Hierarchical Photo Organization using Geo-Relevance

Boris Epshtein
Microsoft Corporation
One Microsoft Way
Redmond, WA, USA
Borisep@microsoft.com

Eyal Ofek
Microsoft Corporation
One Microsoft Way
Redmond, WA, USA
EyalOfek@microsoft.com

Yonatan Wexler
Microsoft Corporation
One Microsoft Way
Redmond, WA, USA
Yonatan.Wexler@microsoft.com

Pusheng Zhang
Microsoft Corporation
One Microsoft Way
Redmond, WA, USA
pzhang@microsoft.com

(a) (b) (c)

Figure 1: Organizing image collections. (a) Typical unorganized display of thumbnails. (b) Images organized according to camera positions (GPS). (c) The proposed hierarchical organization allows meaningful image browsing according to scene semantics. Note that images of the same object may appear at random positions in (a), in different positions based on the camera location in (b) but are placed in the same sub-tree in (c).

ABSTRACT

We present a novel framework for organizing large collections of images in a hierarchical way, based on scene semantics. Rather than score images directly, we use them to score the scene in order to identify *typical views* and *important locations* which we term Geo-Relevance. This is done by relating each image with its viewing frustum which can be readily computed for huge collections of images nowadays. The frustum contains much more information than only camera position that has been used so far. For example, it distinguishes between a photo of the Eiffel Tower and a photo of a garbage bin taken from the exact same place. The proposed framework enables a summarized display of the information and facilitates efficient browsing.

Categories and Subject Descriptors

H.3.1. [**Content Analysis and Indexing**]. H.3.7 [Digital Libraries] Image Indexing and Browsing.

General Terms

Algorithms, Management.

Keywords

Image organization, Clustering, Image databases.

1. INTRODUCTION

The use of digital photography and the internet enables the accumulation and sharing of huge collections of images. A 2003 U.S. Consumer digital Imagery Survey reported that 17% of its respondents take between 3,000 -6,000 images over a 5 years period. Organizing such large collections of photographs is very difficult: unless the user invested in comprehensive tagging of the

images, the best representation of the image content is the image itself.

We address this challenge by leveraging on recent success in the Computer Vision field to recover the viewing frustum of a large collection of cameras (known as camera calibration) [6]. This allows us to take into account not only where photos were taken but also the camera direction and the field of view. These in turn provide us the viewing frustum of each camera. Knowing *what* the photographer chose to depict allows to score places instead of photographs. This in turn allows for hierarchical representation of a scene that is based on actual geometric and geographic scoring which we term Geo-Relevance. Note that this representation is not very sensitive to the accuracy of recovered parameters and in many cases approximate values can be used.

The main shortcoming of current GPS-based approaches is that they are camera-centered, instead of scene-centered. This results in grouping that depends on the camera position instead of the image content. For example, a photo taken near the Eiffel tower may or may not contain the monument itself. By adding two more parameters (viewing direction and angle) per image, we can transform the problem to become scene-centric and rank elements within a scene by importance rather than camera position. By incorporating this information into the grouping process we can make a decision that is scene-centric. The added parameters can be automatically recovered for thousands of images using recent advances in the computer vision field. Alternatively they can also be easily estimated using a simple user interface. Note that we do not rely on object recognition or full 3D reconstruction, which are still under research and so this intermediate representation is useful given the current state of the art methods.

Efficient retrieval of images taken at a specific geographic location is provided nowadays by a number of recently introduced web services and applications, usually by means of geographic tagging. However, handling the retrieved images is hampered by a number of problems:

1. **Large number of images** is associated with locations of high interest. For example, searching for the images of the Statue of Liberty in Google results in about 60,000 images. Currently, these images are displayed over many pages of thumbnails that the user needs to scan manually. A photo containing the statue may appear next to a photo where the statue is somewhere in the background, and next to a photo that doesn't contain the statue at all but was taken next to it.
2. **Grouping of images** that represent the same object in the scene, allows efficient display and browsing of large collections of images. An image might display multiple objects, and thus these grouping should not be exclusive. Image groups that encompass large number of images need to be further subdivided into sub-groups, producing a whole hierarchy of groups and sub-groups.
3. **Representative images** of groups of images need to be chosen manually.
4. **Discoverability** of images in flat collections, either as thumbnails or as in PhotoTourism [6][8] does not work well when large number of images are involved. This suggests a meaningful clustering to organize the data.

We address these problems in the proposed method by incorporating slightly more information (the frustum) than the typical GPS-based approaches. This allows us to perform voting on the scene and subsequently infer a natural, geographically meaningful hierarchical organization of the photos.

1.1 Contribution

The geometric parameters of the cameras, (i.e. their orientation and position) represent the intent of the photographers. Areas that are more interesting are likely to be photographed more. Some views of a scene are expected to be more popular than non-traditional ones. We propose the notion of *Geo-Relevance* that builds upon these observations. In order to compute the geo-relevance for each point of the scene, we use a voting approach. Each camera votes for each point visible from its point of view. Objects that are visible by larger number of cameras will get more votes.

We present a system that enables the discovery of salient geographic objects by accumulating relevance from a set of geo-positioned images, and chooses representative images for those objects. We generate a hierarchical organization of the images. For example, the hierarchical organization of a collection of images of a cathedral might originate from a root node that represents the entire cathedral, while a sub group of images might be clustered into a node that represents a tower, or the entrance gate, etc. At each level of the hierarchy, the existence of child nodes designates the presence of finer level details in the collection.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 presents the notion of geo-relevance of a scene, selection of interesting points in space, and hierarchical subdivision of the collection based on the relevance score. We show experimental results in section 4 and conclusion remarks in section 5.

2. RELATED WORK

Photo organization is the subject of a large number of studies, such as the work of Rodden and Wood [5]. A variety of meta-data, such as capture time, photographer ID or annotations can be used to facilitate organization [1].

In particular, there is a growing body of work on organizing images according to photographer or subject location. Toyama et al. [7] browses images, based on their location on a 2D map. Such map based representation of image thumbnails, suffers from the nature of the picture distribution: some areas may contain more images than can be displayed while others areas might be sparsely populated.

Time Quilt [2], shows the advantage of browsing large image collections by grouping images according to their capture time and representing them by a representative image. However, they do not suggest a way to generate a good representative image.

Naaman et al. [3] cluster images based on the camera location and time of capture. Their hierarchical organization enables more efficient display and search. However, location of a photographer provides limited information on the content of the picture: The subject might be far from the photographer. The authors represent each cluster by a geographic name (usually at coarse scale) and do not select representative images for clusters.

McCurdy and Griswold's RealityFlythrough[3], Snavely et al.'s PhotoTourism [6] and PhotoSynth [8] organize the images in 3D space using full camera parameters. The former uses GPS and orientation sensors while the latter two recover relative parameters by matching features between images. The 3D collections can be browsed using positions of the cameras or the images footprints. These applications are designed to give nice smooth transitions inside a dense cloud of images.

The aforementioned methods do not cluster similar images, thus an access to a specific image in a dense scene might be difficult, as well as discoverability of small details. For example, to be able to discover that one building out of a row of buildings is covered by a closeup images, the user needs to move the mouse over the house, or to move the virtual camera to a close position to the house.

3. OVERVIEW

We would like to facilitate an access to every image in an image collection. On the other hand, we would like to avoid cluttering the display with a large number of thumbnails. To be able to do so, we organize the images in a hierarchical tree, in a way that will represent the semantics of the scene. For example, given a set of images of the Statue of Liberty, we would like to represent them by one overview image of the statue. Browsing further into that collection, we may see representative images of some semantic sub parts, such as the statue head. As we continue browsing into the collection we will eventually see the individual images. At each stage of the interaction, we maintain a limited number of representative images shown to the user. Such organization facilitates discoverability of finer level details as they are represented by the existence of child nodes.

The Algorithm:

1) Define a grid of voting cells on which we accumulate geo-relevance

2) For each input image
   a) Enter the frustum data by the user (in the form of a camera position and an estimated view direction), or recover the frustum data by matching the image to other geo positioned images or models.
   b) Add the contribution of the image frustum to the geo relevance (Section 3.1)

3) For each level of the hierarchy:
   a) Find local maximum points of geo-relevance using mean-shift. (Section 3.3). If we are limited to a display of up to *k* thumbnails per hierarchy level, we will hold the best *k-1* of the maximum points for the current level of the hierarchy, and push the rest to a *k*-th uncategorized node.
   b) Define interest object of each maximum, and find the images that are associated with that object. (Section 3.4)
   c) Find a representative image per interest object (Section 3.5)
   d) Continue to the next level of the hierarchy, and further split the objects to sub-objects.

3.1 Geo-Relevance

Semantic analysis of a scene from images is still an open problem. We address it by relying on the judgment of the multiple photographers that took the actual images. Each photographer aimed the camera according to his perception of what is important in the scene. We define the notion of *Geo-Relevance*, to measure the importance of a spatial point. The geo-relevance of a scene point is proportional to the number of images where that point appears. To estimate the geo-relevance of each point, a voting approach is employed.

The camera parameters associated with each image define a 3D frustum. All objects visible in the image lie within that frustum. In general, we do not know what specific object is the subject of an image. For example, it could be the person in front, or it could be some house in the background. In absence of better information, we add equal amount of contribution to each point within the frustum.

For simplicity, consider a 2D example (although the 3D case and even 4D, temporal data, is completely analogous). Given a position, orientation and internal parameters of one camera, we can draw the projection of the camera's frustum on the ground plane. See figure 5(a). The frustum area represents all the space points that might appear in the image (we will discuss visibility issues later).

A grid of voting cells is used to accumulate the geo-relevance. The frustum of each image, contributes to each cell it covers. The frustum contribution could be uniform (as used in the current implementation) or weighted. For example, one could assume that subjects of an image tend to lie toward the middle of the frame hence weighting this area differently. The focusing distance can be factored in as well, if it is known.

The frustum could be infinite in length, capped by some maximum distance (for example, based on the focal distance of the camera), or intersected with the scene geometry. Scene geometry can be obtained from different sources: Earth terrain model can be obtained from US Geographical Service, models of urban buildings are provided by Google Earth or Microsoft Virtual Earth. Alternatively, an approximate model can be recovered from the images using stereo, as used in this implementation, or provided by user input, etc.

We use the computed geo-relevance as an estimate of the importance of each point. Figure 5(a) shows a geo-relevance score for a sample scene, and a highest relevance position, marked by the red square. We could assume that the peak in the score corresponds to the most important object in the scene.

3.2 Scene Hierarchy

Organizing a collection in a hierarchical fashion enables an efficient access to each image. Rather than sifting through a large collection of thumbnails, the user can reach each image in the collection by traversing a tree of image sub-collections, organized by the geographic scale. For example, a large collection of images of Paris could be divided to smaller sub-collections, including, for example, the Eiffel Tower, the Notre Dame cathedral, etc. The Cathedral sub collection could be further divided to smaller sub collections, such as the main gate, towers, and so on. Such an organization requires knowledge of the structure of the scene. In the absence of such knowledge, the only data that we could use are the images, and the parameters of the cameras.

PhotoTourism [6] showed impressive capability to recover a cloud of 3D points from matches between images of the same scene that were taken by different photographers at different times. Could have such a cloud of points be used to segment the scene to meaningful parts? The cloud of point does have some similarity to the scene geometry. However, the density of those points depends on the texture of the scene objects (smooth surfaces contribute less points), additionally, many images could not be matched (even if their position is known). Even more, the matching process is computationally intensive, which limits the number of participating images. We would like to be able to organize a large collection of images, with known orientation (given, for example, by the users), even without the need to recover structure of the scene.

Given the geo-relevance, one can identify the interesting parts in the scene, without requiring the full reconstruction of the scene structure. Segmenting the scene based on such information will tend to associate images with the nearest interesting spot. Using this segmentation we are able to group together images that show the same subject, or share similar view directions, such as common in many touristic locations.

3.3 Relevance-based Scene Hierarchy

Hierarchical organization of images allows quick browsing. It is desirable to subdivide the image collection into a few sub-collections, each represented by a representative image. We achieve this in two steps: building of a scene hierarchy using geo-relevance and association of images with corresponding hierarchy nodes.

At each level, we subdivide the scene using the mean shift algorithm [9]. Initially a set of points are spread on a regular grid on top of the geo-relevance image (See Figure 6a). The points iteratively update their position using mean-shift: for each point P its updated position P' is computed as the weighted average of points in its neighborhood $N(P)$ of radius r. The points T are weighted by the value of their geo-relevance $R(T)$:

$$P' = \frac{\sum_{T \in N(P)} R(T) * T}{\sum_{T \in N(P)} R(T)} \qquad (1)$$

The mean shift algorithm will converge into the various relevance maxima. Each point will thus move to its basin of attraction, as shown in Figure 6(b). The basins induce the clustering and their size is directly related to the radius r. The value of r is chosen such that the initial number of clusters is below some set value (We used the limit of 5) . All basins can be used as clusters, or only the top k. The remainder, and all points whose neighborhood contains too little relevance, are associated with a one cluster. The process is repeated recursively in subsequent levels where each finer level uses a neighborhood radius r of ½ size used in the previous level, thus producing finer details.

The result is a hierarchical tree representation of the scene.

3.4 Image Collection Association

Once the scene is clustered, each image needs to be associated with a cluster.

Let us define the object of interest within each cluster by thresholding the geo-relevance, so that the object area contains a fixed percentage of the accumulated relevance of the cluster area. In the current implementation, we computed a threshold so that 50% of the relevance score of the region belonged to the object.

To decide whether a frustum should be associated with the cluster R, if the projection of the objects-of-interest of R covers a substantial area of the image. A straight forward way to perform this, is to calculate the 2D view angle of the objects relative to the view angle of the frustum (see Figure 3). In the current implementation, we associate images where the object of interest covers at least 75% of the frustum view angle.

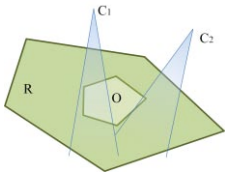


Figure 3: Camera C1 is regarded as a camera that sees the interest object O of cluster R, since the projection of O onto the camera's image will cover most of the image. Camera C2, on the other side, sees O, but on a small portion of its image, thus it is not associated with cluster R.

Images that do not show any of the sub - objects are associated with the uncategorized cluster.

3.5 Representative Images

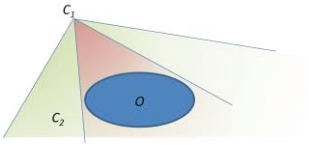


Figure 4: Selection of frustum that best represent an object. The cells that belong to the object are represented by the area O. The score of a specific frustum is the sum of the relevance of all object cells that falls in the frustum, normalized by the view angle. Normalization by the view angles prefers view angles that fit the object, thus preferring red frustum C2 over green frustum C1.

Given a collection of images, associated with cluster N_i, we select one of the images that will represent the cluster. We look for an image that best covers the interest objects associated with the cluster. Other criteria can also be incorporated into this selection, such as images with best contrast, sharpness, etc.

The representing image is chosen by searching for one of the images in the group that maximize the viewed relevance. We apply the following score for a frustum:

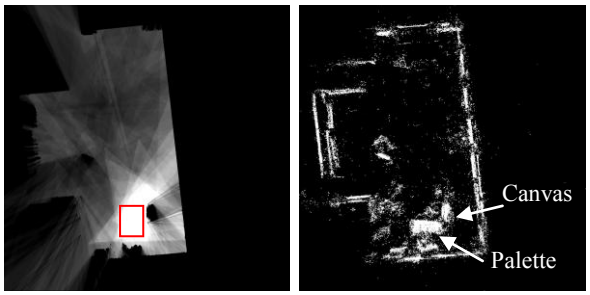
$$score = \frac{1}{viewAngle} \sum_{(Frustum \cap ObjectArea)} S_i. \qquad (2)$$

The score of a frustum is the sum of the relevance of all grid cells S_i, that belong to the object and are covered by the frustum, normalized by the view angle (See eq. 2 and Figure 4). The normalization is used to prefer shots where the view angle fits the object of interest.

4. RESULTS

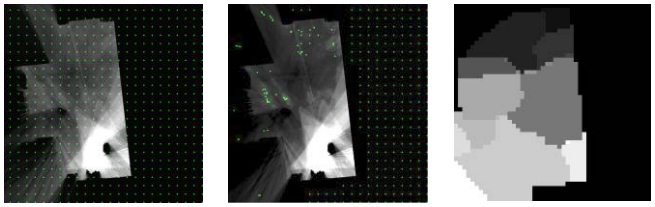
For testing, we used a collection of images, taken at the studio of the artist Faigin and their relative position and orientation, as was recovered by Photosynth [8]. This set has about 200 images, taken in an interior space, and at a large variety of orientations. We accumulated the geo-relevance using a 2D grid of 1000x1000 voting cells that bounds the studio area. For estimation of the visibility, we used a cloud of image features matched between the images (The same features that were used for recovery of the

relative camera parameters).



(a) (b)

Figure 5: (a) Geo-relevance map for the Faigin studio. (b) Feature points of the studio structure. Notice the area of highest relevance, marked by a red rectangle, is the location of the artist, in front of the canvases, contains very few recovered feature points.



(a) (b) (c)

Figure 6: Segmentation of the relevance image. a: Initial regular distribution of points. b. Points positions, after mean-shift iterations, reveals the position of local maxima of relevance. c: The original map is colored according to the final destination of the points.

The computed geo-relevance map is shown in figure 5(a). The cloud of recovered 3D points can give a hint on the structure of the studio, figure 5(b). Notice that the most interesting object in the studio, the artist at work, is not represented in the recovered cloud of 3D points, due to the fact that the artist is not a static object, and was not matched between images.

We recover points of local maxima of geo-relevance, using mean shift (Figure 6). Search radius used was 40 pixels for the first level of hierarchy and 20 pixels for the next level. We identified distinct destinations as separate clusters. The scene segmentation is displayed in Figure 6©.

We associate images with corresponding clusters. In figure 7, two top level nodes of the resulting hierarchy are shown, along with their representative images and a sample of the associated images. Note the large variety of the image view directions and consistency in the depicted subjects.

Figure 1 shows the top two levels of the hierarchy, generated for the studio of the artist Faigin. Each level is limited to 5 sub nodes. The uncategorized node is shown schematically. We displayed a sample of images that are associated with the first and third 2nd level nodes. Notice, that since the association is not exclusive there is an image of the artist in node 3 group.

Figure 7: The top level representative images for the studio scene, ordered from left to right and from top to bottom, along with their corresponding scene areas, and object of interest (white areas).

Figure 9 and 10, shows results on a different image collection, taken at Trafalgar square, London. In contrast to the Faigin set, this set of 200 images, has rather sparse coverage of the location. However, the hierarchy represents different clusters such as Nelson column, or the horseman statue.

5. DISCUSSION AND FUTURE WORK

We have presented a novel framework for hierarchical organization of image collections that follows scene semantics, as it represented in the distribution of the image frustums.

We assume that the images in the collection are all geo-positioned, that is, we know their full camera parameters. There is a growing percentage of images that have location assignment, either using a GPS or by manually positioning images to map location. We envision that in the future we will see accumulation of richer meta-data that will include full camera parameters. The sources for this information may be from cameras equipped with GPS and orientation sensors, manual entry of camera orientation or using matching to other images, such as done by PhotoSynth.

In this work, we used a 2D ground plane as a voting space. One might consider using the surfaces of the scene geometry as a voting surface. This approach has its own merits as well as limitations – for example, it would be impossible to discover interest objects that are not represented in the geometry, such as the artist in the Faigin set. Still, the combination of the two approaches seams as an interesting direction for future research.

One of the most interesting aspects of this work is the analysis of the images subjects. For example, it is possible to generate a slide show of a place that will show the most popular objects in the scene from the most popular view directions. Such a slide show can be used as an overview that represents a place as most people see it. On the other hand, we also wish to look at unique photographs that cover areas that are less seen, or show popular

objects from the most unusual view directions. Such a collection of images may exhibit a more artistic and interesting view of a place.

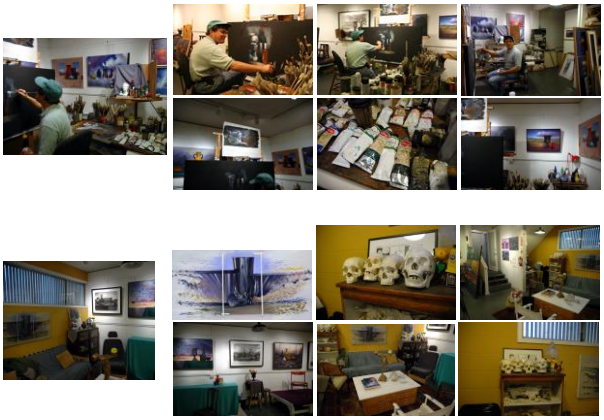


Figure 8: Two of the 1nd level sub collections. The images on the left are the representing images of the groups, and on the right, there are some of the images that were classified as viewing the same area with them. (Notice that due to our use of 2D maps for the relevance, the last picture of the top group does not saw the artist palette, but the space above it).

We wish to scale up the analysis to handle large areas and millions of images. As the contributions of more and more images will add, the geo-relevance will become more informative, and it will be possible to discover finer details. We envision the construction of a hierarchical organization of large areas, and automatic discovery of popular interest areas.

It is natural to build a hierarchical indexing structure to efficiently retrieve photos once photos are organized in the scene hierarchy. Geographic coverages of frustums also provides an opportunity to construct spatial indexes [10,11] based on spatial locality. We will explore further on indexing design on photos using this proposed framework and conduct performance study on larger collections of photos in future work.

6. ACKNOWLEDGMENTS

We would like to thanks Prof. Vincent Tao for his knowledgeable remarks. We would like to thank Dr. Drew Steedly and the Microsoft Live Labs group and Noah Snavely and the CS department of the University of Washington for their contribution of data.

7. REFERENCES

[1] M. Cooper, J. Foote, A. Girgenshon and L. Wilcox. Temporal event clustering for digital photo collections. In Proc. ACM Int. Conf. on Multimedia, 364-373, 2003.

[2] D. Huynh et al. Time Quilt: Scaling up Zoomable Photo Browsers for Large, Unstructured Photo Collections. In Proc. ACM Computer Human Interface (CHI) 2005. 2005.

[3] N. McCurdy and W. Griswold. A system architecture for ubiquitous video. In Proc. Int. Conf. on Mobile Systems, Applications, and Services, 1-14, 2005.

[4] M. Naaman, Y. J. Song, A. Paepcke and H. Garcia-Molina. Automatic organization for digital photographs with geometric coordinates. In Proc. ACM/IEEE-CS Joint Conf. on Digital libraries, 53-62, 2004.

[5] K. Rodden and K. R.Wood. How do people manage their digital photographs? In Proc. On Human Factors in Computing Systems, 409-416, 2003.

[6] N. Snavely, S. Seitz, R. Szeliski. Photo tourism: Exploring Photo collections in 3D. ACM Transactions on Graphics (SIGGRAPH Proceedings), 25(3):835–846, 2006.

[7] K. Toyama, R. Logan and A. Roseway. Geographic location tags on digital images. In Proc. ACM Int. Conf. on Multimedia, pp.156-166, 2003.

[8] PhotoSynth. http://labs.live.com/photosynth/. 2007

[9] D. Comaniciu and P. Meer, Mean Shift: A Robust Approach Toward Feature Space Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5) pp.603-619 , 2002

[10] Spatial Databases: A Tour, S. Shekhar and S. Chawla, Prentice Hall, 2003, ISBN: 013-017480-7

[11] Spatial Databases: With Application to GIS, P. Rigaux, M. Scholl, and A. Voisard, Morgan Kaufmann, 2001

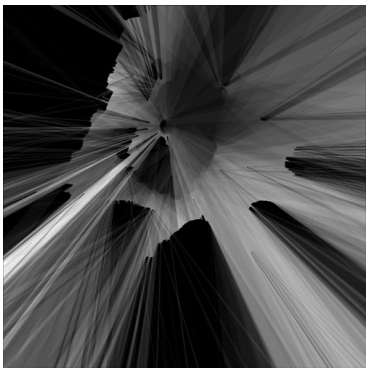


Figure 9: Geo-relevance of the Trafalgar collection. Most of the images are taken from the center of the square outwards.

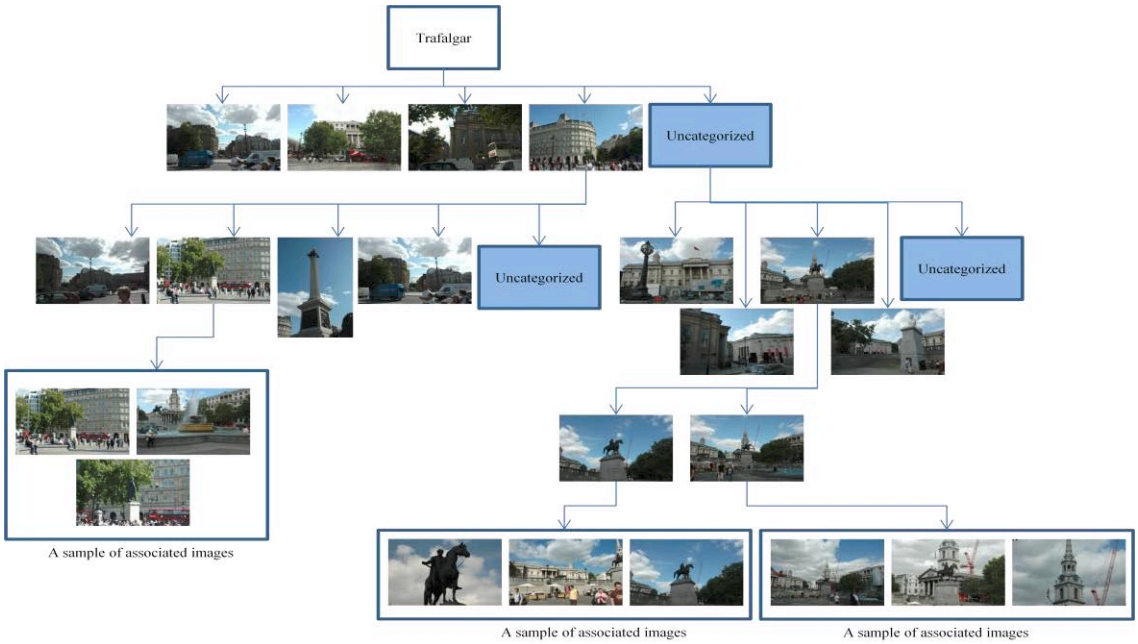


Figure 10: A hierarchical representation of the Trafalgar image collection (~200 images). Notice that the Nelson column was not chosen as a 1st level object, due to the limited coverage of it in the specific image collection. Notice that the images that cover the horseman statue (on the lower right side of the hierarchy) are split into long-shots & close-ups images of the statue