

Human-Computer Interaction for Large-Scale Image Retrieval

Yuli Gao¹, Chunlei Yang², Yi Shen², Jianping Fan²

¹ Hewlett-Packard Labs, Palo Alto, CA 94304, USA, yuli.gao@hp.com

²Dept. of Computer Science, UNC-Charlotte, NC 28223, USA, jfan@uncc.edu

Abstract

As digital cameras and online photo sharing services become more popular, digital images are growing exponentially on the Internet. Thus supporting more effective retrieval from large-scale image collections has become a challenging issue for CBIR community. In this paper, we have developed a human-centered computing framework for assisting users to access large-scale online image collections: (a) filtering out junk images from Google Images; and (b) exploring large-scale Flickr images.

1. Introduction

Two commercial systems are now becoming very popular to support keyword-based retrieval from large-scale image collections: (a) Google Images has incorporated the associated texts for image indexing, but it may return large amounts of junk images; (b) Flickr has exploited social taggings to enable keyword-based image retrieval, but may seriously suffer from the problem of vocabulary discrepancy (image holders and image seekers may use different keywords). To support large-scale image retrieval, we have developed a new human-centered computing framework for: (1) filtering out the junk images from Google Images according to user's personal query intentions; and (2) exploring large-scale manually-tagged Flickr images. Obviously, users are treated one part of this human-centered multimedia computing framework [2-3].

Visual analytics [1], which can seamlessly integrate data analysis and visualization to enable visual-based communication between users and systems, is very attractive for developing new human-centered computing framework to support more effective retrieval from large-scale image collections. In this paper, we have developed a novel visual analytics framework for accessing large-scale image collections. In section 2, we present a novel visual analytics framework to filter out junk images from Google Images. In section 3, a new visual analytics framework is developed for assisting users on exploring large-scale Flickr images. We conclude this paper at section 4.

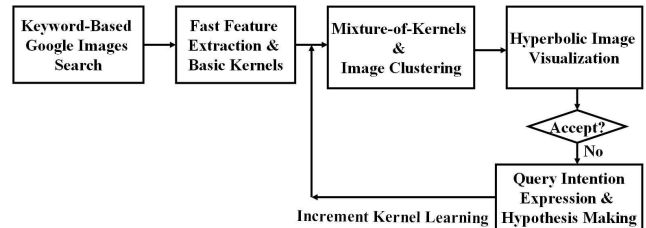


Figure 1: The flowchart for our interactive junk image filtering system.

2. Junk Image Filtering

Google Images search engine has been widely used for accessing large-scale online image collections, but it may seriously suffer from the low precision problem and return large amounts of junk images. To filter out the junk images from Google Images, our visual analytics framework consists of six key components as shown in Fig. 1: (a) Keyword-based image retrieval is first performed on Google Images, and a set of low-level visual features are extracted from the returned images to characterize their visual properties; (b) Multiple basic kernels are combined for characterizing the diverse similarity contexts between the returned images more precisely; (c) One-class SVM is performed to achieve an initial partition of the returned images (majority versus outliers), where the returned images in the outlier group are automatically treated as obvious junk images; (d) The returned images are visualized according to their visual similarity contexts, so that users can explore large amounts of returned images interactively and express their query intentions by clicking few images; (e) The users' inputs are integrated to achieve a new partition of returned images and filter out more junk images via incremental learning.

To characterize the visual properties of the returned images, both global visual features and local visual features are extracted for image content representation. 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 image partition patterns.

The diverse visual similarity contexts between the returned images are characterized more effectively and efficiently by using a linear combination of these three basic

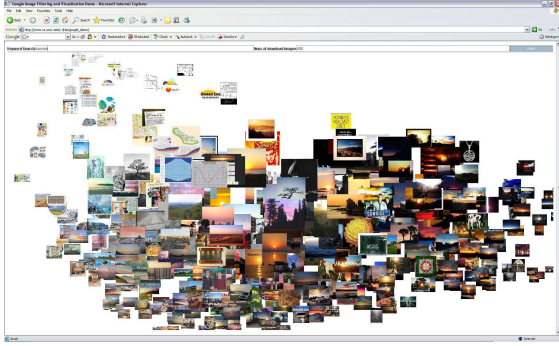


Figure 2: The obvious junk images for “sunrise” are separated from the majority of the returned images effectively and are projected on the top-left corner according to their visual properties.

image kernels (i.e., mixture-of-kernels) [8]:

$$\kappa(x, y) = \sum_{i=1}^3 \beta_i \kappa_i(x, y), \quad \sum_{i=1}^3 \beta_i = 1 \quad (1)$$

where $\beta_i \geq 0$ is the importance factor for the i th basic image kernel $\kappa_i(x, y)$ for image similarity characterization. The importance factors β depend on two issues: (1) significance of the relevant feature subset for visual similarity characterization; and (2) user’s preference or query intention.

One-class SVM is used to determine a smallest enclosing sphere of radius R to cover the majority of the returned images for the given keyword-based query [6]:

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

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

$$\min \left\{ R^2 + \frac{C}{N} \sum_{j=1}^N \xi_j \right\} \quad (3)$$

subject to:

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

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

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 returned images and result in better separation between the majority 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 users’ query intentions and personal preferences) can be obtained by maximizing the margin between the outliers and the majority of the returned images (i.e., the margin between the outliers and the closer

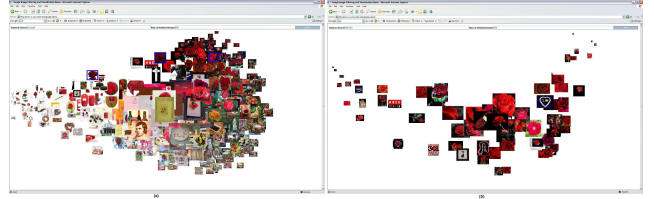


Figure 3: Junk image filtering: (a) the images returned by the keyword-based search “red rose” and the images in blue boundaries are selected as the relevant images by users; (b) the filtered images after the first run of relevance feedback.

support vectors on the boundary of the cluster sphere for the majority of the returned images):

$$\max_{\beta} \left\{ \sum_{l=1}^T \min [\kappa(z_l, x_i) | x_i \in \Omega, R^2(z_l) > R^2, R^2(x_i) = R^2] \right\} \quad (4)$$

where T is the total number of outlying images, Ω is the set of the support vectors locating 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 a precise visualization of the returned images as shown in Fig. 2. Through such a similarity-preserving image visualization process [3-4], users can assess the relevance between the returned images and their real query intentions effectively. As shown in Fig. 3, users can simply click one or few images to indicate their real query intentions, and such the query intentions are integrated to filter out more junk images as shown in Fig. 3. Online system is released at: http://www.cs.uncc.edu/~jfan/google_demo/

3. Personalized Image Retrieval

Social tagging is now becoming very popular for people to organize, share and retrieve large-scale images. Thus it is very important to develop new framework for exploring large-scale Flickr image collections. In this paper, we have developed a novel visual analytics scheme to allow users to explore large-scale Flickr image collections at two levels: (a) a *topic network* is incorporated for users to explore large-scale collections of Flickr images at a semantic level; (b) statistical image sampling and similarity-preserving image visualization [4-5] are integrated to enable image exploration at a perceptual level. After the images and their social taggings are downloaded from Flickr.com, the text terms which are relevant to the image topics (tags for image topic interpretation) are separated automatically by using standard text analysis techniques, and the basic vocabulary of image topics (i.e., keywords for image topic interpretation) are determined automatically.

The topic network consists of two components: (1) image topics; and (2) their inter-topic similarity contexts. The

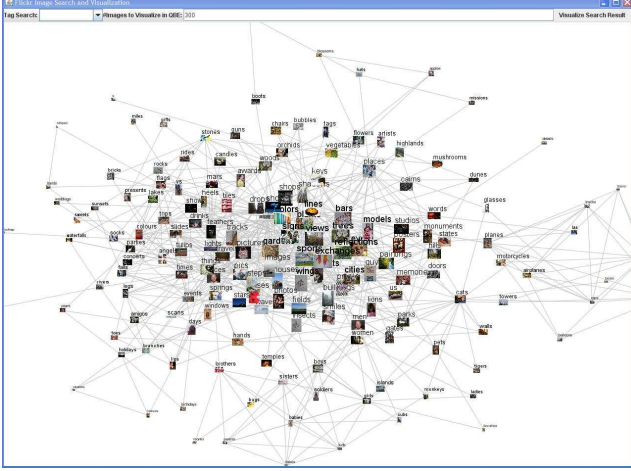


Figure 4: One portion of our topic network for indexing and summarizing large-scale collections of Flickr images at the topic level.

inter-topic similarity contexts consist of: (a) inter-topic semantic contexts; and (b) inter-topic visual contexts. The inter-topic semantic context $\phi(C_i, C_j)$ between two image topics C_i and C_j is defined as:

$$\rho(C_i, C_j) = -\frac{P(C_i, C_j)}{\log P(C_i, C_j)} \quad (5)$$

where $P(C_i, C_j)$ is the co-occurrence probability of the image topics C_j and C_i in the Flickr image collections.

The inter-topic visual context $\gamma(C_i, C_j)$ between the image topics C_i and C_j can be determined by performing canonical correlation analysis [11] on their image sets S_i and S_j :

$$\gamma(C_i, C_j) = \max_{\theta, \vartheta} \frac{\theta^T \kappa(S_i) \kappa(S_j) \vartheta}{\sqrt{\theta^T \kappa^2(S_i) \theta \cdot \vartheta^T \kappa^2(S_j) \vartheta}} \quad (6)$$

where θ and ϑ are the parameters for determining the optimal projection directions to maximize the correlations between two image sets S_i and S_j for the image topics C_i and C_j , $\kappa(S_i)$ and $\kappa(S_j)$ are the kernel functions for characterizing the visual correlations between the images in the same image sets S_i and S_j .

$$\kappa(S_i) = \sum_{x_l, x_m \in S_i} \kappa(x_l, x_m), \quad \kappa(S_j) = \sum_{x_h, x_k \in S_j} \kappa(x_h, x_k) \quad (7)$$

where the visual correlation between the images is defined as their kernel-based visual similarity $\kappa(\cdot, \cdot)$ in Eq. (1).

Both the inter-topic visual context $\gamma(C_i, C_j)$ and the inter-topic semantic context $\phi(C_i, C_j)$ are first normalized into the same interval, and they are further integrated to achieve more precise characterization of their cross-modal inter-topic similarity context $\varphi(C_i, C_j)$:

$$\varphi(C_i, C_j) = \epsilon \cdot \phi(C_i, C_j) + \eta \cdot \gamma(C_i, C_j), \quad \epsilon + \eta = 1 \quad (8)$$

where the first part denotes the semantic context between the image topics C_j and C_i , the second part indicates their inter-topic visual context, $\gamma(C_i, C_j)$ is the visual context between the image sets for the image topics C_i and C_j , ϵ and η are the importance factors for the inter-topic semantic context and the inter-topic visual context.

One part of our topic network for Flickr images is shown in Fig. 4, where each image topic is linked with multiple relevant image topics with larger values of $\varphi(\cdot, \cdot)$ and the geometric closeness between the image topics is related to the strengths of their inter-topic similarity contexts. Thus such a graphical representation of the topic network can reveal a great deal about how these image topics are correlated and how the relevant tags for interpreting multiple inter-related image topics are intended to be used jointly for image tagging. Through *change of focus* of topic network [4], users can interactively explore large-scale collections of Flickr images at the semantic level.

Each image topic on the topic network may contain large amounts of images, thus keyword-based image retrieval may seriously suffer from the problem of information overload. In order to tackle this problem, we have developed a novel framework for personalized image recommendation and it consists of three major components: (a) *Topic-Driven Image Summarization and Recommendation*: The semantically-similar images under the same topic are first partitioned into multiple clusters according to their non-linear visual similarity contexts, and a limited number of images are automatically selected as the most representative images according to their representativeness for a given image topic. Our system can also allow users to define the number of such most representative images for relevance assessment. (b) *Context-Driven Image Visualization and Exploration*: Kernel PCA and hyperbolic visualization are seamlessly integrated to enable interactive image exploration according to their inherent visual similarity contexts, so that users can assess the relevance between the recommended images (i.e., most representative images) and their real query intentions more effectively. (c) *Intention-Driven Image Recommendation*: An interactive user-system interface is designed to allow the user to express his/her time-varying query intentions easily for directing the system to find more relevant images according to his/her personal preferences.

Our visual summarization (i.e., the most representative images) results for the image topic “rose” is shown in Fig. 5, where 200 most representative images for the image topic “rose” are selected for representing the original visual similarity contexts between the images. One can observe that these 200 most representative images can provide an effective interpretation and summarization of the original visual similarity contexts among large amounts of semantically-similar images under the same topic. The underlying the



Figure 5: Our representativeness-based sampling technique can automatically select 200 most representative images to achieve precise visual summarization of 53829 semantically-similar images under the topic “rose”.

visual similarity contexts have also provided good directions for users to explore these most representative images interactively.

It is important to understand that the system alone cannot meet the users’ sophisticated image needs. Thus user-system interaction plays an important role for users to express their image needs, assess the relevance between the returned images and their real query intentions, and direct the system to find more relevant images adaptively. Based on these understandings, our system can allow users to zoom into the images of interests interactively and select one of these most representative images to express their query intentions or personal preferences as shown in Fig. 6(a).

After such the user’s time-varying query interests are captured, the personalized interestingness scores for the images under the same topic are calculated automatically, and the *personalized interestingness score* $\rho^p(x)$ for a given image with the visual feature x is defined as:

$$\rho^p(x) = \rho(x) + \rho(x) \times e^{-\kappa(x, x_c)} \quad (9)$$

where $\rho(x)$ is the original representativeness score for the given image, $\kappa(x, x_c)$ is the kernel-based visual similarity correlation between the given image with the visual features x and the clicked image with the visual features x_c which belong to the same image cluster. Thus the returned images with larger values of the personalized interestingness scores, which have similar visual properties with the clicked image (i.e., belonging to the same cluster) and are initially eliminated for reducing the visual complexity for image summarization and visualization, can be recovered and be recommended to the users adaptively as shown in Fig. 6(b). One can observe that integrating the visual similarity contexts for personalized image recommendation can significantly enhance the users’ ability on finding some particular images of interest even the low-level visual features may not be able to carry the semantics of the image contents directly.



Figure 6: Our interactive image exploration system: (a) the most representative images for the image topic “planes”, where the image in blue box is selected; (b) more images which are relevant to the user’s query intentions of “plane in blue sky”.

4. Conclusions

To support more effective retrieval of large-scale online image collections, we have developed a novel human-centered computing framework for: (a) filtering out junk images from Google Images; and (b) exploring large-scale Flickr images at both the semantic level and the perceptual level. Our experimental results on large-scale Google and Flickr images have obtained very positive results.

References

- [1] J. Thomas, K.A. Cook, *Illuminating the Path: The Research and Development Agenda for Visual Analytix*, IEEE, ISBN-7695-2323-4, 2005.
- [2] N. Sebe, Q. Tian, “Personalized multimedia retrieval: the new trend?” *ACM MIR*, 299-306, 2007.
- [3] A. Jaimes, N. Sebe, D. Gatica-Perez, “Human-Centered Computing: A Multimedia Perspectiv”, *ACM Multimedia*, 2006.
- [4] J. Lamping, R. Rao, “The hyperbolic browser: A focus+content technique for visualizing large hierarchies”, *Journal of Visual Languages and Computing*, vol.7, pp.33-55, 1996.
- [5] B. Moghaddam, Q. Tian, N. Lesh, C. Shen, T.S. Huang, “Visualization and user-modeling for browsing personal photo libraries”, *Intl. J. of Computer Vision*, vol.56, pp.109-130, 2004.
- [6] A. Ben-Hur, D. Horn, H.T. Siegelmann, V. Vapnik, “Support vector clustering”, *Journal of Machine Learning Research*, vol.2, pp.125-137, 2001.
- [7] D.R. Hardoon, S. Szedmak, J. Shawe-Taylor, “Canonical correlation analysis: An overview with application to learning methods”, Technical Report, CSD-TR-03-02, University of London, 2003.
- [8] 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, 2008.