Computing Iconic Summaries of General Visual Concepts

Rahul Raguram   Svetlana Lazebnik
{rraguram, lazebnik}@cs.unc.edu
Department of Computer Science
University of North Carolina at Chapel Hill

Abstract

This paper considers the problem of selecting iconic images to summarize general visual categories. We define iconic images as high-quality representatives of a large group of images consistent both in appearance and semantics. To find such groups, we perform joint clustering in the space of global image descriptors and latent topic vectors of tags associated with the images. To select the representative iconic images for the joint clusters, we use a quality ranking learned from a large collection of labeled images. For the purposes of visualization, iconic images are grouped by semantic "theme" and multidimensional scaling is used to compute a 2D layout that reflects the relationships between the themes. Results on four large-scale datasets demonstrate the ability of our approach to discover plausible themes and recurring visual motifs for challenging abstract concepts such as "love" and "beauty."

1. Introduction

The increasing popularity of photo-sharing websites such as Flickr has sparked a number of recent attempts to organize, browse and query enormous photo collections. A typical way to interact with Flickr (as well as many similar sites) involves querying for images based on a keyword. Unfortunately, the Flickr interface usually makes it difficult to obtain a complete, accurate, and visually compelling summary of query results for a particular category. This is illustrated by Figure 1, which shows the top 24 "most relevant" Flickr images for the query "apple." While most of these images have something to do with "apple" in one of two senses (either the fruit or the computer brand), on the whole, the retrieved results do not represent the corresponding concepts in a particularly salient or "iconic" way.

In this work, we aim to automatically identify iconic images to enable effective summarization, visualization, and browsing of general visual categories. It is very difficult to come up with a formal definition of "iconic image" for arbitrary (and possibly abstract) concepts. Instead, we propose an informal operational definition guided by several intuitions. If we issue a Flickr query for a general concept, the output is bound to contain many images that are either completely irrelevant or do not depict the target category well. However, it is likely that there will be consistent subsets of the search results corresponding to popular or recurring visual motifs, which are often determined by shared cultural associations of Flickr users. For example, images of red hearts or red roses will frequently get tagged with "love." We seek to identify such consistent image groups as being iconic or representative of the query.

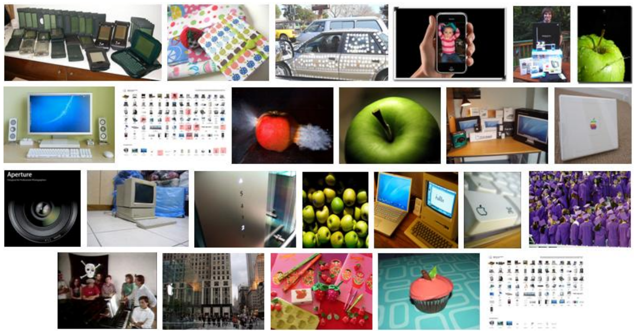


Figure 1. Top 24 "most relevant" images from Flickr for the query "apple" (retrieved on March 28, 2008).

More specifically, we define consistency both in terms of appearance and semantics. To capture similarity of appearance, we use global low-dimensional GIST features introduced by Oliva and Torralba [17]. To capture semantics, we perform probabilistic Latent Semantic Analysis (pLSA) [12] on Flickr tags to identify distinct "themes" within the query results. These themes may reflect polysemy, as in the "apple" example above, or correspond to different sub-categories of the query concept (e.g., "love" can refer to one's romantic partner, child, pet, etc.). By performing joint clustering based on GIST features and pLSA topic vectors, we can zero in on subsets of images that are perceptually and semantically similar. The next step is to choose a representative image from each cluster to serve as an iconic image for the category. For this, we propose aesthetics or image quality as the deciding factor, since higher-quality iconic images (e.g., professional images) should produce more pleasing and compelling visual summaries than lowerquality snapshots. To rank images by quality, we use a variant of the method of Ke et al. [14].

To recap, the main components of our approach are (a) appearance-based clustering, (b) tag-based clustering, and (c) quality ranking. Our contribution is in combining these three elements in a clean and simple way to produce effective visual summaries for *general* concepts. Section 2 will discuss other existing summarization approaches, most of which deal with specific 3D objects or scenes. Section 3 will explain the details of our approach, and Section 4 will demonstrate results on four large datasets, "apple," "beauty," "closeup," and "love." We close in Section 5 by discussing interesting findings made in the course of our research, and by outlining directions for future work.

2. Related Work

Several recent approaches have considered the problem of re-ranking Internet image search results for the purpose of automatically collecting large image databases [3, 9, 16, 22]. By contrast, our goal is to select a small number of salient and representative images to summarize a category at a glance. While dataset collection requires high recall, summarization has low recall by definition. Instead, the emphasis is on precision, as well as as on the somewhat subjective criterion of "representativeness."

For objects or scenes with a rigid 3D structure, the problem of finding a small number of representative views is known as *canonical view selection*. This problem has been addressed both in psychology [18, 5] and computer vision literature [4, 7, 10, 23]. Existing approaches have formulated many different criteria for canonical view selection, some of which are applicable to general and abstract categories (e.g., the canonical view should be similar to many other images in the collection), while others only make sense for objects with a well-defined geometry (e.g., different canonical views should be orthogonal to each other).

In the context of Internet image collections, canonical view selection has been considered for categories that share a common 2D or 3D structure, such as commercial products or logos [13] or famous tourist locations [4, 23]. Berg and Forsyth [4] have defined a canonical or iconic image as "an image that depicts a category member well, from a good aspect and in an uncluttered way." While this definition in itself is general, the approach in [4] is built almost entirely on a hard segmentation to identify *object* regions. As a consequence, the method works only in cases where there is a clear figure/ground separation. We seek to extend the definition of "iconic-ness" to include more abstract categories. Simon et al. [23] have proposed a method for summarizing landmark image collections by directly exploiting 3D scene structure. In this method, the similarity of two views is defined in terms of the number of 3D features they have in common. Unlike this work, we define iconic views without reference to 3D structure or camera motion (since these notions do not make sense for general categories), relying instead on a more holistic and perceptual notion of "shape," as discussed in the next section.

Finally, a few existing approaches explore the aesthetic or artistic aspect of summarization. For example, Rother et al. [20, 19] attempt to combine salient image fragments from a photo album into a visually pleasing collage. We invoke aesthetics for selecting iconic images to represent consistent clusters, but we do not attempt to composite these images. Instead, we use multidimensional scaling (MDS) [15, 21] to compute a 2D layout representing the semantic relationships between the iconic images.

3. The Approach

3.1. Joint clustering of appearance and tags

We think of iconic images as recurring visual motifs that illustrate a given concept, sometimes figuratively or symbolically. Such motifs may either take the form of objects or scenes (e.g., roses or sunsets may symbolize romance), or even abstract compositions with no clear figure/ground separation. We expect such motifs to be distinguished by a stable global spatial organization, and to identify them, we need an image representation that captures this organization in a good way. For this, we use the "GIST" representation of Oliva and Torralba [17], which is a low-dimensional descriptor that characterizes the "shape" of a scene, where a scene is interpreted as a unified entity, as opposed to a sum of its constituent objects. We have also considered augmenting the GIST descriptor with color as in [11], but for the datasets used in our experiments, color seems to be highly correlated with the global textural image characteristics, so that many of the GIST clusters are already quite consistent in terms of color. Thus, we do not use color for clustering, though we do use some color-based features for quality assessment, as explained in Section 3.2.

In the implementation, we compute 960-dimensional GIST descriptors from 128×128 thumbnails of each image in the dataset. Following this, Principal Component Analysis (PCA) is applied to further reduce the dimensionality of the feature vectors to 35. These feature vectors are clustered using k-means. Our experience indicates that a high k ($\approx$ 3000 clusters for the datasets used) produces a cleaner partitioning. Irrelevant and non-typical images tend to fall into small clusters, which are discarded based on a minimium size threshold (five in the implementation).

Next, we need to partition the remaining GIST clusters into semantically consistent subsets using tag information. To this end, we perform probabilistic Latent Semantic Analysis (pLSA) [12] to express the sets of tags associated with all images as weighted combinations of T "topics," each with its own characteristic distribution. Both the topic distributions and the image-specific topic weights are simultaneously computed by the EM algorithm. In our implementation, only tags that occur in more than ten images are retained; the size of the resulting tag vocabulary is approximately 3000 words for each dataset, and the number of topics T is set to 20. To group images based on the output of pLSA, we may either assign each image to the topic with the highest weight (posterior probability for that image), or do an additional step of clustering the entire topic vectors associated with all the images. In our experiments, we have obtained the best results by running k-means clustering on the topic vectors with $k = 30$.

At this point, we have two independently obtained clusterings, one based on GIST descriptors, and one based on pLSA topic vectors. To obtain a joint GIST/pLSA clustering, we simply take their intersection. That is, given a GIST cluster with label i and a pLSA cluster with label j, we form a new joint cluster (i, j) by taking the images that belong to *both* clusters. As before, clusters of fewer than five images are discarded. Empirically, the clusters that survive exhibit a significant degree of visual and semantic consistency.

Note that recent literature contains much more sophisticated approaches for building joint statistical models of words and images [2, 8]. However, these approaches are concerned with the problem of establishing correspondence between individual tags and specific image sub-regions, which is much more challenging than our goal of obtaining a few clusters consistent both in subject matter and appearance. Our simple joint clustering scheme is quite adequate for this task, as demonstrated by our results in Section 4.

3.2. Iconic image selection and visualization

After forming a joint clustering, we need to select iconic images that will represent the clusters. This is done by learning a quality ranking using a subset of the features described by Ke et al. [14]. Specifically, we have implemented edge spatial distribution, color distribution, blur estimation and hue count features, along with low-level features corresponding to contrast and brightness. To train the method, we used the same dataset as [14], consisting of 2500 high-quality and 2500 low-quality images. The method was verified using the same test image dataset as [14] and was found to produce comparable results. For each individual feature, a quality score is obtained as the likelihood ratio of that feature for high-quality vs. low-quality images, and the different feature-based scores are combined using the Naive Bayes model. The image that obtains the highest quality score in each joint GIST/pLSA cluster is selected as the iconic image for that cluster. Since quality ranking is inherently subjective and existing methods [6, 14] are not perfectly reliable, it is difficult to say whether the top-ranked image selected by our method is in fact the "best" one to represent its cluster. However, this ranking step does succeed in making sure that obviously flawed or low-quality images do not appear in our top-level summaries. In the future, we plan to systematically investigate the impact of quality assessment with the help of user studies.

For the purposes of visualization and browsing, we treat our category summaries as a three-level hierarchy. At the top level, we have pLSA clusters, which reflect different semantic aspects of the dataset. At the second level, each pLSA cluster expands into a collection of several iconic images, which represent different visual compositions of the same theme. Finally, each iconic image can be expanded to reveal the rest of the images in its GIST/pLSA cluster, which typically have a very similar appearance to the iconic.

In the figures presented in Section 4, each pLSA cluster at the top level is visualized using four of its top-ranked iconic images. The 2D layout of the pLSA clusters is computed by multidimensional scaling (MDS) [15] using the pLSA topic vectors. Given a matrix of pairwise distances between points in a high-dimensional space, MDS computes a set of points in a lower-dimensional space (two-dimensional, in our case), whose distances approximate the original ones. We perform MDS using the standardized Euclidean distance as a measure of proximity between pLSA topic vectors. The result is a two-dimensional map where related clusters are positioned close to each other, leading to a more intuitive visualization.

It may happen that an individual pLSA cluster will share very few tags with the others, in which case MDS will position it arbitrarily far away. This distorts the whole layout, and moreover, the cluster in question is usually an outlier for the category. In our implementation, we discard clusters that whose top ten tags (except for the original search tag) do not occur in any of the other clusters. For the datasets considered in this paper, we only had one such cluster, corresponding to Detroit architecture in the "love" category.

4. Results

In this section, we show results for four categories: "apple," "love," "beauty," and "closeup." The number of images automatically downloaded from Flickr for each query ranges from 15,000 to 20,000. Figures 2-5 show visual summaries produced by our method for these categories. The top part of each figure shows the pLSA clusters laid out with MDS. The quadruple of iconic images illustrating each pLSA cluster is annotated by the four most probable tags for that cluster (the query keyword, which is the top tag for all clusters, is omitted). The bottom part of each figure shows expansions for a few selected pLSA clusters in terms of all their iconic images, arranged in decreasing order of quality. In turn, a small number of those iconic images are expanded to show the contents of their clusters.

Figure 2 shows results for the keyword "apple." This is the only non-abstract keyword used in our experiments, and

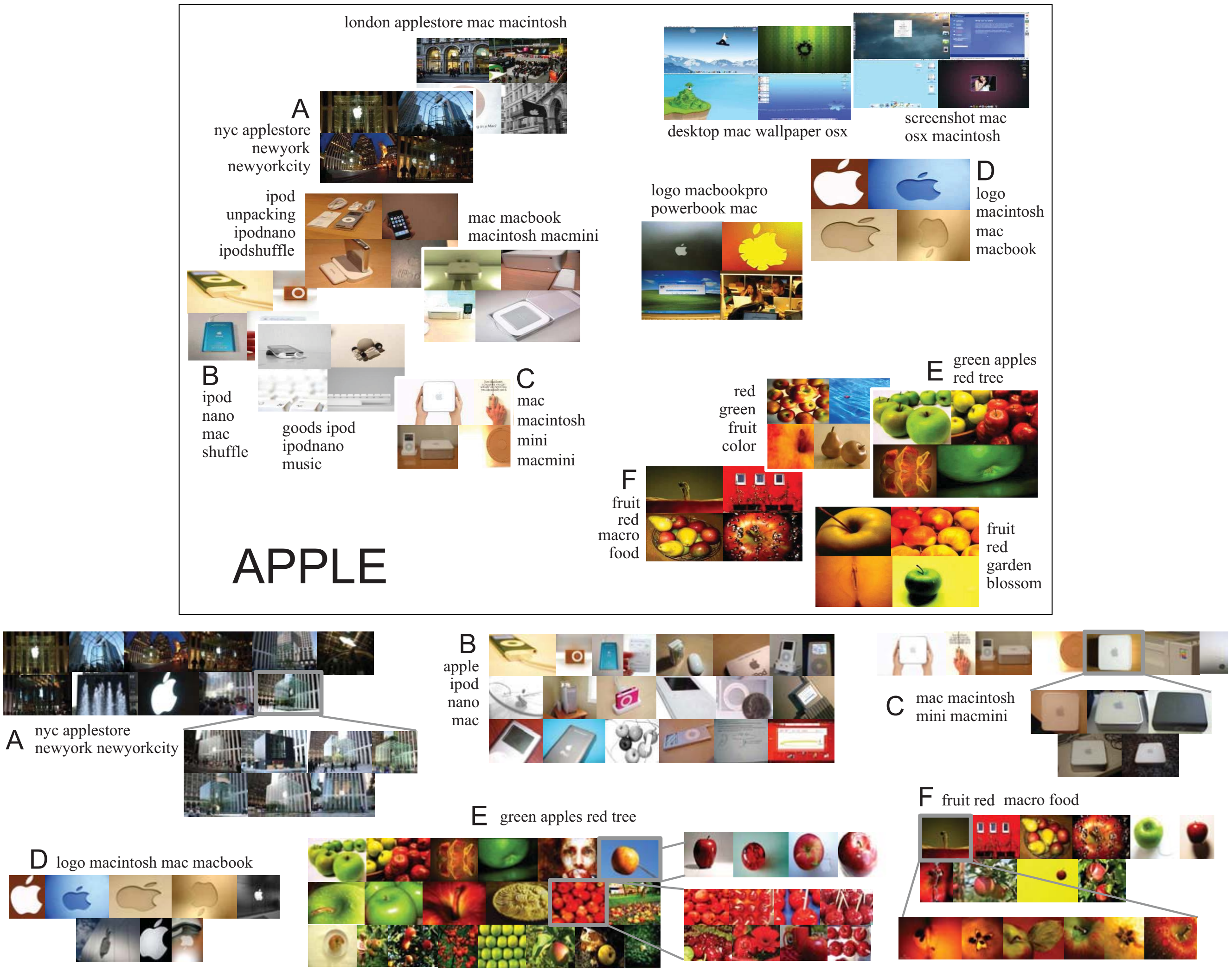

london applestore mac macintosh
A
nyc applestore newyork newyorkcity
desktop mac wallpaper osx
screenshot mac osx macintosh
ipod unpacking ipodnano ipodshuffle
mac macbook macintosh macmini
logo macbookpro powerbook mac
D
logo macintosh mac macbook
B
ipod nano mac shuffle
goods ipod ipodnano music
C
mac macintosh mini macmini
red green fruit color
E
green apples red tree
F
fruit red macro food
fruit red garden blossom
APPLE
A
nyc applestore newyork newyorkcity
B
apple ipod nano mac
C
mac macintosh mini macmini
E
green apples red tree
F
fruit red macro food
D
logo macintosh mac macbook

Figure 2. Apple results (see text).

we chose it because of its obvious polysemy to validate our basic approach. The computed pLSA clusters successfully capture both senses of the keyword, and the MDS layout shows a clear separation between the fruit and the Apple Macintosh clusters. The latter clusters capture the Apple logo, various Apple products, and desktop screenshots from Mac machines. There are even two distinct clusters for the Apple stores in London and New York.

Figure 3 shows results for "beauty." The top-level pLSA clusters for this category correspond to potraits of women (with different clusters zeroing in on glamor shots, nudes, and Japanese girls), as well as pets (predominantly cats), flowers, and nature shots. Not surprisingly, sunsets are very prominent among the landscape shots, as shown in the expanded clusters in parts B and C in the bottom of the figure. Part D also shows an expansion of a few "flower" iconic images, confirming the high degree of visual coherence in the corresponding joint GIST/pLSA clusters.

To test the ability of our method to cope with abstract categories, we tried "closeup" as a search term. In principle, this keyword refers to photographic technique as opposed to subject matter, so it is not a priori clear what specific visual categories or compositions should emerge as iconic in this case. The summary shown in Figure 4 reveals close-ups of faces and their parts (in particular, eyes and lips), close-ups of cats' noses, as well as a number of clusters dedicated to nature imagery, including plants, birds, and insects. There is even a cluster consisting primarily of high-speed pictures of water drops, which is expanded in part A of the figure. To provide a qualitative comparison, we also include a screenshot of the Flickr clusters for the "closeup" tag. We can note that the third Flickr cluster mixes human faces and cat faces, whereas in our summary, humans and cats fall into different pLSA clusters.

Figure 5 shows results for our fourth category, "love." As one would expect, there are large and salient clusters consisting of hearts and roses. There are also clusters corresponding to couples on beaches, babies and dogs. Somewhat surprisingly, there is a clearly defined cluster dedicated to self-love (three of the top four tags include "me," "self," and "selfportrait"). In the bottom right of the figure, we include a screenshot of the Flickr "love" clusters for com-

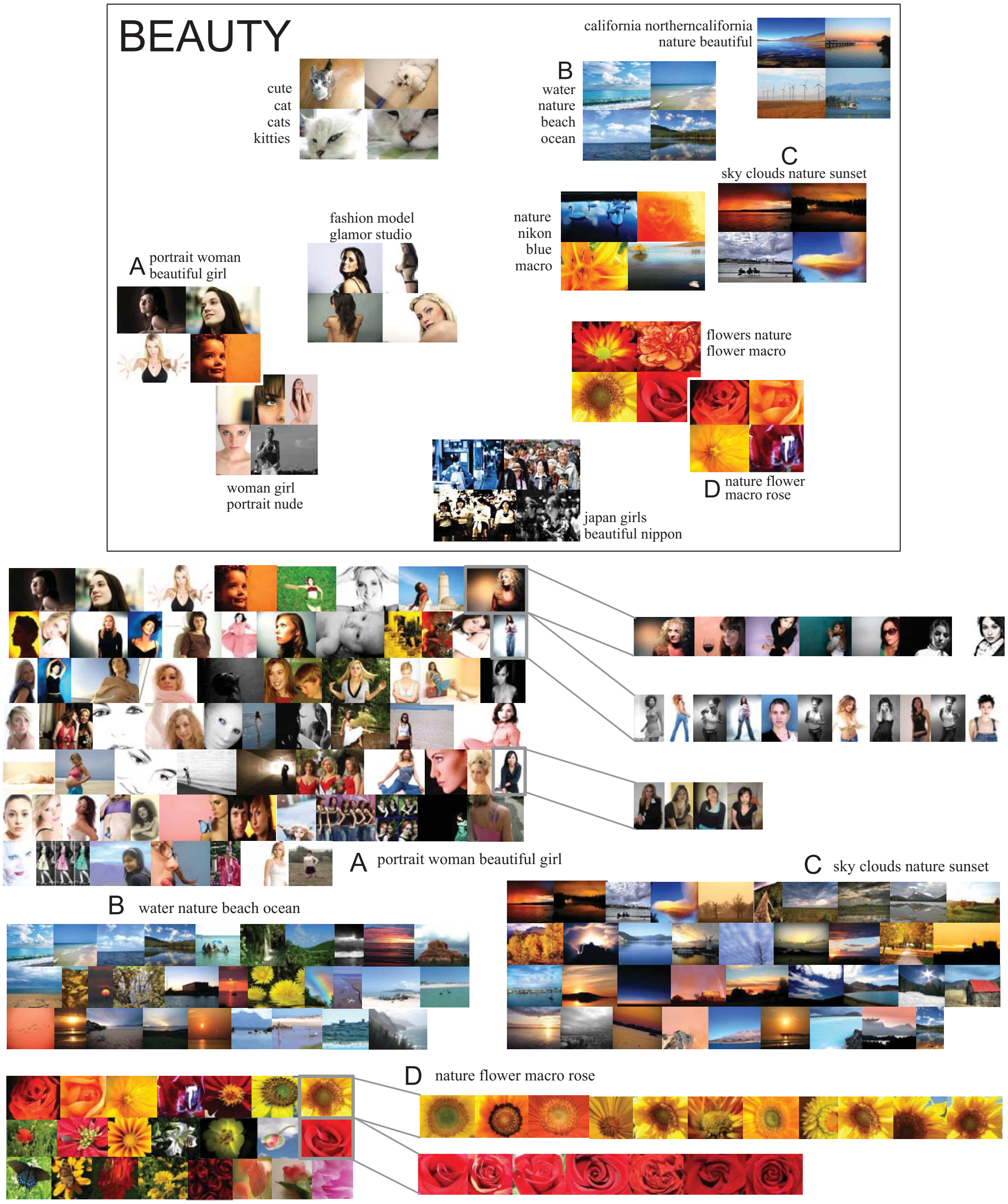

BEAUTY
cute cat cats kitties
california northerncalifornia nature beautiful
B water nature beach ocean
C sky clouds nature sunset
fashion model glamor studio
nature nikon blue macro
A portrait woman beautiful girl
flowers nature flower macro
woman girl portrait nude
D nature flower macro rose
japan girls beautiful nippon
A portrait woman beautiful girl
C sky clouds nature sunset
B water nature beach ocean
D nature flower macro rose

Figure 3. Beauty results (see text).

parison. Flickr obtains a somewhat cleaner cluster for weddings, but their top cluster mixes women, children, and pets, whereas these different love objects are clearly separated in our summary. Our summary also enables us to spot some non-obvious recurring motifs or “visual clichés.” One of these is a picture of a heart or the word “love” scrawled in the sand, as can be seen in part B of the figure. Another is a picture of a ring placed on top of an open book, so that the shadow of the ring takes the shape of a heart – five of the iconic images in the expanded pLSA cluster in part C of the figure are instances of this cliché.

5. Discussion and Future Work

The techniques presented in this paper show promise for helping users to interact with photo-sharing sites like Flickr more effectively. Despite the conventional wisdom that Flickr tags are quite noisy and unreliable, we have found

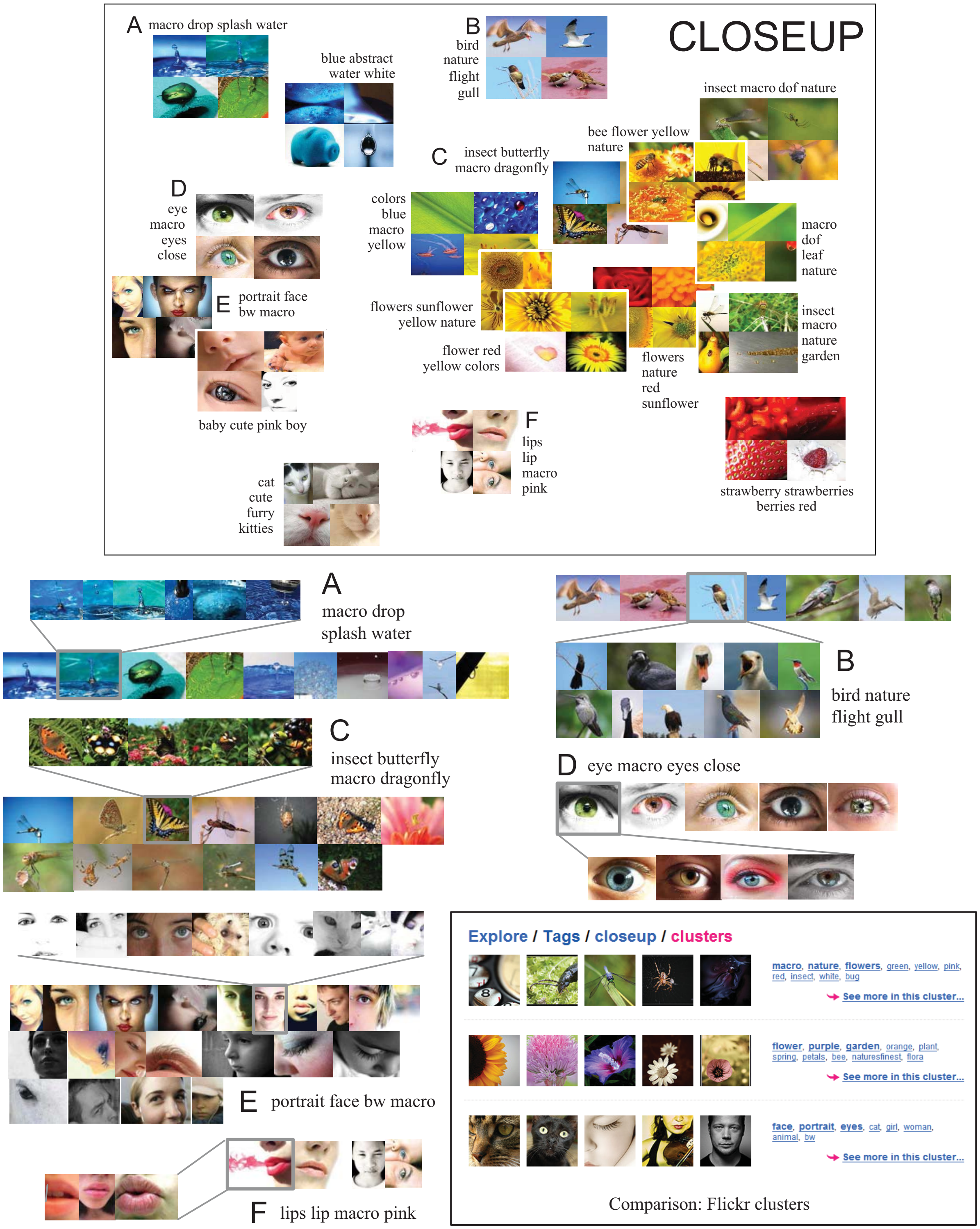
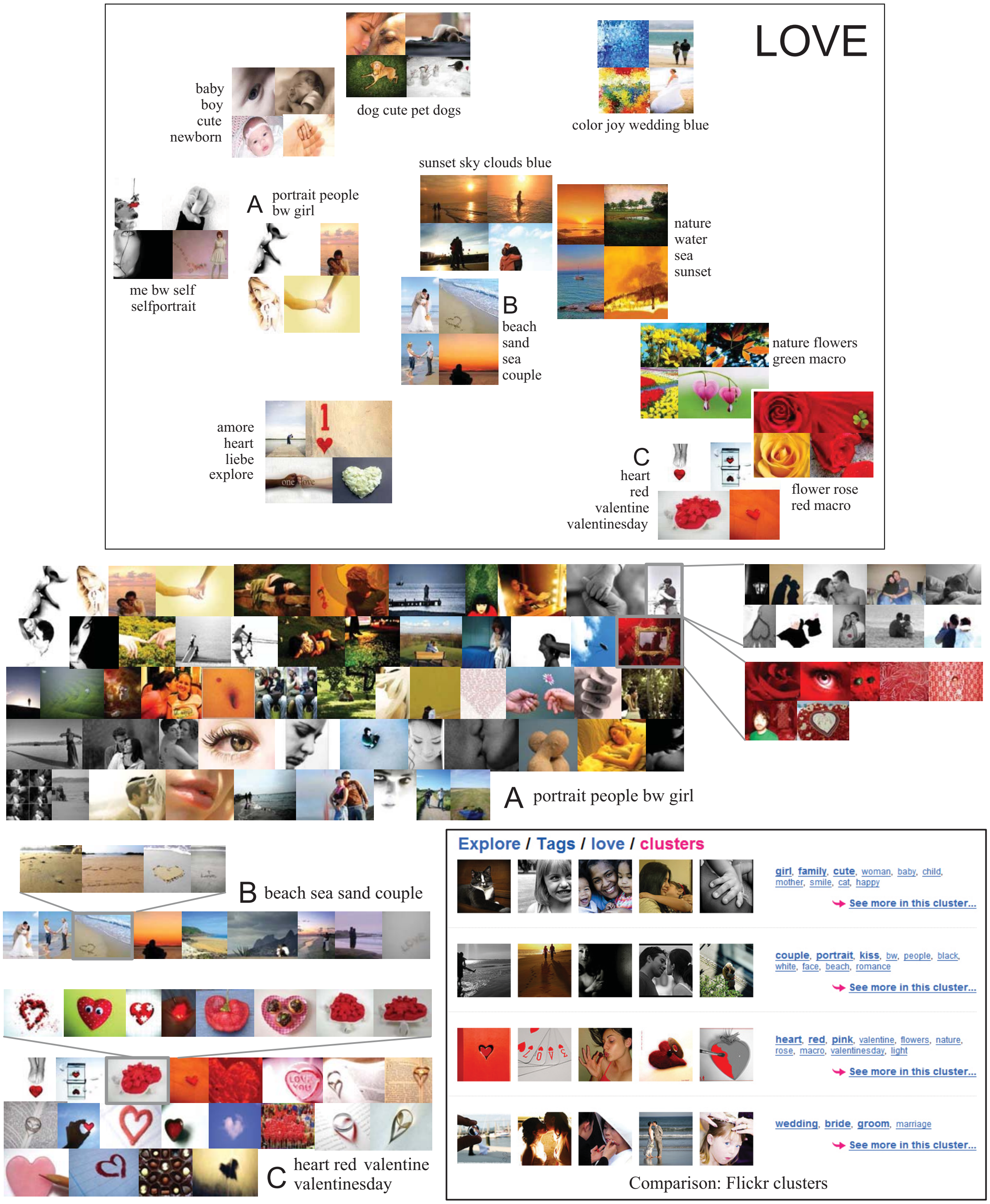

CLOSEUP
A macro drop splash water
blue abstract water white
B bird nature flight gull
insect macro dof nature
bee flower yellow nature
C insect butterfly macro dragonfly
D eye macro eyes close
colors blue macro yellow
macro dof leaf nature
E portrait face bw macro
flowers sunflower yellow nature
insect macro nature garden
flower red yellow colors
flowers nature red sunflower
baby cute pink boy
F lips lip macro pink
cat cute furry kitties
strawberry strawberries berries red
A macro drop splash water
B bird nature flight gull
C insect butterfly macro dragonfly
D eye macro eyes close
E portrait face bw macro
F lips lip macro pink
Explore / Tags / closeup / clusters
macro, nature, flowers, green, yellow, pink, red, insect, white, bug
See more in this cluster...
flower, purple, garden, orange, plant, spring, petals, bee, naturesfinest, flora
See more in this cluster...
face, portrait, eyes, cat, girl, woman, animal, bw
See more in this cluster...
Comparison: Flickr clusters

Figure 4. Closeup results (see text).

that the pLSA clusters found in our data make a great deal of sense. Moreover, joint GIST/pLSA clustering often tends to reveal interesting and unexpected aspects of the target category. While the preliminary results are thus quite encouraging, our approach is currently very simple and needs validation in the form of user studies to determine just how successful our summaries are, as well as to evaluate the impact of different implementation choices, such as quality ranking. We have also identified a few shorcomings that we plan to remedy in future work. In our present implementation, pLSA does not always succeed in identifying *all* the relevant aspects of our target categories. While some


LOVE
baby boy cute newborn
dog cute pet dogs
color joy wedding blue
sunset sky clouds blue
A portrait people bw girl
me bw self selfportrait
nature water sea sunset
B beach sand sea couple
nature flowers green macro
amore heart liebe explore
C heart red valentine valentinesday
flower rose red macro
A portrait people bw girl
B beach sea sand couple
C heart red valentine valentinesday
Explore / Tags / love / clusters
Comparison: Flickr clusters

girl, family, cute, woman, baby, child, mother, smile, cat, happy

See more in this cluster...

couple, portrait, kiss, bw, people, black, white, face, beach, romance

See more in this cluster...

heart, red, pink, valentine, flowers, nature, rose, macro, valentinesday, light

See more in this cluster...

wedding, bride, groom, marriage

See more in this cluster...

Figure 5. Love results (see text).

themes are identified with remarkable precision, others appear to be "smeared" across multiple pLSA clusters. For the case of "apple," we obtain two tight clusters curresponding to the New York and London Apple stores, but images of apple trees or apple blossoms are scattered across multiple clusters. In the case of "love," we obtain an excellent cluster for dogs, but not for brides and grooms.

A related issue is that we do not always have a good understanding of the significance of the clusters produced by our method. For example, is it significant that cats show up under "beauty" while dogs show up under "love"? As another example, for the "closeup" category, we get a nice cluster of strawberry images. It may be that strawberries constitute a major visual cliché for this category, or it may be that there are iconic motifs corresponding to other fruits and berries that simply do not show up in our summary.

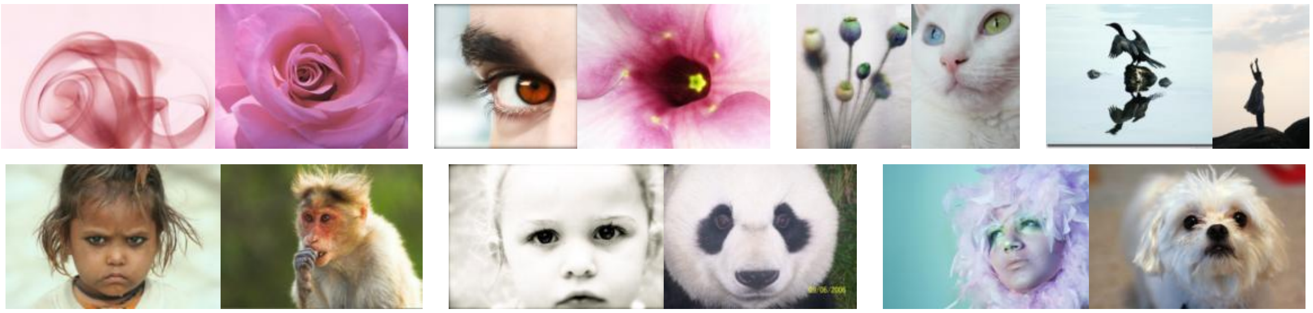


Figure 6. Examples of visual rhymes. Each of the above pairs was placed within the same GIST/pLSA cluster by our approach.

There are really two issues here: one is imperfect clustering leading to selective recall from the downloaded dataset, as discussed above, and another is bias in the dataset itself. Specifically, images from a single non-typical user or group (lovers of strawberry macro photography?) may be overrepresented among the downloaded query results. In the future, we plan to modify our downloading scheme to ensure a more balanced sampling of the general photo pool.

We should remark that failures of pLSA-based clustering are sometimes serendipitous, leading to the fascinating phenomenon of "visual rhymes" [1] or "convergences" [24]. These are pairs of photos that are ostensibly different in subject matter, but share an unexpected visual similarity that becomes a rich source of semantic associations. Figure 6 shows several examples of visual rhymes inadvertently produced by our approach. We can observe that particularly effective rhymes are produced by oppositions such as human/animal, animate/inanimate, organic/inorganic. Rhymes between human and animal faces are especially common among our results.

Several times in the preceding discussion, we have mentioned the notion of a visual cliché. We conjecture that many visual clichés are essentially implicit visual rhymes. For example, two swans with their necks bent toward each other recall the shape of a heart, as does the shadow of a ring placed over an open book, and many human observers can make the connection without having to see these images explicitly paired with an image of a heart. Even though our method starts out knowing nothing about such correspondences, by clustering a large dataset containing many pictures of hearts, swans, and rings tagged by users who share a common cultural background, it can end up "discovering" the underlying associations. In the future, we would like to operationalize the notions of visual rhyme and cliché more precisely and to develop automatic methods for explicitly identifying them. More generally, we are interested in the study of the "visual language" of photographs on the Internet, which encompasses not only computation, but also psychology, sociology, art, and semiotics.

References

[1] http://community.livejournal.com/rhyming_pics.

[2] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[3] T. Berg and D. Forsyth. Animals on the web. In *CVPR*, 2006.

[4] T. L. Berg and D. A. Forsyth. Automatic ranking of iconic images. Technical report, U.C. Berkeley, January 2007.

[5] V. Blanz, M. Tarr, and H. Bulthoff. What object attributes determine canonical views? *Perception*, 28(5):575–600, 1999.

[6] R. Datta, D. Joshi, J. Li, and J. Wang. Studying aesthetics in photographic images using a computational approach. In *ECCV*, 2006.

[7] T. Denton, M. Demirci, J. Abrahamson, A. Shokoufandeh, and S. Dickinson. Selecting canonical views for view-based 3-d object recognition. In *ICPR*, pages 273–276, 2004.

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, volume 4, pages 97–112, 2002.

[9] R. Fergus, F.-F. Li, P. Perona, and A. Zisserman. Learning object categories from Google's image search. In *CVPR*, 2005.

[10] P. Hall and M. Owen. Simple canonical views. In *BMVC*, pages 839–848, 2005.

[11] J. Hays and A. A. Efros. Scene completion using millions of photographs. In *SIGGRAPH*, 2007.

[12] T. Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, 1999.

[13] Y. Jing, S. Baluja, and H. Rowley. Canonical image selection from the web. In *Proceedings of the 6th ACM international Conference on Image and Video Retrieval*, 2007.

[14] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2006.

[15] J. B. Kruskal. Muulti-dimensional scaling by optimizing goodness-of-fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

[16] L.-J. Li, G. Wang, and L. Fei-Fei. OPTIMOL: automatic object picture collection via incremental model learning. In *CVPR*, 2007.

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

[18] S. Palmer, E. Rosch, and P. Chase. Canonical perspective and the perception of objects. *Attention and Performance*, IX:135–151, 1981.

[19] C. Rother, L. Bordeaux, Y. Hamadi, and A. Blake. Autocollage. In *SIGGRAPH*, 2006.

[20] C. Rother, S. Kumar, V. Kolmogorov, and A. Blake. Digital tapestry. In *Proc. CVPR*, pages 589–596, 2005.

[21] Y. Rubner, C. Tomasi, and L. Guibas. A metric for distributions with applications to image databases. In *ICCV*, pages 59–66, 1998.

[22] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[23] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. In *International Conference on Computer Vision*, 2007.

[24] L. Weschler. *Everything that rises: A book of convergences*. McSweeney's, 2006.