Augmenting Image Processing with Social Tag Mining for Landmark Recognition

Amogh Mahapatra[1], Xin Wan[1], Yonghong Tian[2], and Jaideep Srivastava[1]

[1] Department of CS, University of Minnesota, USA
[2] School of EE & CS, Peking University, China
{mahap022,wanxx061}@umn.edu
swantian@gmail.com
srivasta@umn.edu

Abstract. Social Multimedia computing is a new approach which combines the contextual information available in the social networks with available multimedia content to achieve greater accuracy in traditional multimedia problems like face and landmark recognition. Tian et al.[12] introduce this concept and suggest various fields where this approach yields significant benefits. In this paper, this approach has been applied to the landmark recognition problem. The dataset of flickr.com was used to select a set of images for a given landmark. Then image processing techniques were applied on the images and text mining techniques were applied on the accompanying social metadata to determine independent rankings. These rankings were combined using models similar to meta search engines to develop an improved integrated ranking system. Experiments have shown that the recombination approach gives better results than the separate analysis.

Keywords: Social Mutimedia Computing, Landmark Recognition.

1 Introduction

A landmark can be defined as an identifiable location which has some kind of geographical, cultural or historical significance. There are many landmarks across the world which can be classified at various levels, e.g. city scale landmarks to international scale landmarks.In general, every important landmark can be identified with a representative set of images. For example, say a famous monument might have some popular front view images, side view images, night images etc.

Social networks are increasing the amount of multimedia content available online every minute. The number of pictures found on flickr.com for a search on the keyword Paris produces 174,047 results (on 8/15/2010), which is significant since this provides a tremendous image corpus which can be used as a starting point for recognizing Parisian landmarks. With increasing amount of multimedia content present online, the challenge now is to analyze, index and retrieve this content. The images are usually of varying quality, illumination, social relevance and context. The accompanying social metadata is often highly noisy and inaccurate, and the number of images also keeps increasing every day.

K.-T. Lee et al. (Eds.): MMM 2011, Part I, LNCS 6523, pp. 273–283, 2011.

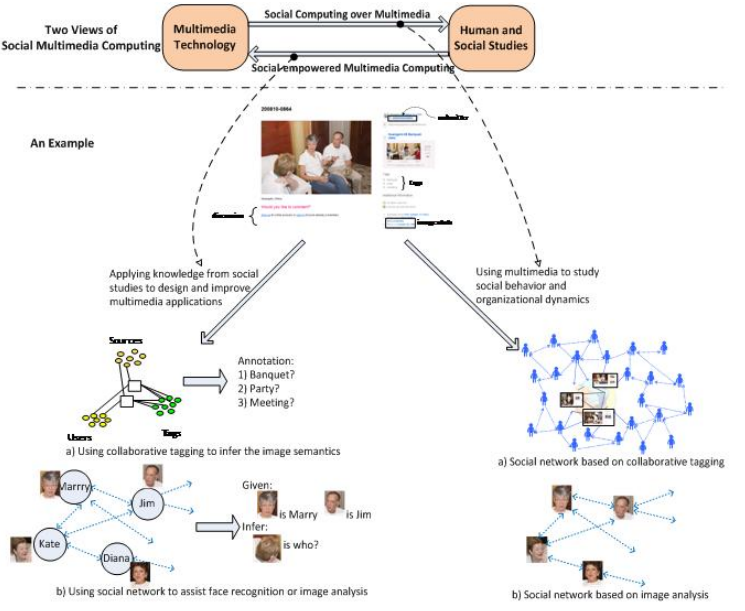

Fig. 1. The Social Multimedia Computing approach[12]

Hence online social networks, despite being rich sources of information, pose the difficult problem of information retrieval according to relevance.(Figure 1).

We will now review the work closely related to ours.Tsai et al. [11], found the context of a given set of images for a pre-defined landmark using GPS info and cell ids, and then did content analysis on images using SIFT[3] features. Their work established that context and content based analysis of social multimedia can improve the accuracy of image recognition. Simon et al.[10] use an unsupervised learning approach to create an image browser. They start with a set of images for a given location, cluster the images based on their visual properties using SIFT and create a browser where scenes can be explored. From each cluster they extract the most representative tags of the images. Their method puts forward a way of finding the most representative user tags. Naaman et al.[4] extract the landmark names across a given area using clustering based on geo-tags, from which they then extract the most representative tags for a given geo cluster. Next they extract the images corresponding to these tags and use image processing algorithms (like SIFT) on these images to find the best images. They used human experts for evaluation. Their work establishes that landmark names can be extracted using geo-tags and the use of a representative user tags as a preprocessing step can further improve the accuracy. Yan et al.[8] use a model where they first detect landmarks, and then filter images using the most representative tags. Finally they use image processing algorithms to mine and recognize landmark images. They created a web scale landmark recognition engine.

In this paper we propose a new approach which starts with a given set of images of a landmark and then clusters and ranks the images based on its visual features. Next an analysis of the tags and other social data is carried out separately. We next combine the rankings obtained from vision analysis and social data analysis using models similar to meta-search engines. In addition to proposing a new approach, we also include other available metadata in addition to geo-tags and user tags like Number of Views and Interestingness in our analysis. We believe the inherent bias in social data and the images can be properly handled using this approach.

2 Problem Statement and Proposed Approach

The Problem Statement: The problem is that of finding a set of images which can visually describe a landmark and which are diverse and yet precise at the same time. The accuracy of this image selection process is measured by how well these images can train a recognition system. If I is the set of images $I = \{i_1, i_2, .., i_n\}$ then the problem can be defined as that of finding a set of representative images I_S such that the following conditions hold, $I_s \in I, |I| >> |I_s|$.

The Proposed Approach: The images and the accompanying social data (e.g., geo-tags, user tags, number of views etc.) were analyzed separately after some initial pre-processing. The rankings generated by these different review techniques were combined using combination models similar to those used by meta search engines to obtain an integrated ranking. An image recognition system was designed to validate our tests. A test bed of images was selected,which contained a set of representative yet dissimilar images of a given landmark. The key point of our proposed approach is that instead of sequentially analyzing images and social data, we analyze them separately. We believe that this approach can handle the inherent bias present in the data extracted from the social networks. Rather than just using user tags and geo-tags we also explored other measures like number of views and Interestingness measure provided by flickr which further improved the accuracy of the proposed approach.(Figure 2)

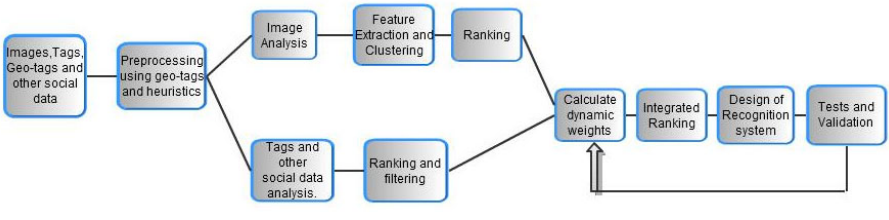

Fig. 2. The System Flow

3 Analysis of Images and Social Metadata

Our dataset is made up of the geo-tagged images downloaded from flickr.com. Some popular landmarks were selected randomly from across the world and their

corresponding images were downloaded from flickr.com (listed in Figure 5).For example, after selecting landmark X, flickr.com was queried using the keyword X.We downloaded 75,000 images for these 25 landmarks.Every image had the following metadata:Photo ID, Photo Owner ID, Title , Date taken, Date Upload, Tags, Geo-Tag(Latitude andLongitude), Views, Interestingness. Afterwards, the following heuristics were applied[4]:

1. The Photo-Owner ID was used to ensure that the number of unique uploaders for a given landmark was above a predefined threshold. This was to ensure that the number of photos included in our analysis did not end up belonging to just a few enthusiastic photographers.
2. The field date taken was used to ensure that the pictures used in our analysis did not belong to just one particular time frame; this was done to remove the local/cultural bias from the data. For example, a given landmark might have a lot of photos corresponding to a local festival.
3. Filtering using Geo Tags: Even after searching for a given keyword like Eiffel Tower from the flickr APIs the number of irrelevant images was still very high. The search results might include some popular restaurant, popular replicas etc. in other parts of the world. Hence, the latitude and longitude of these landmarks was used to filter out the irrelevant images.

3.1 Content Analysis of Images

The proposed image analysis has the following two steps. First, clustering of images was done based on the global features which results in a set of image clusters. This ensures that each cluster would contain different type of images e.g., one cluster could be of images of the landmark during the day while another could be of images taken at night. In the second step the images were ranked based on their local features. The inherent assumption is that the most representative images are similar to each other.

Clustering of Images: Same landmarks can have many images which could be very different from each other. Images could be taken from different parts of a building, from inside or outside, during day or night etc. It is hard to define the similarity between such images. At the same time, the representative image dataset should also include a diverse set of images, so that it can be used to train a recognition system which can recognize different images taken at different angles and illumination. Hence, the first step is to cluster the raw images. Two techniques were used for image clustering, Color Histogram [1] and Gabor features [2].

Color Histogram is a color quantization algorithm based on subjective vision perception. This model first transforms RGB values into HSV, and then describes the characteristics of an image using a histogram. It has relatively lower time and space complexity. For our experiments, H was divided into 8 parts, S and V were divided into three parts each. Hence, the histogram had 72 bins. Gabor wavelets are used to detect the texture features of the images. Mean and standard deviation based on six orientations and four frequencies were used to describe

the Gabor features, which ended up being a forty-eight dimensional array. The color and texture features together result in a 120 dimensional array for each image. Then, K-means algorithm was used to cluster these images into a certain number of groups. Thus, each cluster of images represents different views of the landmark.

Ranking of Images: More often than not the most representative images are visually similar. This implies that they should have a high degree of correlation. Thus the problem of finding the most representative images is the same as finding the images having high similarity with others in the same cluster.We used Scale Invariant Feature Transforms (SIFT) [3] to find the most representative images. SIFT is a local feature; it focuses on the objects in the images, and is invariant to changes in scale, rotation and illumination. We applied the definition of point-wise correspondences [4] to detect the overlap of images. So, two images that focus on similar objects will have a higher number of match points. After extracting all the interesting points from one particular image using SIFT algorithm, all these descriptors are indexed using a k-d tree.

The Best Bin First algorithm [5] was used to search for the match points. The degree of similarity between two images was quantified by the average number of point wise correspondence between one image and all the other images in the same cluster.

3.2 Analysis of User Tags

Most users tag their pictures after uploading them on a social network; if the image happens to be good, it gets tagged by the users too. It has been investigated by [9] that every landmark has a set of representative tags. Some previous approaches have used user tags as a preprocessing step, in their analysis.[8]

Our proposed approach assumes that just one tag is insufficient to perfectly identify a landmark but a landmark can be better identified by a set of tags S. For example, while searching for the most representative images of Eiffel Tower, the tag Eiffel Tower does not give the best results but if we include the images having the set of tags {'paris','france','eiffeltower','eiffel','tower','europe','toureiffel'} then the dataset happens to be more diverse and the recognition accuracy is higher too. The analogy behind this assumption has been derived from the usability aspect of search engines. For example, if we search for the keyword Java on any popular search engine then we get search results about, java the island, java the restaurant and many things other than the programming language. Hence, an end user always uses a particular set of appropriate keywords to refine his search and get to the most representative set of documents.

Find the Most Representative Tags: Noisy and misspelt tags were filtered using a standard thesaurus. Our assumption is that the most representative tags, which are invariant across seasons and local trends, should have a frequency higher than a certain threshold. Hence, a histogram analysis of the tag frequencies followed by a validation feedback loop was used to pick the most representative tags.(Figure 3)

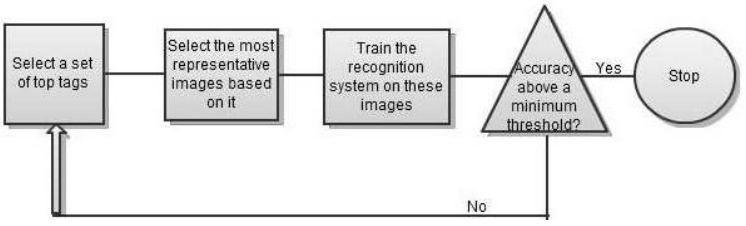

Fig. 3. Flowchart to select the best set of tags

Find the Most Representative Images: The TF-IDF model[13] was used to model the relationship between images and tags. Every image in our set was treated as a document and every tag was treated as word. Next, TF-IDF measure was used to normalize the matrix. The problem hence was reduced to finding the most relevant documents (images) that match our query vector which consists of the most representative tags as obtained previously. Latent semantic indexing was used to do this [14], where every image was treated as a vector in an n-dimensional space. Similarity of each vector was measured against the query vector and a final ranking was obtained.

3.3 Exploring the Number of Views

The data distribution of the number of views across images often gives a distribution where most images show almost the same number of views and some of them show abnormally high number of views(Figure 4). A curious observation is that the images with abnormally high number of views are usually not representative of the landmark; they usually consist of an artistic portrayal or a funny caricature like image. We found out that number of views is not a very good measure to find the most representative images.This was confirmed by training our recognition system on the most viewed images and the results turned out to be very poor.

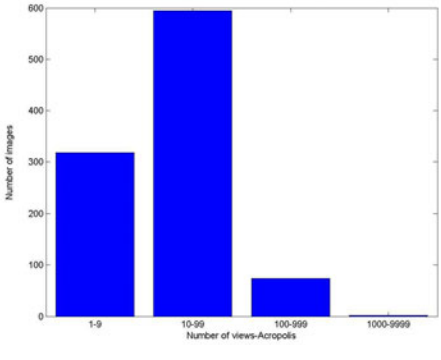

Fig. 4. No. of views of Images of Acropolis

3.4 Interestingness Measure

While downloading images from flickr.com one can specify the images to be sorted according to their interestingness. Interestingness[15] is a measure used by flickr.com to rate its images based on user comments, views, downloads etc. This measure was explored to see if it could yield representative images. After preprocessing and filtering the images we picked the most interesting ones from the remaining data set. Our recognition system was trained on these images and accuracy was measured. The results were ambiguous as the accuracy ranged from high accuracy to low accuracy. This is evident from the results given in Figure 5.

4 Combining Multiple Heterogenous Rankings

User tags, visual features and interestingness measure were used to rank the images independently.The rankings obtained by the above review techniques have different meaning, range and distribution. Hence, re-ranking models similar to those used by the meta-search engines were used to obtain an integrated ranking. The below models are not affected by the information and algorithm used by the review techniques.

The Agreement Model: The idea of the Agreement model[6] is that objects which have received a fair ranking in all reviews deserve a better ranking compared to the ones that have received a high rank in one and a low rank in another. The algorithm is as follows.Suppose we have m review techniques and n images.

1. W is the weighting vector.

$$W = [w_1, ..w_m]; \sum_{p=1}^{m} w_p = 1 \tag{1}$$

2. Agreement score of ith image (s_i) is computed as follows:

$$s_i = \sum_{j=0}^{m} \frac{w_j}{r_i^j} \tag{2}$$

   r_i^j is the ranking of ith image in the jth review.
3. Sort images according to the score.

The Fuzzy Model: Fuzzy Model [7] uses the fact that some reviews might be more important than others in certain situations. Hence, it assigns unequal weights to different reviewers. The algorithm is as follows: The weights were decided both statically and dynamically, as described in the section 5.

1. Let S_i^j be the performance judgment of the ith image according to jth reviewer:

$$S_i^j = |L_j| - P_i^j + 1 \tag{3}$$

   P_i^j is the position of ith image in the list L_j of jth reviewer.

2. Determine fitness score f_j which represents user preference of the jth reviewer $0 \leq f_j \leq L_j$
3. Define the weighting vector W(same as Equation 1) of IOWA operator [16]. The orness of the IOWA operator(for k reviewers) is defined by:

$$orness(W) = \frac{1}{k-1} \sum_{j=1}^{k} ((k-j) * (w_j)) \tag{4}$$

4. Calculate overall performance judgment S_i for each image by aggregating S_i^j using IOWA operator as follows:

$$S_i = IOWA(< u_i^1, S_i^1 >, ..., < u_i^k, S_i^k >) \tag{5}$$

u_i^j is defined based on fitness score f_j of the jth reviewer and performance judgement S_i^j. Finally, sort images according to S_i^j.

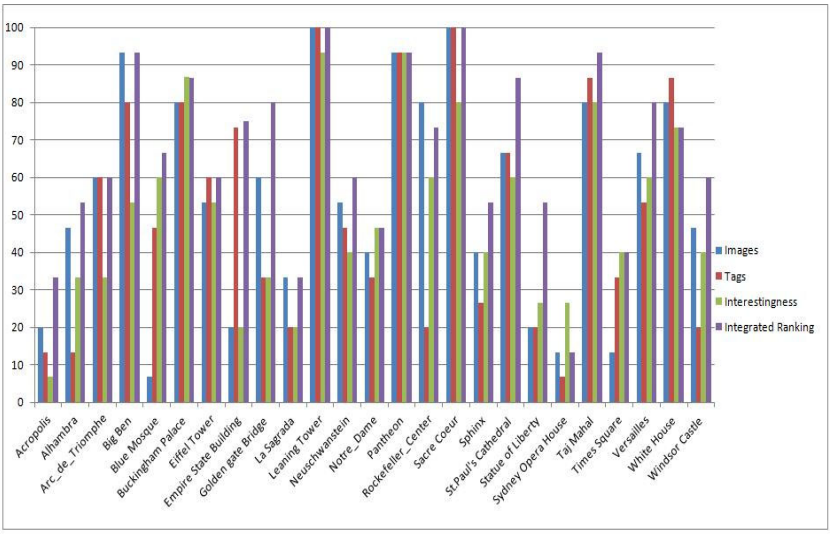

Fig. 5. Recognition accuracy for all landmarks, Y:Percentage, X:Landmarks

5 Evaluation of Proposed Approach

A recognition system was developed to validate the results. The SIFT features from each of the test images were extracted and compared with the k-d tree of each landmarks representative set of images. The landmark which had the highest number of match points with the test image was returned as the result. The system was trained using a variable number of images like the top 50,100, 200 images and it was found that a set of 75 images gave the best results. It was found that for certain landmarks the image analysis showed better results

compared to the social data analysis and vice-versa, hence a dynamic weighting system was adopted in our re ranking models where a particular review system was given a weight based on its observed recognition accuracy.If the recognition accuracy of first method of analysis was found to be A_1 and that of the second method was found to be A_2 then the weights were,

$$W_1 = \frac{A_1}{A_2 + A_1}; W_2 = \frac{A_2}{A_2 + A_1} \tag{6}$$

Six different models were used to integrate the ranks obtained from image and social data analysis as summarized below in the table.(Models 2 and 6 gave the best results respectively)

Models	Agreement Model	Fuzzy Model	Static Weights	Dynamic Weights	Images	Tags	Interestingness
Model 1	Yes	No	Yes	No	Yes	Yes	No
Model 2	No	Yes	Yes	No	Yes	Yes	No
Model 3	Yes	No	No	Yes	Yes	Yes	No
Model 4	No	Yes	No	Yes	Yes	Yes	No
Model 5	Yes	No	No	Yes	Yes	Yes	Yes
Model 6	No	Yes	No	Yes	Yes	Yes	Yes

For the experiments a set of 25 well known landmarks were selected.(listed in Figure 5). Afterwards a set of 75,000 images was downloaded from flickr.com using the landmark names as keywords.A set of diverse images not present in the training dataset was selected as the benchmark test set.

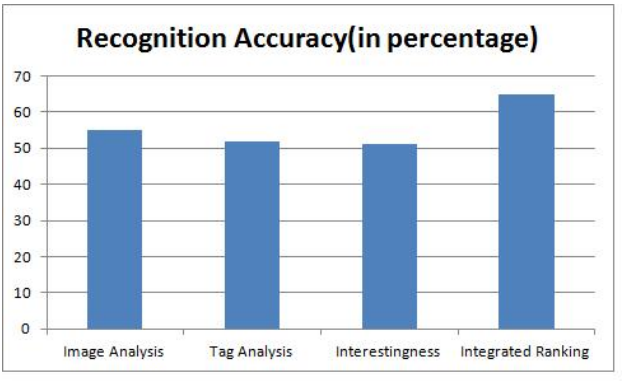

Fig. 6. Overall Recognition Accuracy,Y:Percentage,X:Lines of Analysis

It is evident from Figure 6,that the recognition accuracy increases after the process of integrated ranking.Some landmarks like Leaning Tower and Big Ben showed very high recognition accuracy because of their distinct visual features and popularity. The recognition accuracy was high as 95%. Some landmarks like Acropolis were a little more difficult to detect because of their spread out nature. Some landmarks have a similar skyline or periphery hence the system usually

confuses between them, like in this case, the system was usually confused between Statue of Liberty and Sydney Opera house, because of the accompanying sea. Hence, for better performance we might consider designing a recognition system which first asks for the users location then narrows down its search.

6 Conclusion

We have shown that landmark image mining and recognition can be improved by augmenting image processing with text mining techniques. A potential application of such system could be in various mobile devices where it could be used to recognize images taken by tourists. Such a system could also give tag recommendations to users. Future work,might investigate the origin and lifecycle of social data and incorprorate them into solving traditional multimedia problems.Other useful social information other than tags could be used to gain better insights into the mechanics of community contributed web sites.

Acknowledgments

We would like to express our gratitude to members of the DMR lab in the University of Minneosta, for their valuable inputs during both design and discussion phase.This work is supported by a grant from the Chinese National Natural Science Foundation under contract number 60973055,a grant from the CADAL project and ARL Network Science CTA via BBN TECH/W911NF-09-2-0053.

References

1. Wan, H.L., Chowdhury, M.U.: Image Semantic Classification by Using SVM. Journal of Software 14, 1891–1899 (2003)
2. Zhang, D., Wong, A., Indrawan, M., Lu, G.: Content-based Image Retrieval Using Gabor Texture Feature. In: Proceedings of First IEEE Pacific-Rim Conference on Multimedia (PCM 2000), Sydney, Australia, pp. 392–395 (2000)
3. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. Int'l J. Computer Vision 2(60), 91–110 (2004)
4. Kennedy, L.S., Naaman, M.: Generating diverse and representative image search results for landmarks. In: Proceeding of the 17th International Conference on World Wide Web, WWW 2008, pp. 297–306. ACM Press, New York (2008)
5. Beis, J., Lowe, D.G: Shape indexing using approximate nearest-neighbour search in high-dimensional spaces. In: Conference on Computer Vision and Pattern Recognition, pp. 1000–1006, Puerto Rico (1997)
6. Oztekin, B., Karypis, G., Kumar, V.: Expert Agreement and Content Based Reranking in a Meta-search Environment Using Mearf. In: Proceedings of the 11th International World Wide Web Conference, pp. 333–344, Honolulu, Hawaii, USA, May 7-11 (2002)
7. Wiguna, W.S., Fernández-Tébar, J.J., García-Serrano, A.: Using a Fuzzy Model for Combining Search Results from Different Information Sources to Build a Metasearch Engine. In: International Conference 9th Fuzzy Daysin Dortmund, Germany, pp. 325–334 (2006)

8. Zheng, Y., Zhao, M., Song, H., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T., Neven, H.: Tour the World: building a web-scale landmark recognition engine. In: Proceedings of the 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR (2009)
9. Kennedy, L., Naaman, M., Ahern, S., Nair, R., Rattenbury, T.: How flickr helps us make sense of the world: context and content in community-contributed media collections. In: Proceedings of the 15th International Conference on Multimedia, Augsburg, Germany, September 25-29, pp. 631–640 (2007)
10. Simon, I., Snavely, N., Seitz, S.M.: Scene summarization for online image collections. In: Proceedings of the 11th IEEE International Conference on Computer Vision. IEEE, Los Alamitos (2007)
11. Tsai, C., Qamra, A., Chang, E.Y., Wang, Y.: Extent: Inferring Image Metadata from Context and Content. In: IEEE International Conference on Multimedia and Expo., pp. 1270–1273 (2005)
12. Tian, Y., Srivastava, J., Huang, T., Contractor, N.: Social Multimedia Computing. In: Computer IEEE Computer Society Digital Library, June 30. IEEE Computer Society, Los Alamitos (2010)
13. Ramos, J.: Using TF-IDF to Determine Word Relevance in Document Queries. In: First International Conference on. Machine Learning (2003)
14. Hoffman, T.: Probabilistic Latent Semantic Indexing. In: Uncertainity in Artificial Intelligence, UAI 1999, Stockholm (1999)
15. Explore About Interestingness, `http://www.flickr.com/explore/interesting`
16. Yager, R.R.: Induced aggregation operators. Fuzzy Sets and Systems 137, 59–69 (2003)