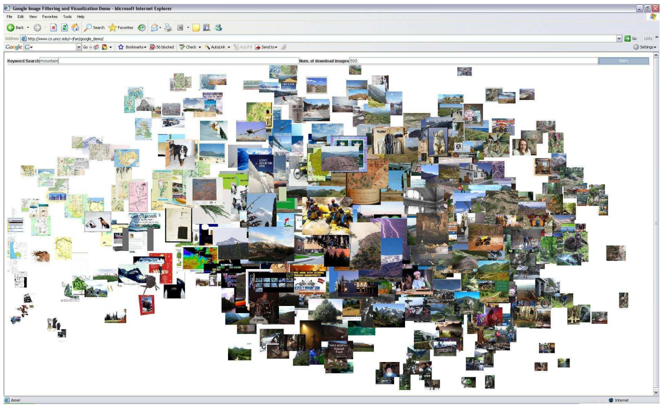
Thus the decision function (i.e., one-class SVM classifier) for image clustering can be determined as:

$$f(x) = R^2 - \sum_{i,j}^{N} \alpha_i \alpha_j \kappa(x_i, x_j) + 2\sum_{j}^{N} \alpha_j \kappa(x_j, x) - \kappa(x, x) \tag{14}$$

We can further define the distance between an image with the visual features x and the center μ^{ϕ} of the cluster sphere:

$$R^2(x) = \kappa(x, x) - 2\sum_{j}^{N} \alpha_j \kappa(x_j, x) + \sum_{i,j}^{N} \alpha_i \alpha_j \kappa(x_i, x_j) \tag{15}$$

$$\kappa(x_i, x_j) = \sum_{i=1}^{3} \beta_i K_i(x_i, x_j), \qquad \sum_{i=1}^{3} \beta_i = 1 \tag{16}$$


Google Image Filtering and Visualization Demo - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://www.cs.uncc.edu/~jfan/google_demo/
Go Links
Google Go Bookmarks 56 blocked Check AutoLink AutoFill Send to Settings
100%

Keyword Search mountain

Num. of download images 500

Fig. 3. The obvious junk images for the keyword-based query “mountain” are separated effectively from the majority group of the returned images and are projected on the left-upon corner.

The image with the visual features x, which locates on the surface of the cluster sphere (i.e., $R^2(x) = R^2$), is treated as a *support vector*. On the other hand, the image with the visual features z is treated as the outlier (i.e., $O(z) = 1$):

$$O(z) = \begin{cases} 1, & R^2(z) > R^2 \\ 0, & otherwise \end{cases} \tag{17}$$

where R is the radius of the smallest enclosing cluster sphere to cover the majority group of the returned images for the given keyword-based query.

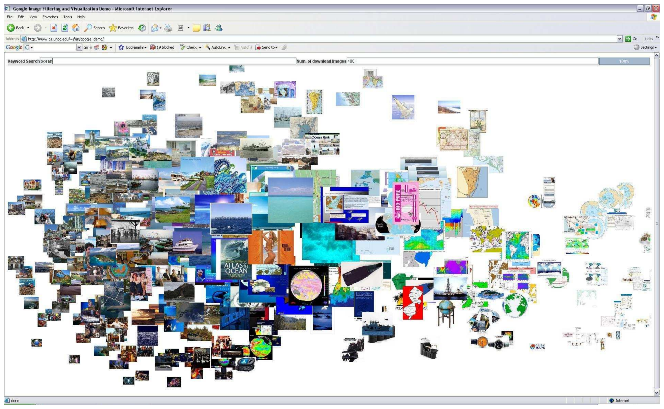
A good combination of these three basic image kernels (i.e., the mixture-of-kernels with the optimal values of these three importance factors β) should be able to achieve more accurate approximation of the diverse visual similarity contexts between the images and result in better separation between the majority group of the returned images and the outliers. Thus the optimal values of the importance factors β for an initial combination of these three basic image kernels (i.e., without considering the users' query intentions and personal preferences) can be obtained by maximizing the margin between the outliers and the majority group of the returned images:

$$\underset{\beta}{max} \left\{ \sum_{l=1}^{T} min \left[\kappa(z_l, x_i), R^2(z_l) > R^2, R^2(x_i) = R^2 \right] \right\} \tag{18}$$

where T is the total number of outlying images, Ω is the set of the support vectors (i.e., the returned images locate on the boundary of the cluster sphere).

After the optimal values for initial combination of these three basic image kernels are obtained, the corresponding mixture-of-kernels is used to create a good partition of the returned images and generate an initial visualization of the returned images, so that users can assess the correctness and the effectiveness of the underlying hypothesis for junk image filtering.

The returned images in the majority group, which have been separated from the outliers (i.e., obvious junk images), are further partitioned into multiple clusters according to their nonlinear visual similarity contexts. In this paper, we have seamlessly integrated one-class SVM algorithm [30] with the kernel K-means algorithm [29] to achieve better partitions of the returned images in the majority group. Our algorithm takes the following major steps: (a) We assume there are τ



Google Image Filtering and Visualization Demo - Microsoft Internet Explorer

Address http://www.cs.uncc.edu/~jfan/google_demo/

Keyword Search ocean

Num. of download images 400

100%

Fig. 4. The obvious junk images for the keyword-based query "ocean" are separated effectively from the majority group of the returned images and are projected on the left-down corner.

clusters for the returned images in the majority group, the centers and the radiuses for these τ clusters are denoted as μ_l^ϕ, R_l, $l = 1, \cdots, \tau$. Ideally, the optimal number of image clusters should be determined by an iterative process, but such iterative process could be computation-sensitive and it is unpractical for enabling online junk image filtering. Thus we set the maximum number of image clusters as $\tau = 10$ heuristically for all the queries, and good performance can be obtained for most ambiguous search terms (ambiguous search terms may have more clusters for different word senses) in our experiments. (b) The optimal partition of the returned images in the majority group is obtained by minimizing the trace of the within-cluster scatter matrix, S_w^ϕ. The scatter matrix is given by:

$$S_w^\phi = \frac{1}{N} \sum_{l=1}^{\tau} \sum_{i=1}^{N} \pi_{li} R_l^2(x_i) \tag{19}$$

where π_{li} is the membership parameter, $\pi_{li} = 1$ if $x_i \in C_l$ and 0 otherwise, C_l is the lth image cluster. $R_l^2(x_i)$ is the radius of the smallest enclosing cluster sphere for the lth cluster and it is defined as:

$$R_l^2(x_i) = \kappa(x_i, x_i) - \frac{2}{N_l} \sum_{j=1}^{N} \pi_{lj} \kappa(x_i, x_j) + \frac{1}{N_l^2} \sum_{j=1}^{N} \sum_{m=1}^{N} \pi_{lj} \pi_{lm} \kappa(x_j, x_m) \tag{20}$$

where $N_l = \sum_{j=1}^{N} \pi_{lj}$. Searching the optimal values of the membership parameters π that minimizes the expression of the trace in Eq. (19) can be achieved effectively by performing kernel K-means algorithm [29].

Partitioning the returned images into multiple clusters according to their nonlinear visual similarity contexts can gain several benefits: (1) It can provide a good global overview of large amounts of returned images (i.e., image clusters and their similarity contexts), which may reveal the interesting or even unexpected distribution trends of the returned images. Thus users can assess the relevance between the returned images and their real query intentions more effectively. (2) It can provide a good insight of large amounts of returned images and achieve more precise visualization of the images with better preservation of both their global similarity structures (i.e., image clusters and their similarity contexts) and their local similarity structures (i.e., nonlinear visual similarity contexts between the returned images in the same cluster). (3) It can support automatic label propagation and reduce the users' efforts significantly for labeling the images to enable incremental kernel learning.

B. Image Projection and Visualization

To incorporate image visualization for assisting users on relevance assessment and hypothesis making, it is very important to develop new visualization algorithms that are able to exploit and preserve the nonlinear visual similarity contexts between the returned images. In this paper, locality preserving projection is used to preserve the nonlinear visual similarity contexts for image projection and visualization [35].

The local similarity matrix Ξ for these N returned images can be obtained by calculating the kernel-based similarity distance between the images and its component Ξ_{ij} is used to characterize the local visual similarity context between two returned images with the visual features x_i and x_j:

$$\Xi_{ij} = \phi(x_i)^T \phi(x_j) = \kappa(x_i, x_j) = \sum_{l=1}^{3} \beta_l K_l(x_i, x_j) \tag{21}$$

By integrating multiple kernels for image similarity characterization, our mixture-of-kernels algorithm can discover the nonlinear visual similarity structures between the images effectively.

Given a set of returned images and their visual features $X = \{x_1, x_2, \cdots, x_N\}$, let A be the transformation matrix and $Y = \{y_1, y_2, \cdots, y_N\}$ be the projection locations of these N returned images on 2-D display screen, we can have the following transformation:

$$y_i = A^T \phi(x_i) \tag{22}$$

Given the local similarity matrix Ξ for all these N returned images, the optimal projection can be obtained by solving the following minimization problem:

$$\begin{aligned} A_{optimal} &= \underset{A}{argmin} \left\{ \sum_{i,j}^{N} (y_i - y_j)^2 \Xi_{ij} \right\} \\ &= \underset{A}{argmin} \left\{ \sum_{i,j}^{N} (A^T\phi(x_i) - A^T\phi(x_j))^2 \Xi_{ij} \right\} \\ &= \underset{A}{argmin} \; A^T \phi(X) \Delta \phi^T(X) A \end{aligned} \tag{23}$$

where Δ is the graph Laplacian,

$$\Delta = D - \Xi, \qquad D_{ii} = \sum_{j}^{N} \Xi_{ij} \tag{24}$$

where D is a diagonal matrix. The optimal solution for Eq. (23) is to ensure that if two images with the visual features x_i and x_j are close in the high-dimensional feature space and their projection locations y_i and y_j are also close each other on 2-D display screen. Thus the algorithms for kernel-based image clustering and locality preserving projection are integrated seamlessly for preserving both the global similarity structures and the local similarity structures between the images, and a nearest neighbor search in the 2D image display space will obtain the same results in the high-dimensional feature space. By integrating the mixture-of-kernels and the locality preserving projection for image clustering and projection, our algorithm can exploit and preserve both the nonlinear image similarity structures (i.e., local visual similarity structures between the returned images in the same cluster) and the image distribution manifolds (i.e., image clusters and their similarity contexts) effectively.

The transformation vector $\overrightarrow{a}$ (i.e., components for the transformation matrix A) for context-preserving image projection, which minimizes the objective function in Eq.(23), is given by the minimum eigenvalue solution to the generalized eigenvalue problem:

$$\phi(X)\Delta\phi^T(X)\overrightarrow{a} = \lambda\phi(X)D\phi^T(X)\overrightarrow{a} \qquad (25)$$

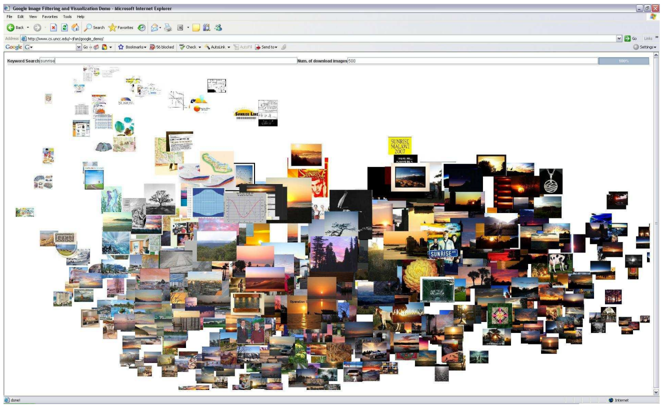
which can further be refined as:

$$\Delta\Xi\overrightarrow{a} = \lambda D\Xi\overrightarrow{a} \qquad (26)$$

The returned images, which are projected by preserving their nonlinear visual similarity contexts, are further laid out on the hyperbolic plane to enable interactive image navigation and exploration. After such context-preserving projection of the returned images is obtained, Poincaré disk model [28] is used to map the returned images on the hyperbolic plane onto a 2D display coordinate to support change of focus and interactive image exploration. Formally, if let φ be the hyperbolic distance of one given image to the center of the hyperbolic plane and ψ be the Euclidean distance of the same image to the center of the display unit circle, the relationship between their derivative is described by:

$$d\varphi = \frac{2s}{1-\psi^2} \cdot d\psi \qquad (27)$$

where s is the scaling factor. Intuitively, the Poincaré mapping makes a unit Euclidean distance correspond to a longer hyperbolic distance as it approaches the rim of the display unit circle. In other words, if the images are of fixed size, they would appear larger when they are closer to the origin of the display unit circle and smaller when they are further away. This property makes it very suitable for visualizing large amounts of returned images because the non-uniformity distance mapping creates an emphasis for the returned images which are in current focus, while de-emphasizing those returned images that can further form the focus point.

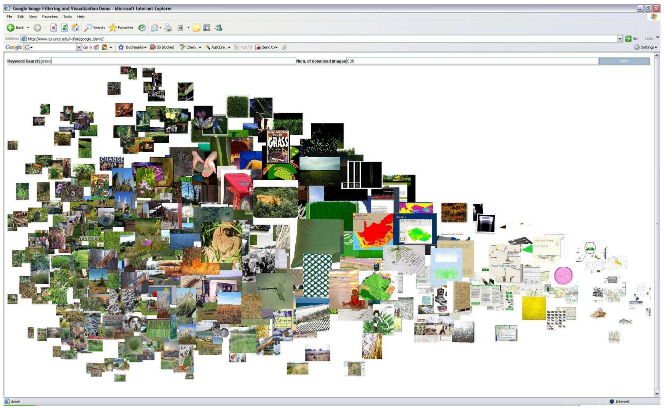

Google Image Filtering and Visualization Demo - Microsoft Internet Explorer

Address http://www.cs.uncc.edu/~jfan/google_demo/

Keyword Search sunrise

Num. of download images 500

Fig. 5. The obvious junk images for the keyword-based query “sunrise” are separated effectively from the majority group of the returned images and are projected on the left-upon corner.


Google Image Filtering and Visualization Demo - Microsoft Internet Explorer

Address http://www.cs.uncc.edu/~jfan/google_demo/

Keyword Search grass

Num. of download images 500

Fig. 6. The obvious junk images for the keyword-based query “grass” are separated effectively from the majority group of the returned images and are projected on the left-down corner.

Our hyperbolic visualization of the returned images for the keyword-based queries “mountain”, “sunrise”, “grass”, and “ocean” are given in Fig. 3, Fig. 4, Fig. 5 amd Fig. 6, where the returned images for multiple clusters are layouted by using context-preserving projection and Poincaré mapping. From these experimental results, one can conclude that: (a) The obvious junk images (i.e., outliers) can be separated effectively from the majority group of the returned images, thus our mixture-of-kernels algorithm can characterize the diverse visual similarity contexts between the returned images more precisely with higher discrimination power and our hidden hypothesis is correct for filtering out the obvious junk images effectively. (b) Our hyperbolic image visualization algorithm can allow users to see the global distribution structures (i.e., image clusters and the similarity contexts) of large amounts of returned images and their nonlinear visual similarity contexts at the first glance. Such global image distribution structures can help users assess the relevance between the returned images and their personal query intentions effectively. (c) The change of focus is implemented for allowing users to navigate and explore the returned images interactively according to their nonlinear visual similarity contexts. Users can change their focuses of the returned images by clicking on any visible image to bring it into focus at the screen center, or by dragging any visible image interactively to any other screen location without losing their nonlinear visual similarity contexts, where the rest of the images can be transformed appropriately. With the power of high interaction and rapid response for exploring and navigating large amounts of returned images according to their nonlinear visual similarity contexts, our hyperbolic image visualization scheme can support more effective solution for users to assess the relevance between the returned images and their real query intentions interactively.

C. *Incremental Kernel Learning*

The initial combination of these three basic image kernels may not be discriminative enough to characterize the diverse visual similarity contexts between the returned images accurately and filter out all the junk images effectively. In addition, such initial mixture-of-kernels has not taken the user's query intentions into consideration, thus it may not be able to filter out all the junk images which may depend on the user's query intentions and personal preferences.

Through interactive navigation and exploration of large amounts of returned images according to their nonlinear visual similarity contexts, users can build up their mental models about which types of returned images they want to look for (i.e., which image clusters are more relevant to their query intentions) and what are the most significant visual similarity contexts between the returned images. After the users find some images of interest via interactive image exploration, our system can allow users to zoom into the images of interest. When the users find some particular images via zooming into the images of interest, they can select one or multiple images to express their query intentions precisely as shown in Fig. 7(a), Fig. 8(a), Fig. 9(a), and Fig. 10(a). After our system capture such users' query intentions, it can automatically update the underlying mixture-of-kernels and change the hypotheses for junk image filtering according to the users' personal preferences. It is worth noting that our system just requires users to click one or few images to express their query intentions and automatic label propagation is used to obtain large amounts of high-quality images for supporting incremental kernel learning. Because the returned images are clustered according to their nonlinear visual similarity contexts, clicking one single image on the screen can obtain large amounts of visually-similar images in the same cluster, which can reduce the users' labeling efforts significantly for enabling incremental kernel learning. If these clicked images belong to the same cluster, the visually-similar returned images for the corresponding cluster will be treated as the relevant images for enabling incremental kernel learning. If these clicked images belong to multiple clusters, the returned images from all these corresponding clusters will be treated as the relevant images for supporting incremental kernel learning.

In this paper, we have developed an incremental kernel learning algorithm to take advantage of users' query intentions and personal preferences for determining more accurate combination of these three basic image kernels (i.e., generating better hypotheses for junk image filtering). The user's query intentions are represented precisely by a set of visually-similar images, which are in the same clusters with the clicked images. These visually-similar images are treated as the relevant images to update the underlying mixture-of-kernels and learn more accurate SVM classifier incrementally for filtering out the junk images according to the users' query intentions and personal preferences.

The junk image filter $f_p(x)$ (i.e., SVM classifier to partition the returned images into the relevant class and the irrelevant class) for the previous hypothesis making loop is defined as:

$$f_p(x) = W_0^T \phi(x) + b \tag{28}$$

where the regularization term W_0 is learned from the boundary images (i.e., which are treated as the support vectors for image clustering in the previous hypothesis making loop), (x_i, y_i), $i = 1, \cdots, L$.

$$W_0 = \sum_{i=1}^{L} \alpha_i^* y_i \phi(x_i) \tag{29}$$

For a given keyword-based query, the SVM classifier for the current hypothesis making loop can be learned incrementally [32]:

$$min \left\{ \frac{1}{2}\|W - W_0\|^2 + \epsilon \sum_{l=1}^{m} [1 - y_l(W^T \cdot \phi(x_l) + b)] \right\} \tag{30}$$

where W_0 is the regularization term for the previous hypothesis making loop, (x_l, y_l), $l = 1, \cdots, m$ are the new training images for the current hypothesis making loop which are obtained via automatic label propagation, ϵ is the penalty term.

The dual problem for Eq. (30) is solved by:

$$min \left\{ \frac{1}{2} \sum_{l=1}^{m} \sum_{h=1}^{m} \alpha_l \alpha_h y_l y_h \kappa(x_l, x_h) - \right.$$
$$\left. \sum_{l=1}^{m} \alpha_l \left(1 - y_l \sum_{i=1}^{L} \alpha_i^* y_i \kappa(x_i, x_l) \right) \right\} \tag{31}$$

subject to:

$$\forall_{l=1}^{m} : 0 \leq \alpha_l \leq \epsilon, \qquad \sum_{l=1}^{m} \alpha_l y_l = 0$$

The optimal solution of Eq. (30) satisfies:

$$W = W_0 + \sum_{l=1}^{m} \alpha_l^* y_l \phi(x_l) = \sum_{i=1}^{L} \alpha_i^* y_i \phi(x_i) + \sum_{l=1}^{m} \alpha_l^* y_l \phi(x_l) \tag{32}$$

where α^* is the optimal value of the weighting factors of the images to optimize the Eq.(31). Thus the SVM classifier $f_c(x)$ for the current hypothesis making loop can be updated incrementally as:

$$f_c(x) = W^T \phi(x) + b = \sum_{i=1}^{L} \alpha_i^* y_i \kappa(x, x_i) + \sum_{l=1}^{m} \alpha_l^* y_l \kappa(x, x_l) + b \tag{33}$$

The SVM classifier, which is learned incrementally by taking the user's query intentions into consideration, is then used to re-partition the returned images into two classes: relevant images *versus* irrelevant images, where the irrelevant images (i.e., junk images according to the user's current query intentions) either can be highlighted automatically or may not be presented to the users. Because the hypothesis for junk image filtering is changed adaptively according to the user's current query intentions, some returned images which are detected as the irrelevant images in the previous hypothesis making loop may become as the relevant images in the current hypothesis making loop. On the other hand, some returned images which are detected as the irrelevant images in the previous hypothesis making loop may become as the relevant images in the current hypothesis making loop because of the change of the user's query intentions and the updating of the decision boundary of the SVM classifier.

In such incremental SVM classifier learning process, the underlying mixture-of-kernels for diverse image similarity

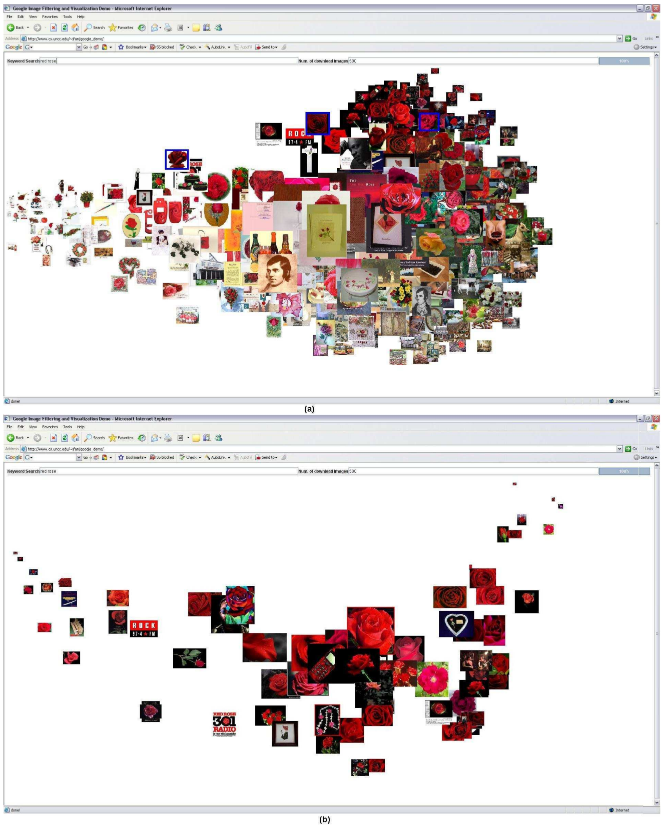

Google Image Filtering and Visualization Demo - Microsoft Internet Explorer
Keyword Search red rose
Num. of download images 500
(a)
(b)

Fig. 7. Junk image filtering: (a) the images returned by the keyword-based search "red rose" and the images in blue boundaries are selected by the user to express his/her query intentions; (b) the filtered images after the first hypothesis making loop.

characterization is also changed adaptively according to the user's query intentions and personal preferences. To obtain the updating rule of the importance factors β for these three basic image kernels, the objective function $J(\beta)$ is defined as:

$$J(\beta) = \frac{1}{2}\sum_{l=1}^{m}\sum_{h=1}^{m}\alpha_l^*\alpha_h^* y_l y_h \sum_{i=1}^{3}\beta_i K_i(x_l, x_h) -$$

$$\sum_{l=1}^{m}\alpha_l^*\left(1 - y_l\sum_{i=1}^{L}\alpha_i^* y_i \sum_{i=1}^{3}\beta_i K_i(x_i, x_l)\right) \qquad (34)$$

For computing the derivatives of $J(\beta)$ with respect to β, we assume that the optimal value of α^* does not depend on β. Thus the derivatives of the objective function $J(\beta)$ can be computed as:

$$\forall_{i=1}^{3}: \quad \frac{\partial J(\beta)}{\partial \beta_i} = \frac{1}{2}\sum_{l=1}^{m}\sum_{h=1}^{m}\alpha_l^*\alpha_h^* y_l y_h K_i(x_l, x_h) +$$

$$\sum_{l=1}^{m}\sum_{i=1}^{L}\alpha_l^*\alpha_i^* y_l y_i K_i(x_i, x_l) \qquad (35)$$

The objective function $J(\beta)$ is convex and thus our gradient method for computing the derivatives of $J(\beta)$ can guarantee to converge. In addition, the importance factors β for these three basic image kernels are updated while ensuring that the constraints on β are satisfied [44].

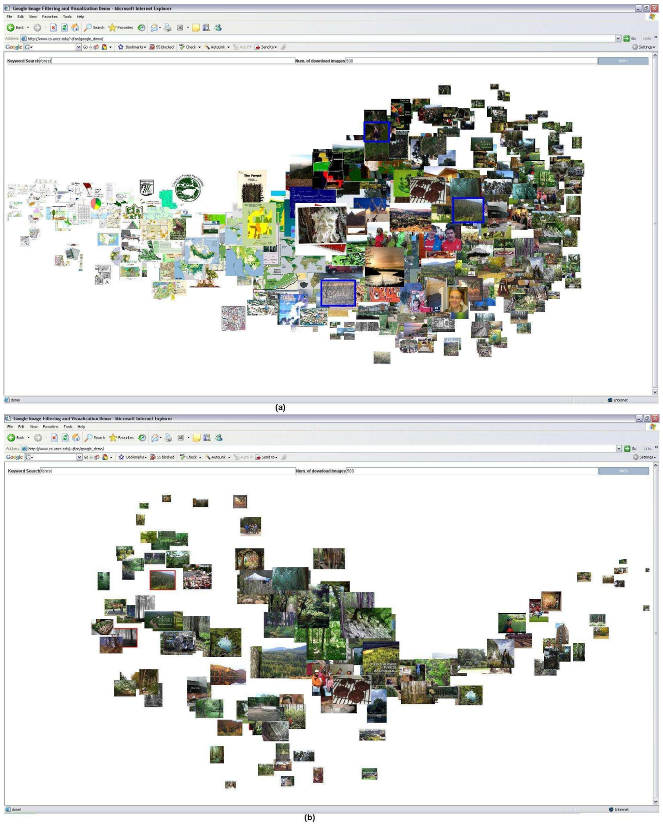

Google Image Filtering and Visualization Demo - Microsoft Internet Explorer
Keyword Search forest
Num. of download images 500
(a)
(b)

Fig. 8. Junk image filtering: (a) the images returned by the keyword-based search "forest" and the images in blue boundaries are selected by the user to express his/her query intentions; (b) the filtered images after the first hypothesis making loop.

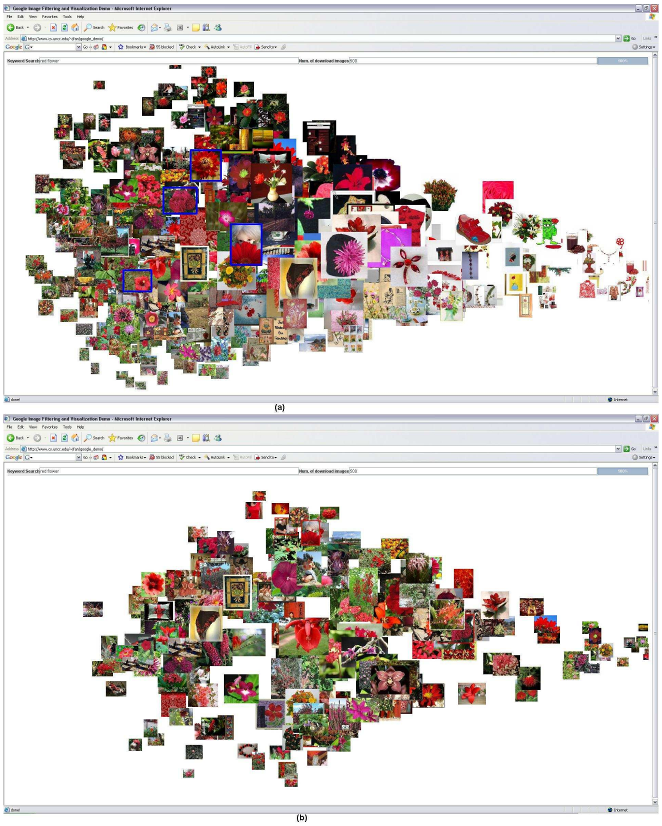
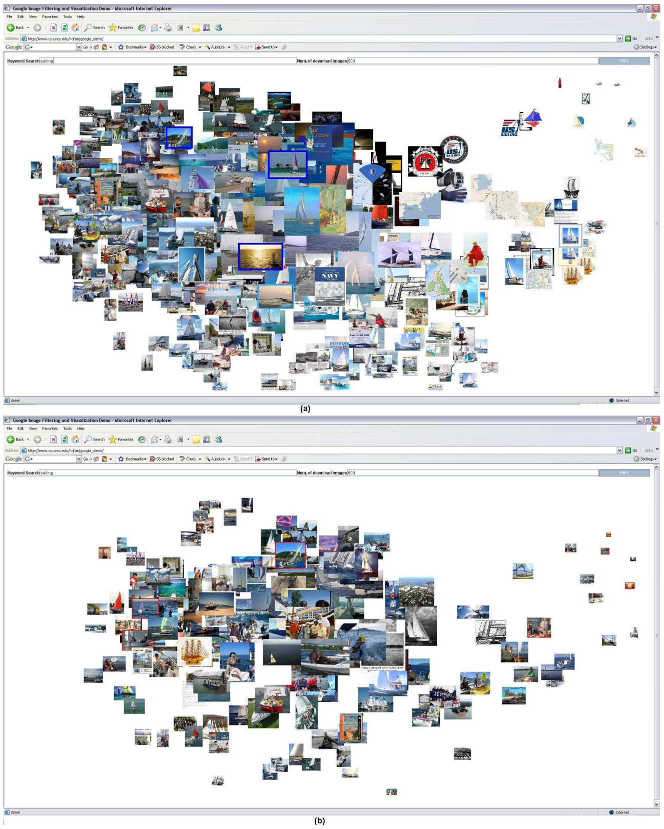

Google Image Filtering and Visualization Demo - Microsoft Internet Explorer
Keyword Search red flower
Num. of download images 500
(a)
(b)

Fig. 9. Junk image filtering: (a) the images returned by the keyword-based search "red flower" and the images in blue boundaries are selected by the user to express his/her query intentions; (b) the filtered images after the first hypothesis making loop.


Google Image Filtering and Visualization Demo - Microsoft Internet Explorer

Address: http://www.cs.uncc.edu/~jfan/google_demo/

Keyword Search: sailing

Num. of download images: 500

Fig. 10. Junk image filtering: (a) the images returned by the keyword-based search “sailing” and the images in blue boundaries are selected by the user to express his/her query intentions; (b) the filtered images after the first hypothesis making loop.

The importance factors β for these three basic image kernels are updated as:

$$\forall_{i=1}^{3}: \quad \beta_i^{t+1} = \beta_i^t + \gamma_t \left[\frac{1}{2} \sum_{l=1}^{m} \sum_{h=1}^{m} \alpha_l^* \alpha_h^* y_l y_h K_i(x_l, x_h) + \sum_{l=1}^{m} \sum_{j=1}^{L} \alpha_l^* \alpha_i^* y_l y_j K_i(x_j, x_l) \right] \qquad (36)$$

where γ_t is the step size for the ith hypothesis making loop, β^{t+1} and β^t are the importance factors for the current and previous hypothesis making loops. The step size γ_t is selected automatically with proper stopping criterion to ensure global convergence [33]. Our incremental kernel learning algorithm is performed until a stopping criterion is met. This stopping criterion can be either based on a maximal number of hypothesis making loops or the variation of β between two consecutive steps.

The updated mixture-of-kernels (i.e., new combination of these three basic image kernels) can characterize the nonlinear visual similarity contexts between the returned images more precisely according to the user’s query intentions and personal preferences, filter out the junk images more effectively and generate more precise visualization of the returned images for next hypothesis making loop. Because the user’s query intentions are interpreted precisely by using the clicked images and their visually-similar images in the same clusters, the relevant images can be highlighted explicitly. The irrelevant images, which have bigger margins with the clusters for the clicked images, are treated as the junk images and are filtered out automatically. As shown in Fig. 11 and Fig. 12, the effectiveness of our incremental kernel learning algorithm can be improved sequentially (i.e., the relevant images and the irrelevant images are more separable) by incorporating the user’s query intentions and personal preferences to learn more accurate SVM classifier for junk image filtering. From these examples, one can observe that most junk images can be filtered out effectively after few hypothesis making loops.

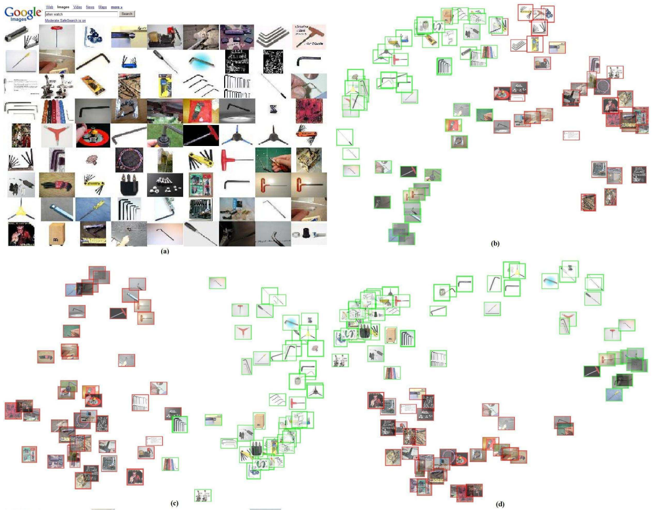


Fig. 11. Hypotheses visualization and assessment for filtering the junk images for the keyword-based query “Allen Watch”: (a) Images returned by Google Images; (b) visualization of the returned images and the hypothesis for junk image filtering (i.e., margin between the relevant images and the irrelevant images) at the first hypothesis making loop; (c) visualization of the returned images and the hypothesis at the second hypothesis making loop; (d) visualization of the returned images and the hypothesis at the third hypothesis making loop.

Our hyperbolic image visualization algorithm, which layouts the returned images according to their nonlinear visual similarity contexts, can allow users to see the margins between the relevant images and the irrelevant images and provide more understandable approach to allow the users to assess the effectiveness of the underlying hypotheses for junk image filtering interactively. Users can easily express their query intentions and personal preferences by selecting one or multiple images interactively to generate better hypotheses and learn more accurate SVM classifier for junk image filtering.

In Fig. 7(b), Fig. 8(b), Fig. 9(b) and Fig. 10(b), the junk image filtering results for several keyword-based queries are given. From these experimental results, one can observe that our system can filter out the junk images effectively. In order to enable social evaluation of our online junk image filtering system, we have released our online system at: ***http://www.cs.uncc.edu/∼jfan/google_demo/***.

It is worth noting that our interactive approach for junk image filtering is significantly different from traditional relevance feedback approaches for image retrieval [6-13]. The traditional relevance feedback approaches require users to label a reasonable number of images into the relevant class or the irrelevant class for learning the modified queries reliably, thus they may bring huge burden on users. In contrast, our interactive junk image filtering approach can obtain large amounts of labeled images easily via automatic label propagation because the returned images are clustered into multiple groups according to their nonlinear visual similarity contexts. Thus clicking one single image on the screen can obtain large amounts of visually-similar images in the same cluster to enable incremental kernel learning and lessen the labeling burden on the users significantly. Most existing relevance feedback approaches use page-based ranked list for displaying the returned images, and the nonlinear visual similarity contexts between the returned images are completely ignored. Thus it is very hard for users to assess the relevance between the returned images and their real query intentions effectively. In contrast, our interactive junk image filtering approach can allow users to see large amounts of returned images and their nonlinear visual similarity contexts at the first glance, thus users can obtain more significant insights, assess the relevance between the returned images and their real query intentions more effectively, and express their query intentions and personal preferences more precisely to generate better hypotheses and learn more accurate SVM classifier for junk image filtering.

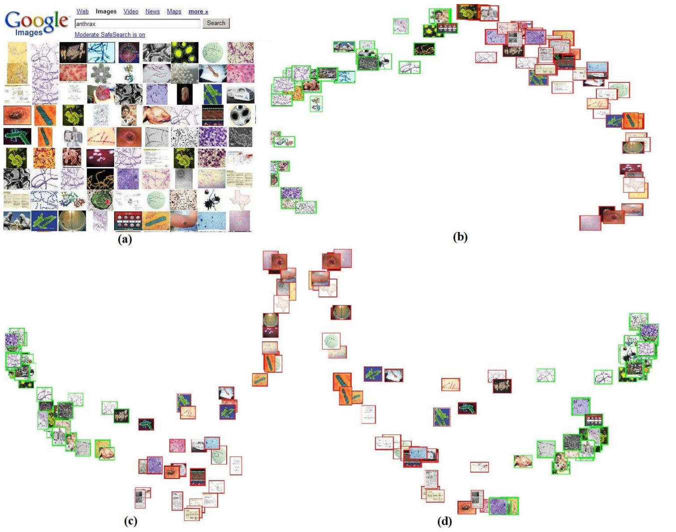


Fig. 12. Hypotheses visualization and assessment for filtering the junk images for the keyword-based query "anthrax": (a) Images returned by Google Images; (b) visualization of the returned images and the hypothesis for junk image filtering (i.e., margin between the relevant images and the irrelevant images) at the first hypothesis making loop; (c) visualization of the returned images and the hypothesis at the second hypothesis making loop; (d) visualization of the returned images and the hypothesis at the third hypothesis making loop.

VI. Algorithm and System Evaluation

In order to evaluate our interactive approach for junk image filtering, we generate the potential query concepts (list of potential keywords) by randomly sampling the keywords from WordNet. However, in WordNet, a lot of keywords are rarely used in our daily life, thus we manually remove those uncommon keywords after random sampling and obtain 2000 commonly-used keywords for our experiments. These 2000 keywords are used as the query concepts to search the images from Google Images search engine. The keywords, which are used to interpret the image concepts at the higher levels of WordNet and have multiple children nodes with different word senses, are treated as the ambiguous text terms to evaluate the performance of our algorithms for image clustering, context-preserving projection and visualization. For a given keyword-based query, our system can automatically download the returned images from Google Images search engine and users are allowed to define the number of returned images they want to look for. In our experiments, 500 returned images are downloaded from keyword-based Google Images search engine for each query.

Our works on algorithm and system evaluation focus on: (a) evaluating the convergence of our incremental kernel learning algorithm; (b) evaluating the accuracy improvement of our incremental SVM classifier when the number of hypothesis making loops increases; (c) evaluating the prediction power of our SVM classifiers for 2000 query concepts (i.e., image concepts) by using large amounts of unseen images; (d) evaluating the computational costs of our algorithms for kernel-based image clustering, context-preserving image projection, and incremental kernel learning for junk image filtering; (e) comparing the performance of junk image filtering algorithm with the keyword-based Google Images search engine.

To evaluate the effectiveness of our algorithms of mixture-of-kernels and incremental kernel learning, the accuracy of the SVM classifiers for different hypothesis making loops is calculated. Given the *confusion matrix* δ for image clustering, the accuracy is definedas:

$$Accurarcy = \frac{\sum_{i=1}^{c} \delta(i,i)}{\sum_{i=1}^{c} \sum_{j=1}^{c} \delta(i,j)} \qquad (37)$$

where $c = 2$ is the number of clusters (i.e. relevant *versus* irrelevant clusters). As shown in Fig. 13, the performance of our junk image filtering algorithm (i.e., the performance of the underlying SVM classifiers for different hypothesis making loops) can generally improve with the number of hypothesis making loops, but it becomes stable after 4 hypothesis making loops. Thus our proposed algorithm for incremental kernel learning has very good convergence property.

On average, our interactive approach for junk image filtering can achieve over 75% accuracy for image retrieval by filtering out the junk images from Google Images search results. Compared to the original 58% average accuracy of Google Images search engine, our interactive approach for junk image filtering can achieve a significant improvement on the search results.

To evaluate the generalization performance of the mixture-of-kernels which is learned incrementally, the optimal mixture-of-kernels (which are obtained after 5 hypothesis making loops) is used to train the SVM image classifiers for the relevant image concepts (i.e., the keywords for query formulation). The *benchmark metric* for classifier evaluation includes *precision* ρ and *recall* ϱ. They are defined as:

$$\rho = \frac{\theta}{\theta + \varpi}, \qquad \varrho = \frac{\theta}{\theta + \eta} \qquad (38)$$

where θ is the set of true positive images that are related to the corresponding image concept and are classified correctly, ϖ is the set of true negative images that are irrelevant to the

TABLE I

The accuracy comparison for Google Images search engine with and without performing our online junk image filtering. ρ and ϱ are precision and recall with SVM classifiers for junk image filtering, $\bar{\rho}$ and $\bar{\varrho}$ are precision and recall for Google Images search engine.

query	ρ	ϱ	$\bar{\rho}$	$\bar{\varrho}$
blood	16.7%	100%	15.5%	98%
mahogany	39.5%	100%	36.2%	96.8%
lamp	32.5%	79.2%	31.9%	78.3%
screw	44.3%	88.5%	42.8%	80.2%
amati	21.2%	64.0%	20.2%	61.1%
africa	40.7%	76.5%	38.5%	73.2%
african	33.3%	65.0%	32.8%	64.5%
apc	56.3%	87.5%	55.8%	86.5%
laurel	70.6%	100%	69.5%	98.3%
badger	71.0%	100%	69.5%	98.3%
besseya	95.2%	100%	75.0%	80.0%
afghan	88.2%	92.0%	62.1%	73.2%
terrier	96.2%	100%	70.2%	78.1%
oil plan	96.6%	100%	73.8%	76.3%
22 karat gold	68.1%	70.0%	50.0%	48.0%
african grey	94.3%	96.2%	72.5%	73.2%
daisy	91.3%	92.6%	68.8%	68.6%
elephant	95.9%	96.6%	70.0%	68.5%
violet	88.3%	83.3%	60.2%	60.8%
aertex	62.8%	56.0%	30.9%	32.8%

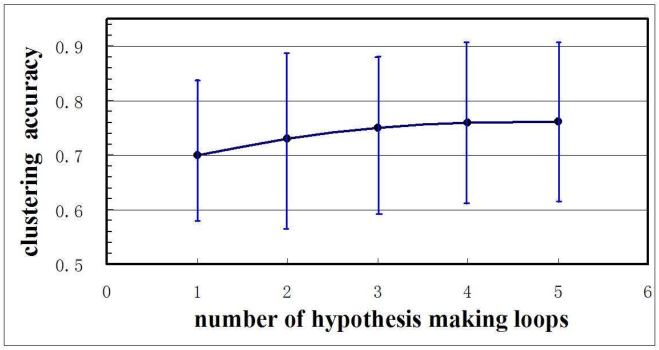


Fig. 13. Clustering accuracy as a function of the number of hypothesis making loops. The solid line represents the average clustering accuracy while the error bar shows the standard deviation over all 2000 keyword-based queries.

corresponding image concept and are classified incorrectly, and η is the set of false positive images that are related to the corresponding image concept but are misclassified. The performances of our SVM image classifiers are given in Table 1 for the 10 queries with most significant improvement of the performance and 10 queries with the least improvement of the performance among these 2000 query concepts. From this table, one can observe that the classification accuracies on large amounts of unseen images for different image concepts (queries) are significantly improved by performing incremental kernel learning. Our incremental kernel learning algorithm can also reduce the computational cost significantly as shown in Fig. 14.

To achieve kernel-based image clustering, the kernel matrix for the returned images should be calculated and the computational cost largely depends on the number of returned images. In our current system, we allow users to define the number of returned images they want to see. The computational cost for achieving kernel-based image clustering is approximated as $O(\tau N^3)$, where N is the total number of the returned images and τ is the number of image clusters. We have obtained

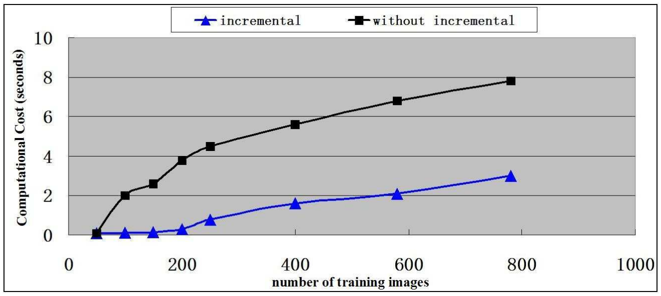


Fig. 14. The empirical relationship between the computational cost (seconds) and the number of training images for SVM classifier training.

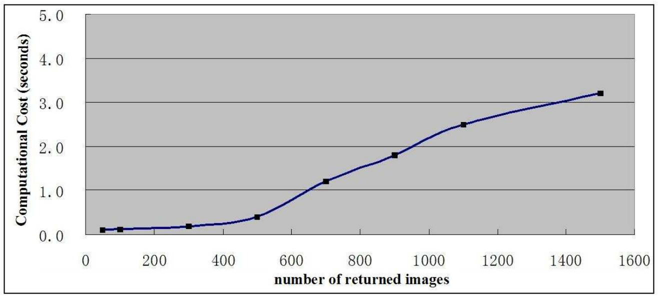


Fig. 15. The empirical relationship between the computational cost Ω_1 (seconds) and the number of returned images.

the empirical relationship between the computational cost Ω_1 (CPU time) and the number of returned images as shown in Fig. 15. One can observe that the computational cost Ω_1 increases exponentially with the number of returned images. When the number of returned images is less than 500, our system can almost achieve image clustering nearly in real time. It was reported that most people may just scan the first few pages of Google search results, thus it is reasonable to assume that people may just want to look for less than 500 returned images.

After the returned images are partitioned into multiple clusters via kernel-based clustering, our system can perform locality preserving projection (LPP) to obtain the similarity-preserving projection of the returned images on the hyperbolic plane. The computational cost for performing LPP is approximated as $O(\frac{N^2}{\tau})$, where N is the number of returned images for the given keyword-based query and τ is the number of image clusters. As shown in Fig. 16, we have obtained the empirical relationship between the computational cost Ω_2 and the number of returned images. One can observe that the computational cost Ω_2 exponentially increases with the number of returned images. When the number of returned images is less than 500, the computational cost Ω_2 is acceptable for supporting interactive image exploration and achieving junk image filtering nearly in real time.

VII. Conclusions

In this paper, we have proposed an incremental kernel learning algorithm to filter out large amounts of junk images from keyword-based Google Images search results. To achieve more accurate partition and visualization of the returned images, multiple kernels are seamlessly integrated for diverse image similarity characterization. A hyperbolic image visualization approach is incorporated to allow users to assess the relevance

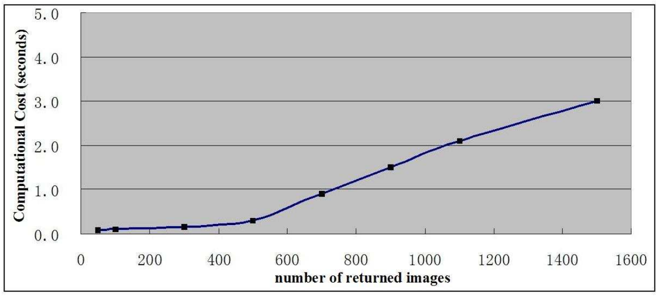


Fig. 16. The empirical relationship between the computational cost Ω_2 (seconds) and the number of returned images.

between the returned images and their real query intentions interactively and express their query intentions more precisely. To reduce the computational cost for junk image filtering, an incremental kernel learning algorithm is developed for SVM image classifier training by translating the users' query intentions to determine more accurate combination of these three basic image kernels, achieve more accurate characterization of diverse visual similarity contexts between the returned images, generate more accurate partition of the returned images, create more precise visualization of the returned images, and make better hypotheses for junk image filtering. Experiments on diverse keyword-based queries from Google Images search engine have obtained very promising results and our online junk image filtering system is also released on our web site for public evaluation.

VIII. Acknowledgment

The authors would like to thank the reviewers for their insightful comments and suggestions to make this paper more readable. We also want to thank the Associate Editor Dr. Yong Rui for his handling the review process of our paper. Jianping Fan is supported by National Science Foundation under 0601542-IIS and 0208539-IIS. Hangzai Luo is supported by Shanghai Pujiang Program under 08PJ1404600 and National Science Foundation of China under 60803077. Jinye Peng is supported by Program for New Century Excellent Talents in University(NCET-07-0693).

References

[1] R. Fergus, P. Perona, A. Zisserman, "A visual category filter for Google Images", Proc. ECCV, 2004.

[2] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, "Learning object categories from Google's image search", Proc. IEEE CVPR, 2006.

[3] D. Cai, X. He, Z. Li, W.-Y. Ma, J.-R. Wen, "Hierarchical clustering of WWW image search results using visual, textual, and link information", ACM Multimedia, 2004.

[4] X.-J. Wang, W.-Y. Ma, G.-R. Xue, X. Li, "Multi-modal similarity propagation and its application for web image retrieval", ACM Multimedia, 2004.

[5] B. Gao, T.-Y. Liu, T. Qin, X. Zhang, Q.-S. Cheng, W.-Y. Ma, "Web image clustering by consistent utilization of visual features and surrounding texts", ACM Multimedia, 2005.

[6] X. He, W.-Y. Ma, O. King, M. Li and H.J. Zhang, "Learning and inferring a semantic space from user's relevance feedback", ACM Multimedia, 2002.

[7] S. Tong, E.Y. Chang, "Support vector machine active learning for image retrieval", ACM Multimedia, pp.107-118, 2001.

[8] K. Goh, E. Chang, W. Lai, "Concept-dependent multimodal active learning for image retrieval", ACM Multimedia, pp.564-571, 2004.

[9] Y. Rui, T.S. Huang, M. Ortega, S. Mehrotra, "Relevance feedback: A power tool in interactive content-based image retrieval", *IEEE Trans. on CSVT*, vol.8, no.5, pp.644-655, 1998.

[10] D. Tao, X. Tang, X. Li, X. Wu, "Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval", *IEEE Trans on PAMI*, vol. 28, no.7, pp. 1088-1099, 2006.

[11] D. Tao, X. Tang, X. Li, Y. Rui, "Direct kernel biased discriminant analysis: A new content-based image retrieval relevance feedback algorithm", *IEEE Trans. on Multimedia*, vol. 8, no.4, pp.716-727, 2006.

[12] X. Zhou, T. Huang, "Small sample learning during multimedia retrieval", Proc. IEEE CVPR, pp.11-17, 2001.

[13] Y. Chen, X. Zhou, T.S. Huang, "One-class SVM for learning in image retrieval", Proc. IEEE ICIP, 2001.

[14] J. Fan, Y. Gao, H. Luo, "Multi-level annotation of natural scenes using dominant image components and semantic image concepts", ACM Multimedia, 2004.

[15] M.R. Boutell, J. Luo, C.M. Brown, "Scene parsing using region-based generative models", *IEEE Trans. on Multimedia*, vol.9, no.1, pp.136-146, 2007.

[16] R. Zhang, Z. Zhang, "Hidden semantic concept discovery in region based image retrieval", Proc. IEEE CVPR, 2004.

[17] M. Swain, D. Ballard, "Color indexing", *Int. Journal of Computer Vision*, 1991.

[18] Y. Deng, B.S. Manjunath, C. Kenney, M.S. Moore, H. Shin, "An efficient color representation for image retrieval", *IEEE Trans. on Image Processing*, vol.10, pp.140-147, 2001.

[19] A.B. Torralba, A. Oliva, "Semantic organization of scenes using discriminant structural templates", Proc. IEEE ICCV, 1999.

[20] R. Fergus, P. Perona, A. Zisserman, "Object class recognition by unsupervised scale-invariant learning", Proc. IEEE CVPR, 2003.

[21] J. Fan , Y. Gao, H. Luo, G. Xu, "Statistical modeling and conceptualization of natural images", *Pattern Recognition*, vol.38, no.6, pp.865-885, 2005.

[22] Y. Gao, J. Fan, H. Luo, X. Xue, R. Jain, "Automatic image annotation by incorporating feature hierarchy and boosting to scale up SVM classifiers", ACM Multimedia, Santa Barbara, CA, 2006.

[23] W-Y. Ma, B. S. Manjunath, "Texture features and learning similarity", Proc. IEEE CVPR, pp.425-430, 1996.

[24] T. Chang, C.Kou, "Texture analysis and classification with tree-structured wavelet transform", *IEEE Trans. on Image Processing*, vol.2, 1993.

[25] D. Lowe, "Distinctive image features from scale invariant keypoints", *Intl Journal of Computer Vision*, vol.60, pp.91-110, 2004.

[26] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, L. J. Van Gool, "Modeling scenes with local descriptors and latent aspects", Proc. IEEE ICCV, pp.883-890, 2005.

[27] J. Fan, H. Luo, Y. Gao, R. Jain, "Mining multi-level image semantics via hierarchical classification", *IEEE Trans. on Multimedia*, vol. 10, no.1, pp.167-187, 2008.

[28] J. Fan, D.A. Keim, Y. Gao, H. Luo, Z. Li, "JustClick: Personalized image recommendation via exploratory search from large-scale Flickr images", *IEEE Trans. on CSVT*, vol. 19, no.2, pp.273-288, 2009.

[29] M. Girolami, "Mercer kernel-based clustering in feature space", *IEEE Trans. on Neural Networks*, vol.13, no.3, pp.780-784, 2002.

[30] A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, "Support vector clustering", *Journal of Machine Learning Research*, vol.2, pp.125-137, 2001.

[31] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, M. Jordan, "Learning the kernel matrix with semidefinite programming", *Journal of Machine Learning Research*, vol.5, pp.27-72, 2004.

[32] J. Fan, Y. Gao, H. Luo, "Integrating concept ontology and multi-task learning to achieve more effective classifier training for multi-level image annotation", *IEEE Trans. on Image Processing*, vol. 17, no.3, pp.407-426, 2008.

[33] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet, "More efficiency in multiple kernel learning", Proc. ICML, 2007.

[34] Y. Chen, J. Wang, R. Krovetz, "CLUE: cluster-based retrieval of images by unsupervised learning", *IEEE Trans. on Image Processing*, vol.14, no.8, pp.1187-1201, 2005.

[35] X. He, P. Niyogi, "Locality preserving projection", Proc. NIPS, 2003.

[36] D. Stan, I. Sethi, "eID: A system for exploration of image databases", *Information Processing and Management*, vol.39, pp.335-361, 2003.

[37] Y. Rubner, L. Guibas, C. Tomasi, "The earth's mover distance, multidimensional scaling, and color-based image retrieval", Proc. of ARPA Image Understanding Workshop, 1997.

[38] J.A. Walter, D. Webling, K. Essig, H. Ritter, "Interactive hyperbolic image browsing-towards an integrated multimedia navigator", ACM SIGKDD, 2006.

[39] K. Rodden, W. Basalaj, D. Sinclair, K. Wood, "Does organization by similarity assist image browsing?", ACM SIGCHI, 2001.

[40] J. Fan, D.A. Keim, Y. Gao, H. Luo, Z. Li, "A novel approach to enable semantic and visual Summarization for exploratory image search", ACM Conf. on Multimedia Information Retrieval (MIR'08), 2008.

[41] J. Kustanowitz, B. Shneiderman, "Hierarchical layouts for photo libraries", *IEEE Multimedia*, 2006.

[42] D. Heesch, A. Yavlinsky, S. Ruger, "NN^k networks and automated annotation for browsing large image collections from the world wide web", demo at ACM Multimedia, 2006.

[43] B. Moghaddam, Q. Tian, N. Lesh, C. Shen, T.S. Huang, "Visualization and user-modeling for browsing personal photo libraries", *Intl. Journal of Computer Vision*, vol.56, pp.109-130, 2004.

[44] M. Bilenko, S. Basu, R.J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering", Proc. ICML, pp.81-88, Banff, Canada, 2004.

[45] D. Cohn, R. Caruana, A. McCallum, "Semi-supervised clustering with user feedback", Technical Report TR2003-1982, Cornell University, 2003.

[46] E. Xing, A.Y. Ng, M.I. Jordan, S. Russell, "Distance metric learning with application to clustering with side-information", in *Advances in Neural Information Processing Systems* 15, MIT Press, 2003.

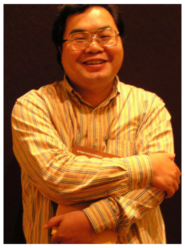
[47] S. Hoi, R. Jin, M.R. Lyu, "Learning non-parametric kernel matrices from pairwise constraints", Proc. ICML, pp.361-368, 2007.

[48] Y. Gao, J. Fan , H. Luo, S. Satoh, "A novel approach for filtering out junk images from Google search results", Intl Conf. on Multimedia Modeling, 2008.

Yuli Gao received his BS degree in computer science from Fudan University, Shanghai, China, in 2002. At the same year, he joined University of North Carolina at Charlotte to pursue his PhD degree on Information Technology. He got his PhD degree at 2007 and then joined HP Labs. His research interests include computer vision, image classification and retrieval, and statistical machine learning. He got an award from IBM as emerging leader in multimedia at 2006.

Hangzai Luo received the BS degree in computer science from Fudan University, Shanghai, China, in 1998. At the same year, he joined Fudan University as Lecturer. At 2002, he joined University of North Carolina at Charlotte to pursue his PhD degree on Information Technology. He got his PhD degree at 2006 and joined East China Normal University as Associate Professor at 2007. His research interests include computer vision, video retrieval, and statistical machine learning. He got second place award from Department of Homeland Security at 2007 for his excellent work on video analysis and visualization for homeland security applications.

Daniel A. Keim received the PhD degree in computer science from the University of Munich in 1994. He is a full professor in the Computer and Information Science Department, University of Konstanz. He has been an assistant professor in the Computer Science Department, University of Munich, and an associate professor in Computer Science Department, Martin Luther University Halle. He also worked at AT&T Shannon Research Labs, Florham Park, New jersey. In the field of information visualization, he developed several novel techniques, which use visualization technology for the purpose of exploring large databases. Dr. Keim has published extensively on information visualization and data mining, he has given tutorials on related issues at several large conferences, including Visualization, SIGMOD, VLDB, and KDD, he was program cochair of the IEEE Information Visualization Symposia in 1999 and 2000, the ACM SIGKDD Conference in 2002, and the Visual Analytics Symposium in 2006. Currently, he is on the editorial board of the IEEE Transactions on Knowledge and Data Engineering, the Knowledge and Information System Journal, and the Information Visualization Journal. He is a member of IEEE Computer Society.



Jianping Fan received his MS degree in theory physics from Northwestern University, Xian, China in 1994 and his PhD degree in optical storage and computer science from Shanghai Institute of Optics and Fine Mechanics, Chinese Academy of Sciences, Shanghai, China, in 1997.

He was a Postdoc Researcher at Fudan University, Shanghai, China, during 1998. From 1998 to 1999, he was a Researcher with Japan Society of Promotion of Science (JSPS), Department of Information System Engineering, Osaka University, Osaka, Japan. From September 1999 to 2001, he was a Postdoc Researcher in the Department of Computer Science, Purdue University, West Lafayette, IN. At 2001, he joined the Department of Computer Science, University of North Carolina at Charlotte as an Assistant Pofessor and then become Associate Professor. His research interests include image/video analysis, semantic image/video classification, personalized image/video recommendation, surveillance videos, and statistical machine learning.