WHAT CAN VISUAL CONTENT ANALYSIS DO FOR TEXT BASED IMAGE SEARCH?

Gang Hua

Microsoft Live Labs Research
One Microsoft Way, Redmond, WA 98052
ganghua@microsoft.com

ABSTRACT

Modern image search engines such as Google, Yahoo!, Microsoft Live image search are all text metaword based. To search for images, the users type in a text query and the search engines rank the result images almost sorely based on the text meta-words. The abundant visual information in the images themselves is largely neglected. Recently, we have observed several new features released in the aforementioned image search engines, especially Microsoft Live image search, which are clearly based on the analysis of the visual content. We summarize some of these features, give insights about how they are designed, and motivate new content analysis based features for text based image search engines.

***Index Terms*—** image search, visual content analysis

1. INTRODUCTION

The abundant research efforts in image search have been striving for content based image search [10, 17, 16, 7, 15, 5] due to its intellectual challenges. However, what fly commercially are still meta-word based, such as Google, Yahoo!, and Microsoft Live image search. Although there have been recent efforts on commercializing content based image search system, such as Idee TinEye[1] and Snaptell[2], these three major commercial image search engines still account for almost all the image search traffic on the internet, which amounts to hundreds of millions of queries daily.

In text based image search engine, the users type in a text query word, and the search engines automatically rank the images based on the text keywords associated with them, including the surrounding texts, the HTML alternative texts, or the titles of the host webs. In this process, the rich visual content has largely remained being unexplored. Hence the visual relevances of the result images are largely determined by how accurate the meta-words are in describing the visual content. This makes it vulnerable to spam such as web-stuffing attacks.

Adding rich meta-words to the images based on visual content analysis is natural to resolve this complication. There are several areas that content analysis techniques from computer vision and multimedia processing could help: (1) rich image search and browsing experiences [6, 11]; (2) security and privacy protection [9]; and (3) chronicle, product [14], location, and people search [1] in image search vertical. All of these three aspects are directly related to providing more relevant image search results.

However, there are fundamental challenges in applying content analysis in web scale. First, although most content analysis processes are running during crawling time, we still need to ensure that they process the indexed images in real-time to ensure the refreshing cycle to be agile enough. Hence image processing using MMX, SSE2, GPU, and multi-cores would greatly help. Besides, some advanced visual content based search or browsing experiences require to match visual features on the fly. This poses two constraints: first the visual features must be concise to save indexing storage as well as downloading time; and second, the matching algorithm should be real time so as to ensure smooth user experiences.

We summarize some advanced content based features in the various mainstream image search engines, especially those from Microsoft Live image search, who has been showing leading status in this space although their query share is not quite as great as Google image search. We give insights on how these features are designed and suggest new directions for content analysis based features.

Sec. 2 presents a overview of the architecture of text based image search engines. Sec. 3 is devoted to some advanced content analysis based features in them, followed by a prospect of future directions in Sec. 4. We conclude in Sec. 5.

2. TEXT-BASED IMAGE SEARCH ENGINE

Fig. 1 presents an overview of the architecture of modern text based image search engines. There are two processes: the *off-line* index generation process, and the *online* index serving process. Index generation is in charge of how to build an efficient and scalable (web-scale) indexing structure of all web images. While index serving is responsible for efficiently serving the gigantic amount of web image indices to the users based on their text query input.

In the index generation process, a *crawler* surfs the internet and builds the inverted file index for each identified image URL. The thumbnail image is also created and related

[1] http://www.tineye.com/
[2] http://www.snaptell.com/

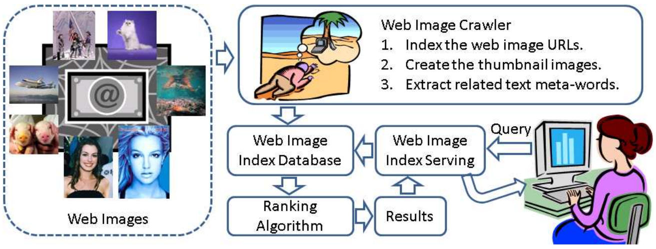

Fig. 1. Architecture of text based image search engines.

text meta-words are also extracted from the hosting web of the image URLs simultaneously. The crawling priority of the web images would be determined by the importance of the hosting webs determined, for example, by the PageRank algorithm [2, 13]. Both Google and Microsoft Live image search have built the indices of billions of web images.

Here inverted file system have to be leveraged in order to perform efficient retrieval in index serving time. Some of the text meta-words will be used to build the inverted file structure and the others will be used as ranking features. To store all the indices of the web image URLs, distributed file storage systems, such as the Google file system [12], are often adopted to handle the gigantic amount of data.

In index serving time, the search engine will run an index serving service and based on the input queries from the users. It performs forwarded look-up and gathers all related image URLs. Then a ranking algorithm, for example, the RankNet [4] or λ-Rank [3], would rank all the result image URLs and determine their relevances to the text queries the users typed in. The list of images retrieved, ordered by their relevances, are presented to the users. To be able to efficiently access the enormous amount of image indices, hierarchical job distribution and aggregation similar to that of MapReduce [8] is often leveraged.

Except the query click logs, the main features feeded into the ranking algorithms are those extracted from the text-meta words. This leaves large spaces for visual content analysis techniques to improve the image search relevances. In the next section, we will start reviewing some of the recent image search features in main-stream image search engines that are clearly powered up by visual content analysis.

3. IMAGE SEARCH BY CONTENT ANALYSIS

As discussed, visual content analysis could help to provide rich image search and browsing experiences. Indeed, this type of content analysis features is the most visible one when compared with the other two, i.e., security and privacy protection, and chronicle, product, location and people image search. Therefore our reviews will be focused on it.

The first feature we review is the *faceted image search* feature empowered by content analysis techniques such as face detection [18] and photo and graphics image recognition. Fig. 2 presents some screen shots to demonstrate such features from Microsoft Live image search. The left side of Fig. 2 shows the results page when you type in the text query “George Bush”. In the right pane of the web page, there is a list of faceted search features, among which there are two that are called *Style* and *Face*, respectively. The enlarged view is presented in the middle of Fig. 2. Under *Style*, if we click on *illustration*, we will see the result page on the right side of Fig. 2. As we can observe, the results include mostly cartoon images or edited images of president George Bush.

Fig. 3 demonstrated another faceted image search property of *faces*. The left side shows the result pages when we type in the query of “Paris”. As we can see, most result images are Effei Tower and Arc de Triomphe. After we click on the *Face* facet, the result images shown in the right side are almost all “Paris Hilton”. It is clear that this happened because the *Face* facet is filtering images based on if they contain faces or not. “Paris Hilton” is one of the most popular celebrity search queries in image search vertical. That explains why the result page on the right side of Fig. 3 contains exclusively Paris Hilton’s images. One can easily understand that these two features must have been enabled by running photo/graphics image recognition, as well as face detection to assign meta-words to the indexed images during crawling.

The next feature is more leaning toward content based image retrieval. On Microsoft Live image search, after type in any query, such as “Beijing Olympics” as showing in Fig. 4 and mouse-hover any result image, we will see a pop-out window, which refers to the region of the second result image. We show the enlarged view in the middle of Fig. 4. In the pop-out window, there is a URL called “show similar images”. Click on it, and it brings us to the result page in the right side of Fig. 4, where the result image we mouse hovered will be in the first place and the other following result images are visually quite similar to it. This feature uses the image specified as the query image to re-rank all the other result images in the database. It is clear that the matching of visual similarities happened on the fly so that the calculation must be efficient enough to ensure good user experiences. There are similar academic prototypes such as the re-ranking system of [6] and the CueFlick [11] system. Imprezzeo[3] also presented some demos like it on their web-site.

Some other features which are more hidden including porn filters. It is clear that all mainstream image search engines utilized some content analysis algorithms to perform porn filtering besides text keyword filtering. We will not further extend the discussion here in that space.

4. PROSPECT OF FUTURE DIRECTIONS

We remark on future research directions of applying computer vision or content analysis technologies for web-scale image

[3] http://www.imprezzeo.com

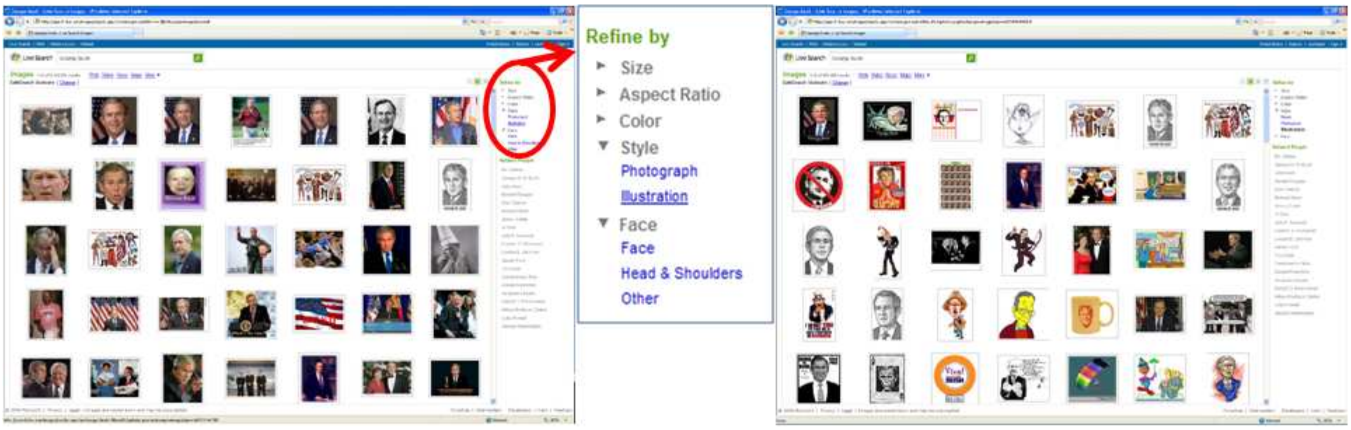

Fig. 2. Faceted image search based on photo and graphics image recognition.

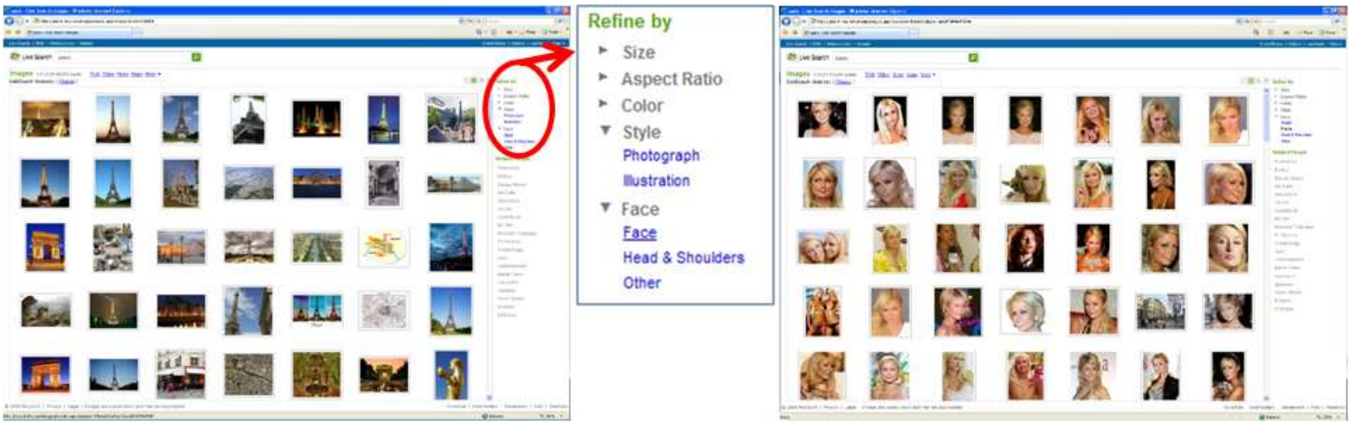

Fig. 3. Faceted image search based on face detection.

search. The discussion is a mixed one considering both real application scenarios, and the key research technologies we need to develop or challenges we need to address.

• The evolving computer vision research, especially object category recognition, scene recognition, as well as face recognition, would show great value in providing rich meta-words for improving image search relevance, or more facets for faceted search.

• Porn image is a fundamental offense in web image search. It is especially detrimental to kids. It has been proven that porn filtering purely based on text words is not enough. Hence content based porn filter has a huge role to play. Unfortunately, research in this topic is not that active in academia. The main reason is its notorious difficulty. The other reason is that it is not that a "decent" research topic.

• To improve the richness and completeness of the meta-words of one web image, it would be very beneficial to aggregate the meta-words of nearly duplicated images. State-of-the-art image matching technologies [15, 5] has indeed made this a very achievable task. The nearly duplicate images just function as the cross hyper-links between two web pages.

• People search is one of the major query categories in web image search. Human or face detection and recognition will greatly improve users' search and browsing experiences in this space.

• Many users use image search to browse beautiful images. They would especially appreciate high quality images. Research on non-reference visual quality assessment of web images would greatly help improve their browsing experiences.

• Query-by-image is still the end goal for large scale image search. The state-of-the-art technologies only enable us to perform a decent job on near-duplicate image detection [15, 5]. The semantic gap is still the fundamental, if not an unsolvable challenge, to this goal.

• To make the image features indexable, the features usually need to be vector quantized to integers [15, 5]. It is still an open problem what is the best way of carrying out the quantization.

• Once the image features become indexable, it should be quite straightforward to leverage state-of-the-art text information retrieval (IR) technologies for image retrieval.

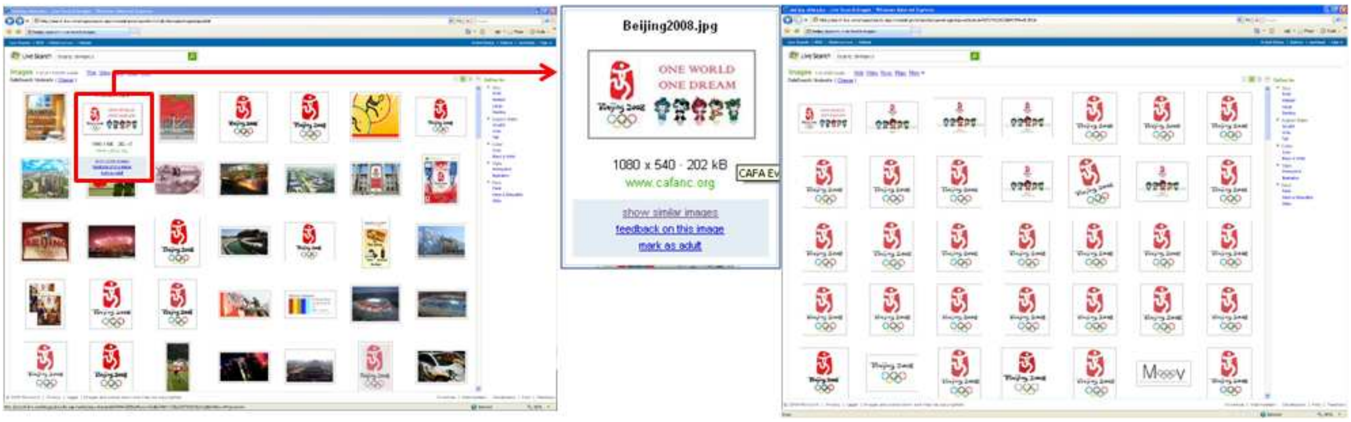

Fig. 4. Show similar images.

Further progress of research work in the aforementioned areas would greatly improve the quality of mainstream text based image search engines.

5. CONCLUSION

In this paper, we present an extensive review on what and how visual content analysis technologies can help the quality of modern text based image search engines. In particular, we review several advanced features on image search released by Microsoft Live image search and give insights on how they are designed. We conclude by identifying some of the key research topics which would further improve users' experiences in using text meta-words based image search engines.

6. REFERENCES

[1] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 848–854, Washington D.C., June 2004.

[2] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engin. In *Proc. of the Seventh International World Wide Web Conference*, Brisbane, Australia, April 1998.

[3] C. Burges, R. Ragno, and Q. V. Le. Learning to rank with non-smooth cost functions. In *Advances in Neural Information Processing Systems 19*, Vancouver, Canada, 2007.

[4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proc. of the 22nd International Conference on Machine Learning*, pages 89–96, New York, NY, USA, 2005. ACM.

[5] O. Chum, J. Philbin, and A. Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. In *Proc. of British Machine Vision Conference*, Leeds, UK, September 2008.

[6] J. Cui, F. Wen, and X. Tang. Real time google and live image search re-ranking. In *Proc. of ACM International Conference on Multimedia*, pages 729–732, Vancouver, BC, Canada, October 2008.

[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2):1–60, 2008.

[8] J. Dean and S. Ghemawat. Mapreduce: simplified data processing on large clusters. *Communication of ACM*, 51(1):107–113, 2008.

[9] M. M. Fleck and D. A. Forsyth. Finding naked people. In *Proc. of European Conf. on Computer Vision*, volume II, pages 592–602, Cambridge, UK, April 1996.

[10] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The qbic system. *Computer*, 28(9):23–32, 1995.

[11] J. Fogarty, D. S. Tan, A. Kapoor, and S. Winder. Cueflik: Interactive concept learning in image search. In *Proc. of CHI 2008 Conference on Human Factors in Computing Systems*, pages 29–38, Leeds, UK, September 2008.

[12] S. Ghemawat, H. Gobioff, and S.-T. Leung. The google file system. In *Proc. of 19th ACM Symposium on Operating Systems Principles*, Lake George, New York, October 2003.

[13] T. Haveliwala. Efficient computation of pagerank. Technical Report 1999-31, Stanford InfoLab, 1999.

[14] Y. Jing and S. Baluja. Pagerank for product image search. In *Proc. of 17th Internetional World Wide Web Conference*, Beijing, China, April 2008.

[15] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 2161–2168, New York City, New York, June 2006.

[16] Y. Rui and T. Huang. Optimizing learning in image retrieval. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 236–243, SC, USA, June 2000.

[17] Y. Rui and T. S. Huang. A novel relevance feedback technique in image retrieval. In *Proc. of the seventh ACM international conference on Multimedia (Part 2)*, pages 67–70, New York, NY, USA, 1999. ACM.

[18] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001.