Automatic Ranking of Iconic Images

Tamara L. Berg
University of California, Berkeley
Berkeley, CA 94720
millert@cs.berkeley.edu

D. A. Forsyth
University of Illinois, Urbana Champaign
Urbana, IL 61801
daf@cs.uiuc.edu

Abstract

*We define an iconic image for an object category (*e.g. *eiffel tower) as an image with a large clearly delineated instance of the object in a characteristic aspect. We show that for a variety of objects such iconic images exist and argue that these are the images most relevant to that category. Given a large set of images noisily labeled with a common theme, say a Flickr tag, we show how to rank these images according to how well they represent a visual category. We also generate a binary segmentation for each image indicating roughly where the subject is located. The segmentation procedure is learned from data on a small set of iconic images from a few training categories and then applied to several other test categories. We rank the segmented test images according to shape and appearance similarity against a set of 5 hand-labeled images per category. We compute three rankings of the data: a random ranking of the images within the category, a ranking using similarity over the whole image, and a ranking using similarity applied only within the subject of the photograph. We then evaluate the rankings qualitatively and with a user study.*

1. Introduction

There are now many popular websites where people share pictures. Typically, these pictures are labelled, with labels indicating well-known objects depicted. However, the labellings are not particularly accurate, perhaps because people will label all pictures in a memory card with a particular label. This means, for example, that the photograph of the Eiffel Tower and a photograph of a friend taken in a nearby cafe will both have the label `eiffel tower`. Our user study results show that about half of the pictures for the categories we used on Flickr represent the category poorly.

All this means that these collections are hard to use for training object recognition programs, or, for that matter, as a source of illustrations, etc. We would like to rank such sets of images according to how well they depict the category. We refer to an image that depicts a category member well, from a good aspect and in an uncluttered way, as an **iconic image**. We believe that such iconic representations should exist for many categories, especially landmarks as we study in this paper, because people tend to take many photographs of these objects and among this large number there will be many taken from similar characteristic views.

In this paper, we show that iconic images can be identified rather accurately in natural datasets by segmenting images with a procedure that identifies foreground pixels, then ranking based on the appearance and shape of those foreground regions. This foreground/background segmentation also yields a good estimate of where the subject of the image lies.

1.1. Previous Work

There has been some previous work on automatically determining the subject of photographs. Li *et al.* [12] automatically determine the object of interest in photographs. However, their focus is on images with low depth of field. Banerjee and Evans [2] propose an in-camera main subject segmentation algorithm that uses camera controls to automatically determine the subject. Since we collect our images from the web we cannot use this method. The work most related to ours in this area is Luo *et al.* [13] who use region segmentation and probabilistic reasoning to automatically determine subjects in unconstrained images, although they do this in a very different manner than our method.

Segmentation is well studied and cannot be reasonably surveyed in the space available. Most related to our work are segmentation algorithms involving Markov Random field models dating back to Geman and Geman [7] and studied by many others since. We use a Markov Random field segmentation described by Boykov and Kolmogorov [4]. This is an implementation of a min-cut/max-flow algorithm to compute a two label segmentation efficiently.

There has been extensive work on ranking images for content retrieval [1, 8, 11, 15] and on automatically reranking search results [5, 6, 14]. We focus on the area of ranking iconic images and include a notion of where the object lies within the picture as an aid to doing this task.

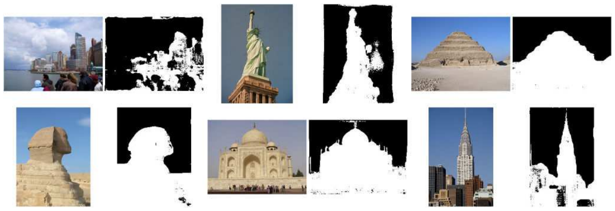

Figure 1. Some example segmentations of photographs into object and background labels. Our segmentation procedure is learned from a small set of 110 hand segmented iconic images from a few training categories (`eiffel tower`, `golden gate bridge`, `colosseum` and `stonehenge`). It is then applied to test images of previously unseen categories. While it is quite difficult to build a foreground/background segmentation algorithm that works on all images in general, our segmenter works well on iconic images with large, clearly delineated objects.

Another related paper from Ke *et al.* [9] concentrates on the problem of ranking images according to their photographic quality. This quality measure is somewhat related to our notion of iconic images which ideally should also be high quality images.

Many people believe that segmentation and recognition are linked in some natural way. There have been some papers showing that segmentation can help improve recognition results. Barnard *et al.* [3] show that different possible segmentations can be judged according to how well they predict words for regions and that word prediction can be improved by using these segmentations. Liebe and Schiele [10] use segmentation as a way of integrating individual image cues and show that this multi-cue combination scheme increases detection performance compared to any cue in isolation.

We integrate aspects from all of these areas, automatically detecting the subject of photographs using segmentation methods and re-ranking images according to how well they represent a visual category.

1.2. Data

Our dataset consists of photographs collected from Flickr for a set of 13 categories. We use all public photos uploaded over a period of one year containing that category in any associated text. Each category contains between 4,000 and 40,000 images. Four categories are used for training our segmentation algorithm: `colosseum`, `eiffel tower`, `golden gate bridge` and `stonehenge`. Nine categories are used for testing: `capital building`, `chrysler building`, `empire state building`, `lincoln memorial`, `sphinx`, `statue of liberty`, `sydney opera house`, `taj mahal` and `pyramid`.

2. Computing Segmentations

The goal of the segmentation portion of our method is to automatically detect the region of the image corresponding to the subject of the photograph. As such, we want to compute a binary segmentation of subject and background. Because this segmentation has only two labels we can use a very efficient min-cut/max-flow algorithm developed by Boykov and Kolmogorov [4]. Images are modeled as a Markov Random Field where for an image, each pixel corresponds to a node of the graph, with edges between each node and the source and sink nodes, as well as edges between the pixel and its four neighboring pixels in the image.

Segmentation parameters are learned on a set of training images from 4 training categories and then applied to new images from test categories. The features used to compute our segmentations will be described in section 2.1 and computing the unary and binary potentials for the edge weights will be described in section 2.2.

2.1. Image Features

We compute 7 features describing each pixel: focus, texture, hue, saturation, value, vertical position and horizontal position. These features were selected because we tend to believe that the subject of a photograph is more likely to be sharp, textured, more vivid in color and brighter than the background. We also believe that the subject will be more likely to lie in either the middle of the photo or be placed at one of the intersections suggested by the rule of thirds (a common rule of good subject placement in photographs).

Focus is computed in a 3x3 window around each pixel as the average ratio of high pass energy to low pass energy. Texture is also computed in a 3x3 window by computing the average texture response to a set of 6 bar and spot filters. Hue, saturation and value correspond to their respective values at each pixel. Location for each pixel is represented as its x location and y location divided by the image width and height respectively. Each of these features has a value ranging between 0 and 1.

2.2. Learning Potentials

We use training data to learn how our features contribute to the probability of subject versus background and to the probability of a discontinuity between neighboring pixel labels. We use 110 training images from 4 categories (`colosseum`, `eiffel tower`, `golden gate bridge` and `stonehenge`) that have been hand segmented into object and background. These training images were selected to be highly iconic images with large, clearly delineated subjects.

There are two types of potentials necessary for our segmentation algorithm. The unary potentials correspond to the probability of a pixel being subject (edge weights between pixels and the source node) and the probability of a pixel being background (edge weights between pixels and the sink node). The second potential type are the binary potentials between neighboring nodes. These correspond to the probability of the labels being the same between neighboring nodes.

All feature vectors in the training images are clustered together using k-means clustering with 1000 clusters. The probability of subject and background, $P(source|pixel)$ and $P(sink|pixel)$, are computed for each cluster as the percentage of training pixels within the cluster labeled as object and background respectively. The probability of two neighboring pixels having the same label, $P(same|pixel_i, pixel_j)$ where i and j are neighboring pixels, is computed as the percentage of such occurrences given the pixel's cluster index and the neighboring pixel's cluster index.

2.3. Segmentation Results

For a test image, features are computed for each pixel. These features are associated with the index of the closest cluster center. Each pixel then inherits the source and sink probabilities of its cluster index. Each pair of neighboring pixels is assigned the pre-computed probability of having the same label given their cluster indices. We compute the edges for the image's graph as the logs of these probabilities (where edges have symmetric weights) and run the min-cut/max-flow algorithm on them.

We don't expect the segmentation to work perfectly for images in general as determining figure/ground segmentations is quite a difficult task. However, by definition the images that are iconic should have a large object instance in the midst of a fairly uncluttered background. Thus, these images should be relatively easy to segment. As long as our segmenter works on these images it should help to determine which of the large pool of images are the representative ones.

In figure 1 we show some segmentation results on 6 example images. In each of these images the segmentation algorithm is able to automatically determine the subject of the photograph. Doing this allows us to compute similarity between images using the appearance of only those parts of the image that correspond to the object of interest which will be used in our ranking task, section 3.2. The segmentation also gives us an idea of the support of the object which is used to find objects with similar shapes.

3. Ranking Images

For each test category we select 5 iconic ground truth images as training. We compute rankings against the training images using three alternative methods and compare their results. As a baseline computation, the first ranking that we compute is a random ranking of the images. The second ranking uses similarity in appearance to the ground truth images for the appropriate category. The last ranking that we compute uses our figure/ground segmentations to compute similarity based on appearance and shape.

3.1. Ranking Without Segmentations

To rank the images we use the same 7 dimensional feature vectors as used for segmentation. These vectors have some idea of color, location, focus and texture. For each training and test image we compute the average over all pixels in the image of these feature vectors. The test images are then compared to all training images using the normalized correlation of their average feature vectors. The test images are ranked according to their maximum correlation value to any training image.

3.2. Ranking With Segmentations

For our ranking with segmentation information we compare test images to training images using similarity in shape and appearance. First the segmentation is run on all of the training and test images.

Shape similarity between test and training images is computed as the normalized correlation between their binary segmentation masks. This should give larger values to shapes that are more similar, though it is a somewhat rough measure of shape similarity.

Appearance vectors are calculated by taking the average feature vector within the region marked as object. Appearance similarity between two images is then computed as the normalized correlation between average feature vectors. Because the appearance is computed only over the region marked as object, this measure is more robust to changes in background than the similarity computed for the ranking without segmentation.

Test images are then ranked according to their maximum correlation to any training image where correlation to a training image is computed as the sum of their appearance and shape correlations.

4. Results

We have produced ranked results for 9 test categories. We judge our rankings qualitatively by showing some highly ranked photos for our three methods. More results of this nature can be viewed in the supplementary material associated with our paper. We also judge our results quantitatively according to the results of a user study which compares the goodness of our top ranked images to top ranked images ranked using the two alternative methods.

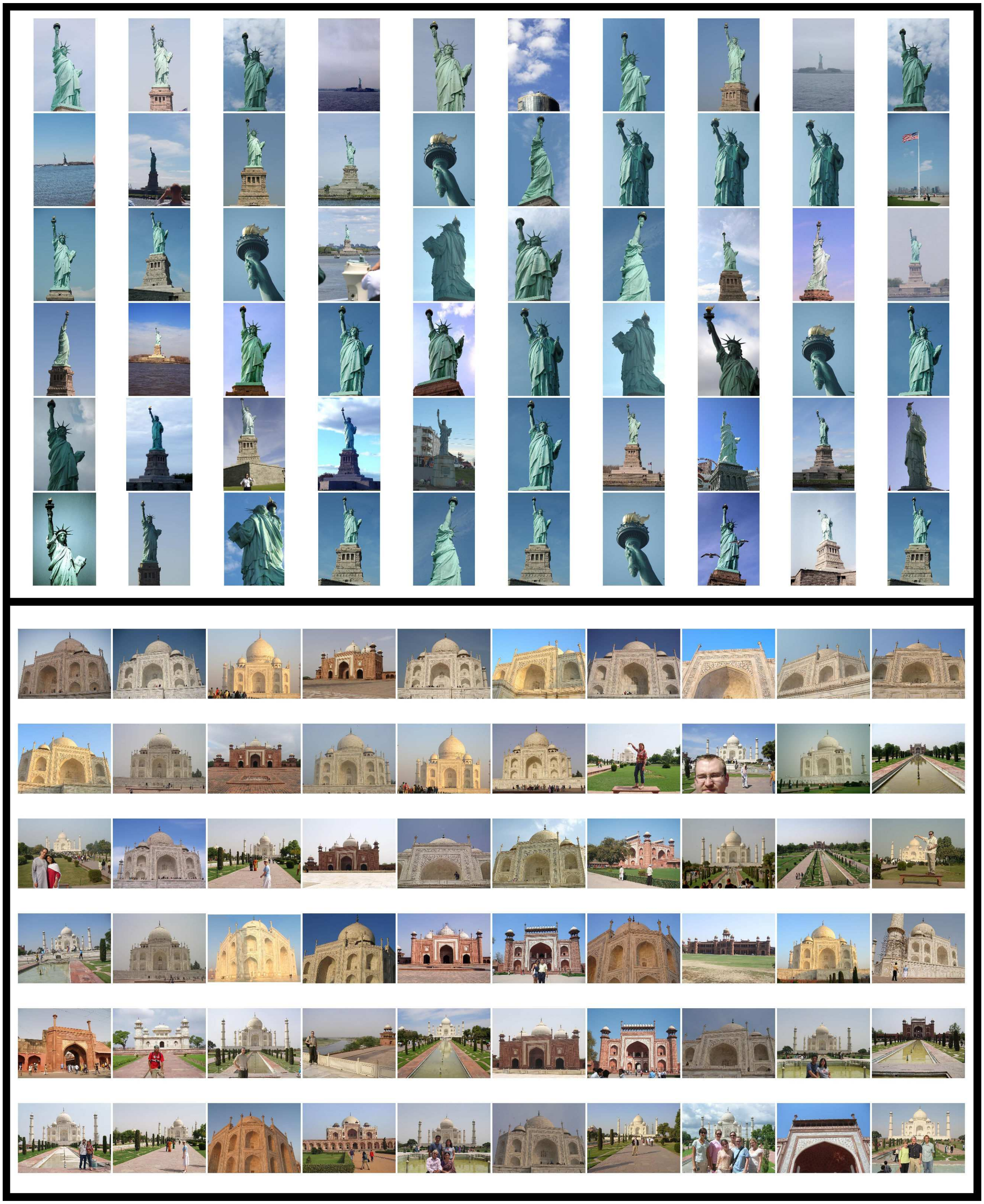

Figure 2. The top 60 ranked images (ranked left to right) for the `statue of liberty` and `taj mahal` categories. Several iconic representations of the statue of liberty are highly ranked including the iconic torch. Images of the taj mahal are highly ranked despite color variations. Some of the highly ranked buildings are incorrect, consisting of pictures of another (red) building on the `taj mahal` grounds because this building is similar in appearance and shape. Errors like these might be difficult for non-domain experts to spot.

	With Segmentation			Without Segmentation			Random		
category	1s	2s	3s	1s	2s	3s	1s	2s	3s
pyramid	0.7879	0.1919	0.0202	0.4242	0.3636	0.2121	0.2600	0.2300	0.5100
lincoln	0.7273	0.2121	0.0606	0.4200	0.3000	0.2800	0.3061	0.2959	0.3980
chrysler	0.6417	0.1917	0.1667	0.3000	0.3917	0.3083	0.2500	0.3083	0.4417
statue	0.6364	0.2818	0.0818	0.4909	0.2545	0.2545	0.2110	0.3211	0.4679
taj	0.5152	0.2525	0.2323	0.4227	0.2784	0.2990	0.2727	0.2727	0.4545
sphinx	0.3737	0.3232	0.3030	0.4286	0.3571	0.2143	0.1579	0.2316	0.6105
sydney	0.2828	0.2929	0.4242	0.2900	0.2600	0.4500	0.2800	0.3300	0.3900
capital	0.2653	0.1735	0.5612	0.1684	0.1474	0.6842	0.1250	0.1354	0.7396
empire	0.1700	0.3300	0.5000	0.2300	0.2600	0.5100	0.1400	0.2800	0.5800
average	0.4889	0.2500	0.2611	0.3528	0.2903	0.3569	0.2225	0.2672	0.5102

Table 1. Results of our user study. Users were asked to rate randomly sampled images the top 100 images for each type of ranking as to how well they represented each category where 1 corresponded to "Very Well", 2 "Moderately Well", 3 "Poorly", and 4 "Don't know". The above numbers correspond to the percentage of each rating by the users for our ranking with segmentation (**1st 3 columns**), ranking without segmentations (**2nd 3 columns**), ranking randomly (**3rd 3 columns**). As can be seen from the random results, almost half the images collected from Flickr are judged to be poor representations of the category. So, being able to select the good images from among these is an important task. Our ranking that incorporates segmentation information performs better than both a random ranking and the ranking without segmentation on 6 of the 9 categories and does quite well on several of the categories (`pyramid`, `lincoln memorial`, `chrysler building`, `statue of liberty` and `taj mahal`). For example, 79% of the top 100 rated `pyramid` images received ratings that they represented the category "Very Well" and 73% of the top 100 `lincoln memorial` pictures were rated "Very Well". From these numbers we can see that segmentation makes a clear, obviously useful difference for our system. Other categories such as the `sydney opera house` and the `empire state building` are more challenging because the object is often presented only in cluttered scenes where a segmentation into figure/ground is quite difficult. None of the rankings perform very well on these images.

4.1. Ranked Images

In figure 2 we show the top 60 ranked images (ranked left to right) for the `statue of liberty` and `taj mahal` categories. These images have been ordered using our method of ranking which includes figure/ground segmentation information. Many of the top ranked images from these categories correspond to good representations of the category. Several of the highly characteristic aspects are represented in these images including the highly iconic torch. Images of the Taj Mahal are highly ranked by our system despite color variations depending on time of day. A few of the highly ranked buildings are incorrect, showing images of another (red) building on the Taj Mahal grounds. This building is highly ranked because it has a very similar appearance and shape. Errors like these might be difficult for non-domain experts to spot.

In figure 3 we show the top ranked images using segmentation, the top ranked images without using segmentation, and the top images for a random ranking of the images (separated by red lines). The four quadrants each show a different test category where the upper left contains images from the `chrysler building` category, the upper right the `lincoln memorial`, the lower left the `pyramid` category and the lower right the `capital building` category.

For the `chrysler building` category the difference between our ranking including segmentation (top), the rankings without segmentation (middle), and the random ranking (bottom) is startling. Our method is able to extract images containing iconic photographs of the building whereas the two other rankings show views where even if the building is present, it is present in a much less iconic context. The ranking without segmentation seems to select images that have approximately the right overall make-up (when judged based on color for example), but since it is considering the whole image equally it is not able to make the distinction between skyline images and iconic close up images containing only the Chrysler building.

Our rankings for the `lincoln memorial` and the `pyramid` category are also significantly better than those of the random ranking and the ranking without segmentation. For the `lincoln memorial` category, we are able to rank multiple characteristic aspects (both the outdoor view of the memorial building and the inside view of Lincoln's statue). Even though the method of ranking without segmentation was presented with the same training images it still produces a much less compelling ranking. This is true for the `pyramid` category as well.

`Capital building` was our most muddled category. This was partially due to the fact that during collection we had in mind images depicting the U.S. **Capitol** building in Washington D.C., but incorrectly spelled the query as **capital** building. The term capital building can be used to refer to any state (etc) capital building. Therefore, the images collected tend to depict different capitals from around the

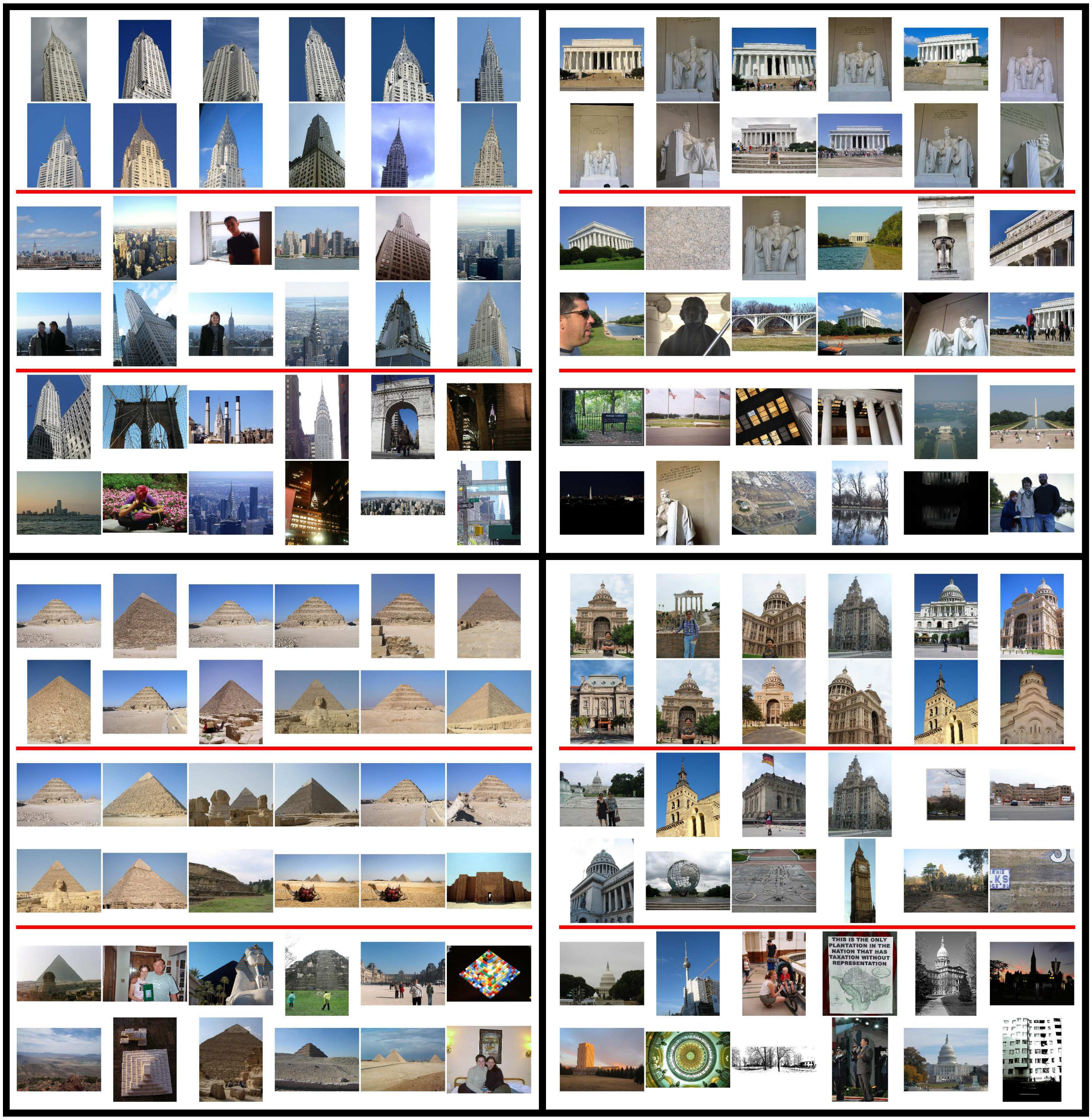
THIS IS THE ONLY PLANTATION IN THE NATION THAT HAS TAXATION WITHOUT REPRESENTATION

Figure 3. Each quadrant contains images from a category of objects ranked in three ways (separated by red lines): by our system using appearance and shape of segmented objects, by a ranking using appearance similarity across the whole image, and by a random ranking of images within the category. The **upper left** quadrant contains images from the `chrysler building category`, the **upper right** the `lincoln memorial` category, the **lower left** the `pyramid category`, and the **lower right** the `capital building` category. Notice that our system performs quite favorably compared to the appearance and random based rankings. For some categories (`chrysler building`, `pyramid`, `lincoln memorial`) it does quite well. Notice that for the `lincoln memorial` class we are able to rank multiple characteristic aspects (both the outdoor view of the Memorial and Lincoln's statue). The ranking without segmentation performs much less favorably on this category because it has no information about what areas of the image need to be similar (the regions containing the object) and which can vary (background). This is also true for the `chrysler building` in which the ranking without segmentation seems to pick images based on their color similarity rather than images that share a common object. Even for the somewhat ill-defined `capital building` category, our system finds domed buildings, many of which are capital buildings of various locations.

globe including the Wisconsin, and Texas capital buildings. Many of these buildings actually have similar shapes to the U.S. Capitol building and so are hard to distinguish. As can be seen in figure 3 the top images ranked for this category don't all depict the U.S. Capitol building, but do tend to be photographs of quite similar looking domed buildings.

4.2. User Ranking

We also judge our performance based on user ratings. Twenty-three volunteers (mostly graduate and undergraduate students) with no idea of the purpose of the experiment were asked to label a random selection of images sampled from the top 100 images from each type of ranking. For each image, the user was asked to label it according to how well it represented a category, where 1 corresponds to a rating of "Very Well", 2 to "Moderately Well", 3 to "Poorly", and 4 to "Don't Know". Besides the written instructions we also provided a visual aid of several example images from a training category, eiffel tower, labeled as 1, 2 or 3.

We show the tallied results for each of the three rankings in table 1. For each ranking method and for each category, the table shows the percentage 1s, 2s, and 3s assigned to the top 100 images from that ranking.

According to the numbers for the random ranking, about 50% of the images that we collected from Flickr are judged to be poor examples of the category name. Being able to automatically select the high quality images from this noisy set is an important and nontrivial task.

If we measure performance as the percentage of the 100 top-ranked images that received a rating of 1, then we see that our ranking with incorporated segmentation information performs better than both a random ranking and the ranking without segmentation on 6 of the 9 test categories. We do quite well on several of the categories (`pyramid`, `lincoln memorial`, `chrysler building`, `statue of liberty` and `taj mahal`). For example, 79% of our 100 top-ranked `pyramid` images receive ratings indicating that they represent the category "Very Well" and 73% of our 100 top-ranked `lincoln memorial` pictures are rated "Very Well". From these figures we can see that segmentation makes a clear, obviously useful difference for our system.

Other categories such as the sydney opera house and the empire state building are more challenging because the object is often presented only in cluttered scenes where a segmentation into figure/ground is quite difficult. None of the rankings perform very well on these images.

We use a t-test to determine whether the difference in sample means is significant for the three different ranking methods. The t-test is calculated as the ratio of the difference between the sample means to the variability of the values. We compute this for the average percentage of images ranked as representing the category "Very Well" (labeled as 1s). For our ranking including segmentation versus the ranking without segmentation, t is calculated to be 1.6429, giving about a 7% chance of observing these results if the means were actually equal. For the our ranking with segmentation versus the random ranking, t is calculated to be 3.05 or about a 3% chance of observing these results given equal means. This suggests that the difference between our ranking and the two alternative methods is a statistically significant difference.

Some comments that the users had were related to the confusion in exactly what the definition of a category is. They were presented with just the category name and so some were unsure how to rate images showing other versions of the category than the standard meaning (*e.g.* photographs of a sphinx house cat in the sphinx category). There was also much confusion about the capital building category mostly because of the capitol, capital problem mentioned previously. Most users labeled images with the U.S. capitol building in mind rather than the broader definition of capital building.

5. Conclusion & Future Work

We have shown that it is possible to rank images according to how well they represent a given category. We use the fact that iconic representations of a category should appear with high frequency and similar appearance in a set of images linked by the fact that they have all been associated with a common label (Flickr tag). We have also demonstrated that incorporating a rough idea of where the object is located in the image can improve our performance significantly.

The user comments we received reinforce the fact that notion of a category is a confusing and slippery thing. More study should be put into determining what is meant by a category.

For future work we would like to rank images in a completely unsupervised manner. We tried various methods of ranking including clustering and ways to select ground truth images according to how iconic they seemed or how similar they were to the bulk of images. None of our attempts were successful and seemed to indicate that this is a harder problem than it might seem. One last thing we would like to work on is some functional definition of iconicness according to perceptual cues of figure/ground like surroundedness and above/below.

References

[1] J. Bach, C. Fuller, R. Humphrey, and R. Jain. The virage image search engine: An open framework for image management. In *SRIVD*, 1996. 1

[2] S. Banerjee and B. Evans. Unsupervised automation of photographic composition rules in digital still cameras. In *Conf*

on Sensors, Color, Cameras and Systems for Digital Photography, 2004. 1

[3] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, , and D. Forsyth. The effects of segmentation and feature choice in a translation model of object recognition. In *CVPR*, 2003. 2

[4] Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *PAMI*, Sept. 2004. 1, 2

[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *ICCV*, Oct. 2005. 1

[6] R. Fergus, P. Perona, and A. Zisserman. A visual category filter for google images. In *ECCV*, May 2004. 1

[7] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *PAMI*, 6:721–741, 1984. 1

[8] T. Gevers and A. Smeulders. Content-based image retrieval by viewpoint-invariant color indexing. In *IVC*, 1999. 1

[9] Y. Ke, X. Tang, and F. Jing. The design of high-level features for photo quality assessment. In *CVPR*, 2006. 2

[10] B. Leibe, K. Mikolajczyk, and B. Schiele. Segmentation based multi-cue integration for object detection. In *BMVC*, 2006. 2

[11] J. Li and J. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *PAMI*, pages Vol 25, no. 9, 1075–1088, 2003. 1

[12] J. Li, J. Wang, R. Gray, and G. Wiederhold. Multiresolution object-of-interest detection of images with low depth of field. In *CIAP*, 1999. 1

[13] J. Luo, S. Etz, A. Singhal, and R. Gray. Performance-scalable computational approach to main subject detection in photographs. In *HVEI VI*, 2001. 1

[14] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *ACM multimedia*, 2001. 1

[15] R. Veltkamp and M. Tanase. Content-based image retrieval systems: A survey. In *T.R. Utrecht University*, 2000. 1

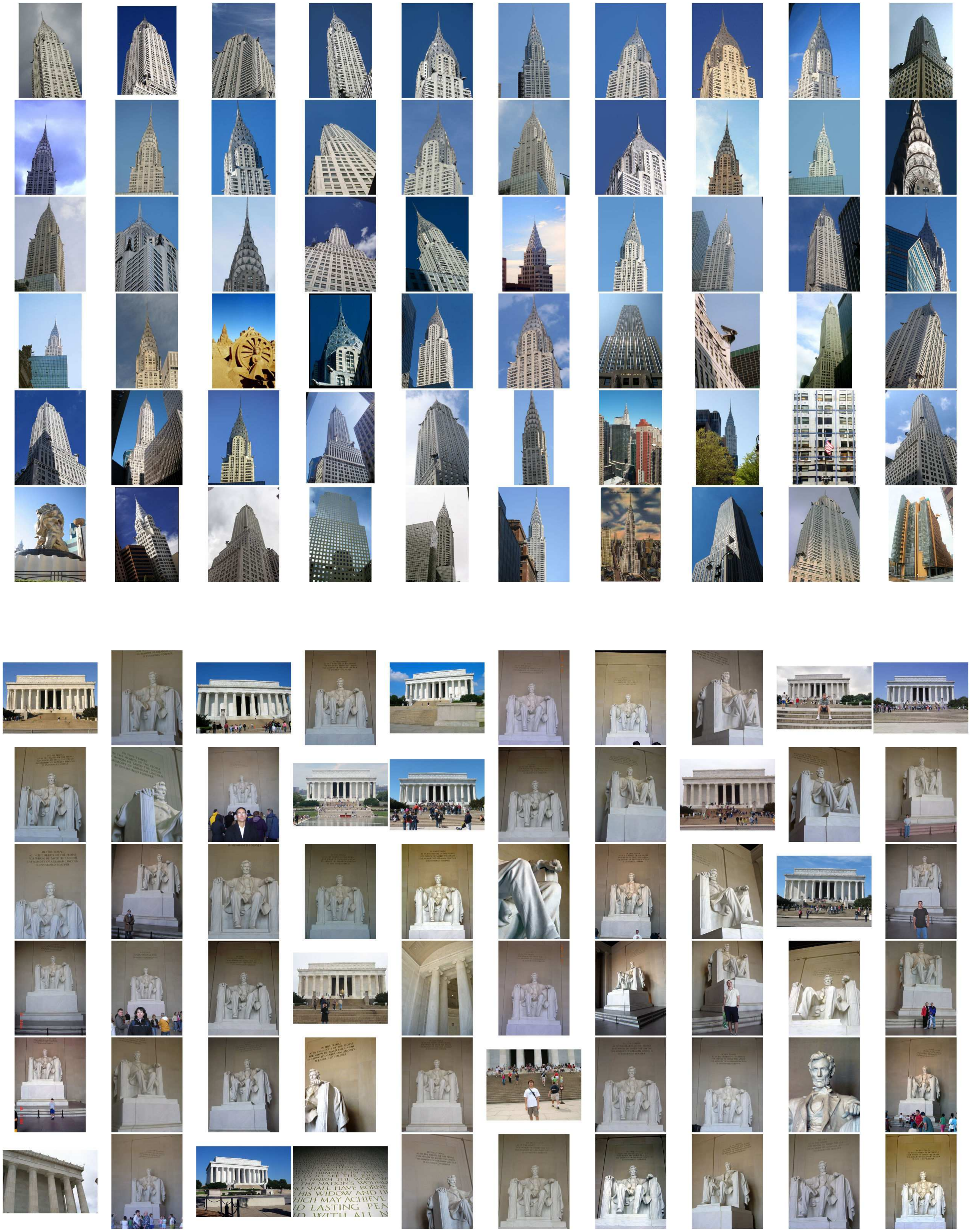

Figure 4. The top 60 ranked images (ranked left to right) for the `chrysler building` and `lincoln memorial` categories.

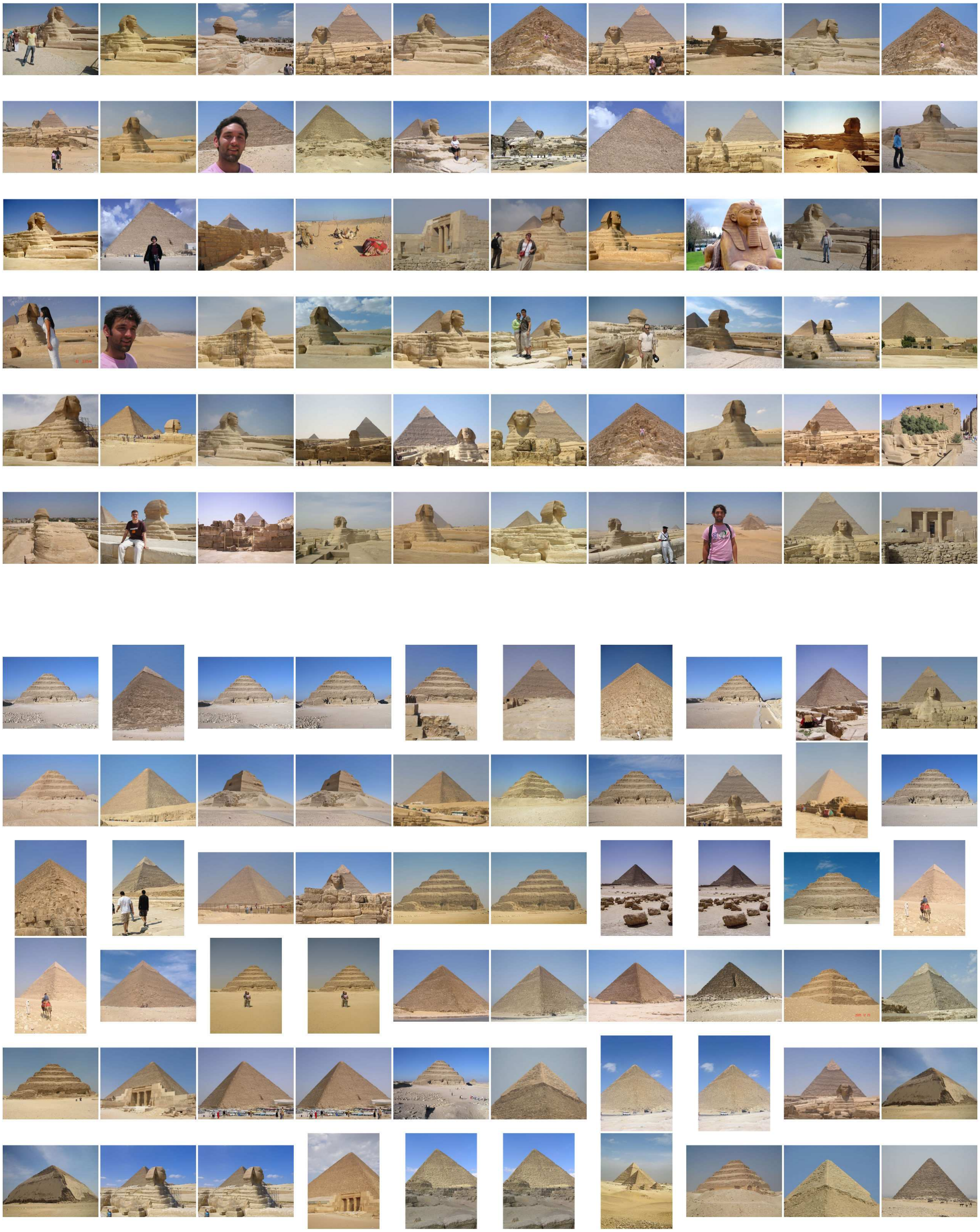

Figure 5. The top 60 ranked images (ranked left to right) for the `sphinx` and `pyramid` categories.

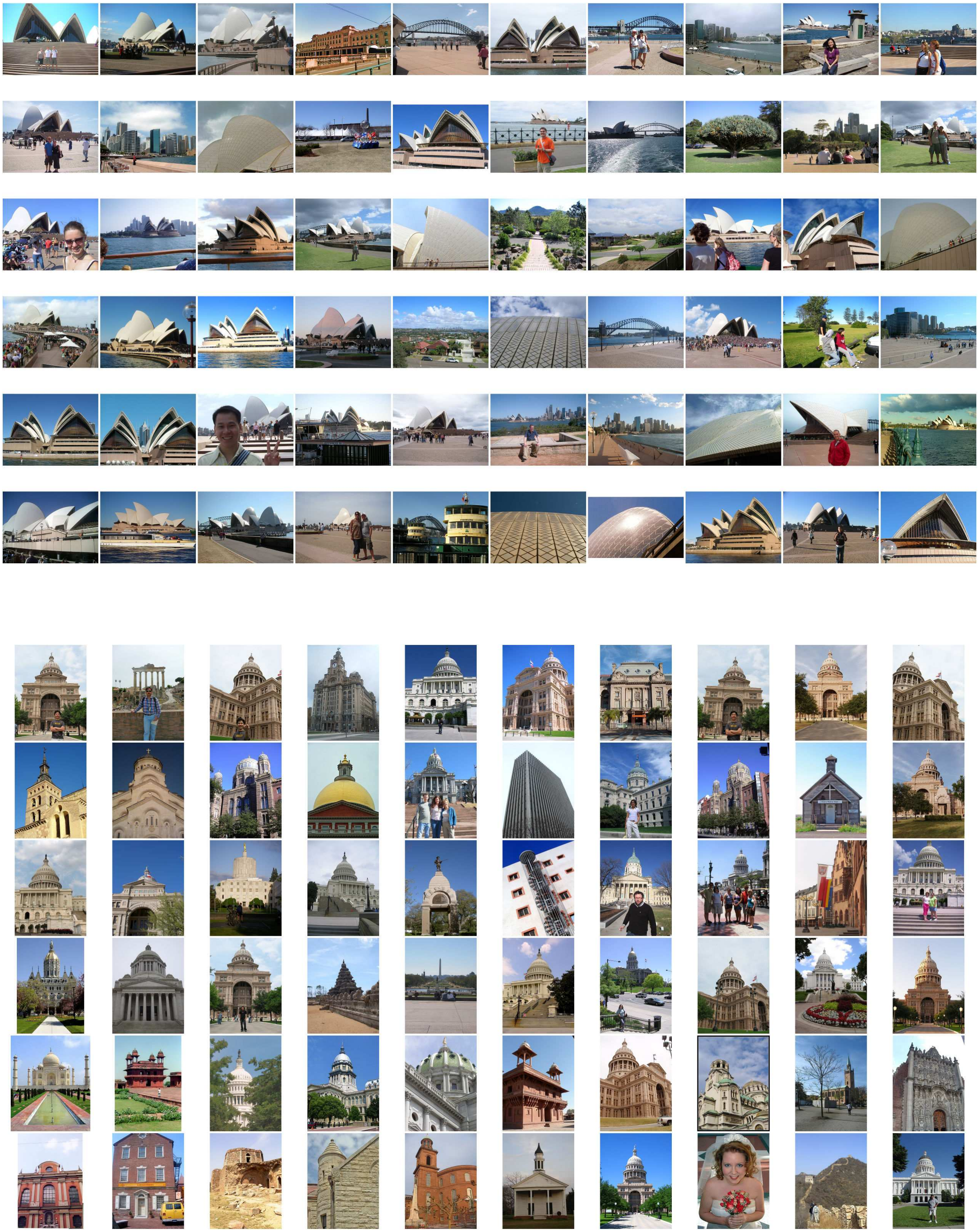

Figure 6. The top 60 ranked images (ranked left to right) for the `sydney opera house` and `capital building` categories.