Content Based Image Retrieval: A Survey

Nakul Agarwal[1]

***Abstract*— The explosive increase and ubiquitous accessibility of visual data on the web have led to the prosperity of research activity in image search or retrieval. With the ignorance of visual content as a ranking clue, methods with text search techniques for visual retrieval may suffer inconsistency between the text words and visual content. Content-based image retrieval (CBIR), which makes use of the representation of visual content to identify relevant images, has attracted a lot of attention in recent two decades. Such a problem is challenging due to the intention gap and the semantic gap problems. Numerous techniques have been developed for content-based image retrieval in the last decade, and the purpose of this paper is to briefly summarize and categorize those algorithms. I conclude with a few promising directions for future research.**

I. Introduction

With the universal popularity of digital devices embedded with cameras and the fast development of Internet technology, billions of people are projected to the Web sharing and browsing photos. The ubiquitous access to both digital photos and the Internet sheds bright light on many emerging applications based on image search. Image search aims to retrieve relevant visual documents to a textual or visual query efficiently from a large-scale visual corpus. Although image search has been extensively explored since the early 1990s [1], it still attracts lots of attention from the multimedia and computer vision communities in the past decade, thanks to the attention on scalability challenge and emergence of new techniques. Traditional image search engines usually index multimedia visual data based on the surrounding meta data information around images on the web, such as titles and tags.

However, text based image retrieval brings along a lot of problems with itself. First and foremost is the problem of image annotations. Image search engines have a huge ($\sim$millions) database of images and it's infeasible for each and every image to be manually annotated for retrieval. Even if one somehow manages to label all these images, it would probably be only in one uniform language, which is a limitation. Second is the problem of human perception, which applies to both the stages of image annotation and query formation. An image is likely to be perceived differently by different people. Fig. 1 shows an example of this subjectivity of human perception. The image in the figure can be thought of as an image of a *"lotus"*, *"flowers in a pond"* or *"Nelumbus Nucifera"*, which is the biological name of lotus. Now during retrieval, if the user does not input a text query that matches the perception of the annotator, he or she won't retrieve the desired result. Third is the problem of deeper (abstract) needs. Sometimes, it is hard to describe the images in terms of text. Textual information may be inconsistent with the visual content. In such cases, it is easier to tap into the visual features of these images for a description.

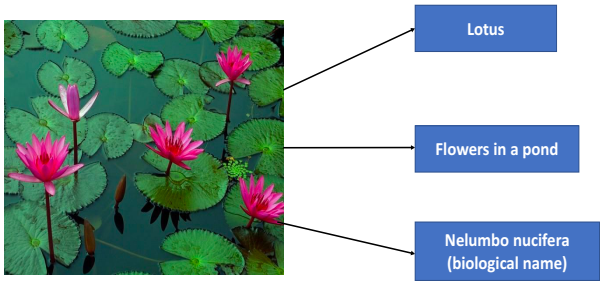

Fig. 1. Subjectivity of human perception. Different annotations of the same image for text-based image retrieval is shown above.

Because of the above reasons, content-based image retrieval (CBIR) is preferred and has been witnessed to make great advance in recent years.

In content-based visual retrieval, there are two fundamental challenges, i.e., intention gap and semantic gap. The intention gap refers to the difficulty that a user suffers to precisely express the expected visual content by a query at hand, such as an example image or a sketch map. The semantic gap originates from the difficulty in describing high-level semantic concept with low-level visual feature [2], [3], [4]. To narrow those gaps, extensive efforts have been made from both the academia and industry.From the early 1990s to the early 2000s, there have been extensive study on content-based image search. The progress in those years has been comprehensively discussed in existing survey papers [5], [6]. Around the early 2000s, the introduction of some new insights and methods triggers another research trend in CBIR. Specially, two pioneering works have paved the way to the significant advance in content-based visual retrieval on large-scale multimedia database. The first one is the introduction of invariant local visual feature SIFT [7]. SIFT is demonstrated with excellent descriptive and discriminative power to capture visual content in a variety of literature. It can well capture the invariance to rotation and scaling transformation and is robust to illumination change. The second work is the introduction of the Bag-of-Visual-Words (BoW) model [8]. Leveraged from information retrieval, the BoW model makes a compact

[1] University of California, Merced nagarwal2@ucmerced.edu

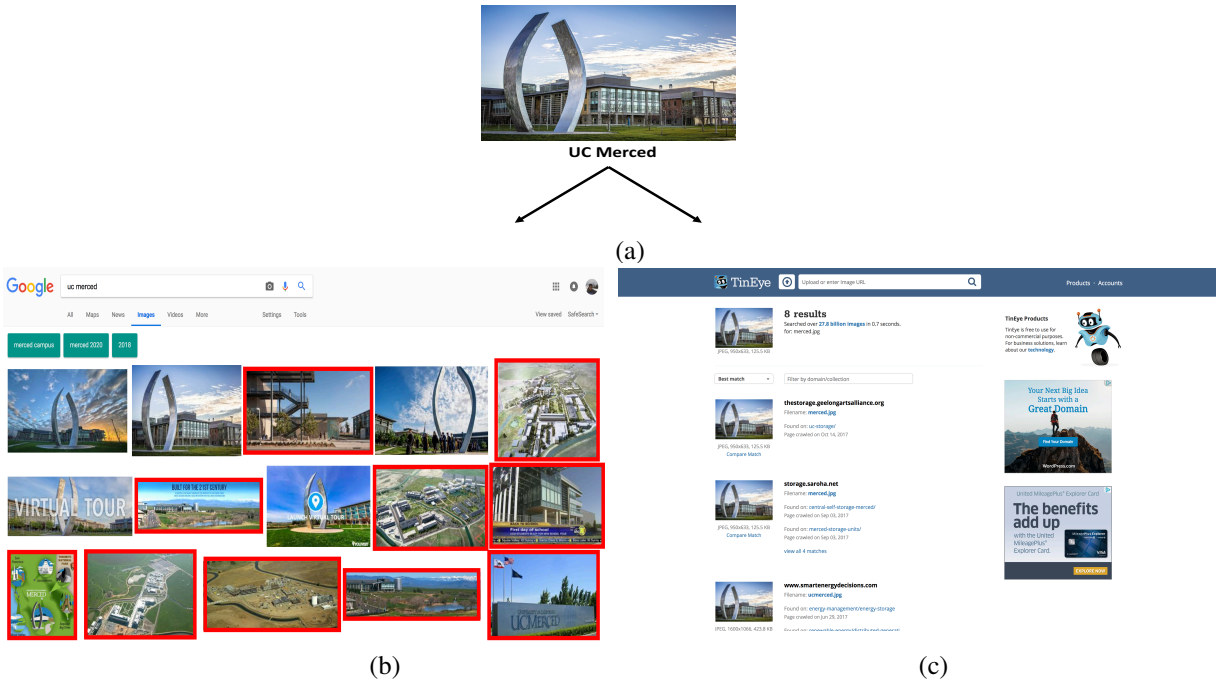

Fig. 2. Top image retrieval results for 'UC Merced' as input query (a) using different retrieval methods. Text-based image retrieval is shown in (b) using Google as search engine, with images outlined in red as false positives. In (c), content-based image retrieval is shown using Tineye[1] as search engine.

representation of images based on the quantization of the contained local features and is readily adapted to the classic inverted file indexing structure for scalable image retrieval.

Based on the above pioneering works, the last decade has witnessed the emergence of numerous work on multimedia content-based image retrieval [9], [10], [11], [12], [13]. Meanwhile, in industry, some commercial engines on content-based image search have been launched with different focuses, such as Tineye[1], Ditto[2], Snap Fashion[3], ViSenze[4], Cortica[5], etc. Tineye was launched as a billion-scale reverse image search engine in May, 2008. Until January of 2017, the indexed image database size in Tineye has reached up to 17 billion. To show a the difference between the results retrieved using text based and content based image retrieval, I use Google Images and Tineye respectively as shown in Fig. 2 (b,c). I use 'UC Merced' as the query in this case, which takes the form of text for google images and a representative image for Tineye. As you can glean from Fig. 2, Google Images returns quite a few false positives, whereas Tineye retrieves only the relevant results. False positives in this case refers to any image that doesn't contain the symbol of UC Merced. This clearly shows the advantage of content based image retrieval.

Different from Tineye, Ditto is specially focused on brand images in the wild. It provides an access to uncover the brands inside the shared photos on the public social media web sites. Technically speaking, there are three key issues in content-based image retrieval: image representation, image organization, and image similarity measurement. Existing algorithms can also be categorized based on their contributions to those three key items.

Image representation originates from the fact that the intrinsic problem in content-based visual retrieval is image comparison. For convenience of comparison, an image is transformed to some kind of feature space. The motivation is to achieve an implicit alignment so as to eliminate the impact of background and potential transformations or changes while keeping the intrinsic visual content distinguishable. In fact, how to represent an image is a fundamental problem in computer vision for image understanding. There is a saying that An image is worth a thousand words. However, it is nontrivial to identify those words. Usually, images are represented as one or multiple visual features. The representation is expected to be descriptive and discriminative so as to distinguish similar and dissimilar images. More importantly, it is also expected to be invariant to various transformations, such as translation, rotation, resizing, illumination change, etc.

In this paper, I focus on giving an overview on content based image retrieval. Recently, there have been some surveys related to CBIR [14], [2], [3]. In [14], Zhang et al. surveyed image search in the past 20 years from the perspective of database scaling from thousands to billions. In [3], Li et al. made a review of the state-of-the-art CBIR techniques in the context of social image tagging, with focus on three closed

[1]http://tineye.com/
[2]http://ditto.us.com/
[3]https://www.snapfashion.co.uk/
[4]https://www.visenze.com/
[5]http://www.cortica.com/

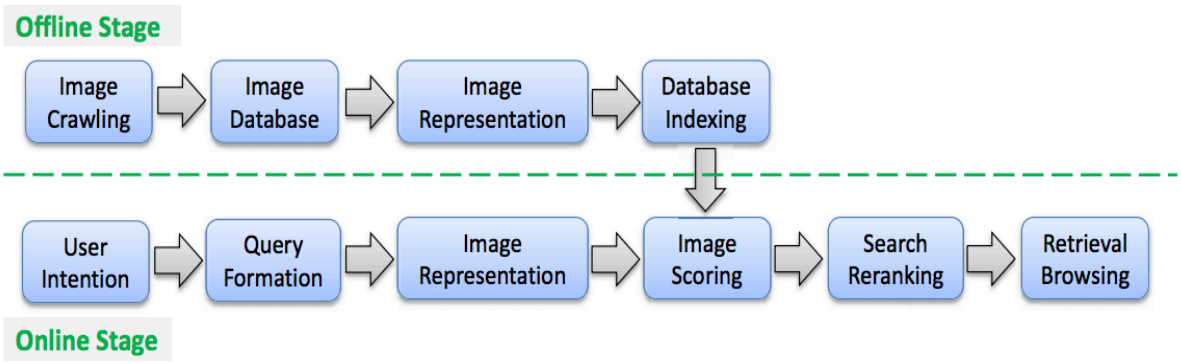

Fig. 3. The general framework of content-based image retrieval. The modules above and below the green dashed line are in the off-line stage and on-line stage, respectively. In this paper, I focus the discussion on four components, i.e., query formation, image representation, database indexing, and image scoring

linked problems, including image tag assignment, refinement, and tag-based image retrieval. Another recent related survey is referred in [2].

In the following sections, I first briefly review the generic pipeline of content-based image retrieval. Then, I discuss four key modules of the pipeline, respectively. Finally, I discuss future potential directions and conclude this survey.

II. General Pipeline

Content-based image search or retrieval has been a core problem in the multimedia field for over two decades. The general flowchart is illustrated in Fig. 3. Such a visual search framework consists of an off-line stage and an on-line stage. In the off-line stage, the database is built by image crawling and each database image is represented into some vectors and then indexed. In the on-line stage, several modules are involved, including user intention analysis, query formation, image representation, image scoring, search reranking, and retrieval browsing. The image representation module is shared in both the off-line and on-line stages. This paper will not cover image crawling, user intention analysis [15], and retrieval browsing [16], of which the survey can be referred in previous work [6], [17]. In the following, I will focus on the other four modules, i.e., query formation, image representation, database indexing, and image scoring.

In the following sections, I'll briefly describe the four modules.

III. Query formation

At the beginning of image retrieval, a user expresses his or her imaginary intention into some concrete visual query. The quality of the query has a significant impact on the retrieval results. A good and specific query may sufficiently reduce the retrieval difficulty and lead to satisfactory retrieval results. Generally, there are several kinds of query formation, such as query by example image, query by sketch map, query by color map, query by context map, etc. As illustrated in Fig. 4, different query schemes lead to significantly distinguishing results. In the following, I will briefly discuss each of those representative query formations.

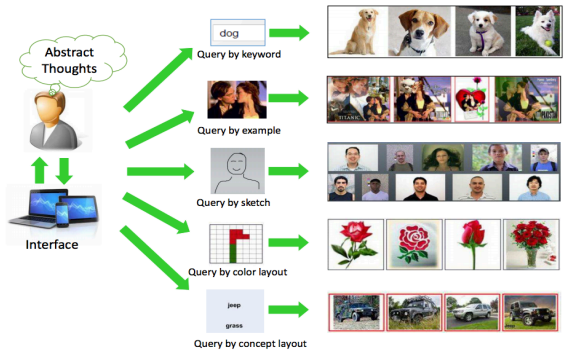

Fig. 4. Illustration of different query schemes with the corresponding retrieval results.

The most intuitive query formation is query by example image. That is, a user has an example image at hand and would like to retrieve more or better images about the same or similar semantics. For instance, a picture holder may want to check whether his picture is used in some web pages without his permission. Since the example images are objective without little human involvement, it is convenient to make quantitative analysis based on it so as to guide the design of the corresponding algorithms. Therefore, query by example is the most widely explored query formation style in the research on content-based image retrieval [8], [18].

Besides query by example, a user may also express his intention with a sketch map [19], [20]. In this way, the query is a contour image. Since sketch is more close to the semantic representation, it tends to help retrieve target results in users mind from the semantic perspective. Another query formation is color map. A user is allowed to specify the spatial distribution of colors in a given gridlike palette to generate a color map, which is used as query to retrieve images with similar colors in the relative regions of the image plain [21].

The above query formations are convenient for uses to

input but may still be difficult to express the users semantic intention. To alleviate this problem, Xu et al. proposed to form the query with concepts by text words in some specific layout in the image plain [22], [23]. Such structured object query is also explored in [24] with a latent ranking SVM model. This kind of query is specially suitable for searching generalized objects or scenes with context when the object recognition results are ready for the database images and the queries.

IV. Image Representation

In content based image retrieval, the key problem is how to efficiently measure the similarity between images. Since the visual objects or scenes may undergo various changes or transformations, it is infeasible to directly compare images at pixel level. Usually, visual features are extracted from images and subsequently transformed into a fix-sized vector for image representation. Considering the contradiction between large scale image database and the requirement for efficient query response, it is necessary to pack the visual features to facilitate the following indexing and image comparison. To achieve this goal, quantization with visual codebook training are used as a routine encoding processing for feature aggregation/pooling. Besides, as an important characteristic for visual data, spatial context is demonstrated vital to improve the distinctiveness of visual representation.
Based on the above discussion, I can mathematically formulate the content similarity between two images X and Y in Eq. 1.

$$S(\mathrm{X},\mathrm{Y}) = \sum_{x\in \mathrm{X}}\sum_{y\in \mathrm{Y}} k(x,y) \tag{1}$$

$$= \sum_{x\in \mathrm{X}}\sum_{y\in \mathrm{Y}} \phi(x)^T\phi(y) \tag{2}$$

$$= \psi(\mathrm{X})^T\psi(\mathrm{Y}) \tag{3}$$

Based on Eq. 1, there emerge three questions:
1) Firstly, how to describe the content image X by a set of visual features $\{x1, x2, ...\}$?
2) Secondly, how to transform feature sets X = $\{x1, x2, ...\}$ with various sizes to a fixed-length vector $\psi(\mathrm{X})$?
3) Thirdly, how to efficiently compute the similarity between the fixed-length vectors $\psi(\mathrm{X})^T\psi(\mathrm{Y})$?

The above three questions essentially correspond to the feature extraction, feature encoding aggregation, and database indexing, respectively. As for feature encoding and aggregation, it involves visual codebook learning, spatial context embedding, and quantization. The database indexing is left to the next section for discussion.

V. Database Indexing

Image index refers to a database organizing structure to assist for efficient retrieval of the target images. Since the response time is a key issue in retrieval, the significance of database indexing is becoming increasingly evident as the scale of image database on the Web explosively grows. Generally, in CBIR, one kind of indexing technique is popularly adopted, i.e., inverted file indexing. In the following, I will briefly discuss related retrieval algorithms in this category.

A. Inverted File Indexing

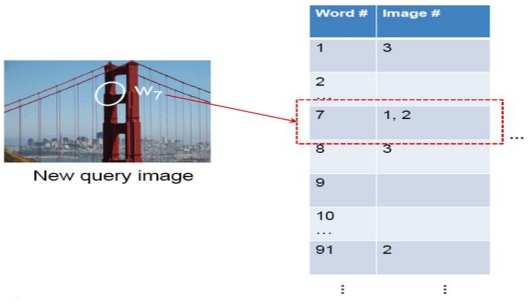

Fig. 5. A query image is efficiently matched to database images that share visual words using inverted file indexing structure.

Inspired by the success of text search engines, inverted file indexing [25] has been successfully used for large scale image search [8], [18]. In essence, In the inverted file structure, each visual word is followed by an inverted file list of entries. Each entry stores the ID of the image where the visual word appears, as shown in Fig. 5, along with some other clues for verification or similarity measurement. In on-line retrieval, only those images sharing common visual words with the query image need to be checked. Therefore, the number of candidate images to be compared is greatly reduced, achieving an efficient response.

VI. Image Scoring

In multimedia retrieval, the target results in the index image database are assigned with a relevance score for ranking and then returned to users. The relevance score can be defined either by measuring distance between the aggregated feature vectors of image representation or from the perspective of voting from relevant visual feature matches.

A. Distance Based Scoring

With feature aggregation, an image is represented into a fix-sized vector. The content relevance between images can be measured based on the Lp-normalized distance between their feature aggregation vectors, as shown in Eq. 4.

$$D(I_q, I_m) = (\sum_{i=1}^{N} |q_i - m_i|^p)^{\frac{1}{p}} \tag{4}$$

where the feature aggregation vectors of image I_q and I_m are denoted as $[q_1, q_2, ..., q_N]$ and $[m_1, m_2, ..., m_N]$, respectively, and N denotes the vector dimension. In [18], it is revealed that L1-norm yields better retrieval accuracy than L2-norm with the BoW model. Lin et al. extended the above feature distance to measure partial similarity between images with an optimization scheme [26].

When the BoW model is adopted for image representation, the feature aggregation vector is essentially a weighted visual

word histogram obtained based on the feature quantization results. To distinguish the significance of visual words in different images, term frequency (TF) and inverted document/image frequency (IDF) are widely applied in many existing state-of-the-art algorithms [18], [8].

B. Voting Based Scoring

In local feature based image retrieval, the image similarity is intrinsically determined by the feature matches between images. Therefore, it is natural to derive the image similarity score by aggregating votes from the matched features. In this way, the similarity score is not necessarily normalized, which is acceptable considering the nature of visual ranking in image retrieval.

In [13], the relevance score is simply defined by counting how many pairs of local feature are matches across two images. In [27], Jegou et al formulated the scoring function as a cumulation of squared TF-IDF weights on shared visual words, which is essentially a BOF (bag of features) inner product [27]. In [28], the image similarity is defined as the sum of the TF-IDF score [12], which is further enhanced with a weighting term by matching bundled feature sets. The weighting term consists of membership term and geometric term. The former term is defined as the number of shared visual words between two bundled features, while the latter is formulated using relative ordering to penalize geometric inconsistency of the matching between two bundled features. In [29], Zheng et al propose a novel Lp-norm IDF to extend the classic IDF weighting scheme.

VII. Future direction

Despite the extensive research efforts in the past decade, there is still sufficient space to further boost content based visual search. In the following, I will discuss several directions for future research, on which new advance shall be made in the next decade.

A. Deep Learning in CBIR

Despite the advance in content-based visual retrieval, there is still significant gap towards semantic-aware retrieval from visual content. This is essentially due to the fact that current image representation schemes are hand-crafted and insufficient to capture the semantics. The success of deep learning in large-scale visual recognition [30], [31], [32] has already demonstrated such potential.

To adapt those existing deep learning techniques to CBIR, there are several non-trivial issues that deserve research efforts. Firstly, the learned image representation with deep learning shall be flexible and robust to various common changes and transformations, such as rotation and scaling. Since the existing deep learning relies on the convolutional operation with anisotropic filters to convolve images, the resulted feature maps are sensitive to large translation, rotation, and scaling changes. It is still an open problem as whether that can solved by simply including more training samples with diverse transformations. Secondly, since computational efficiency and memory overhead are emphasized in particular in CBIR, it would be beneficial to consider those constraints in the structure design of deep learning networks. For instance, both compact binary semantic hashing codes and very sparse semantic vector representations are desired to represent images, since the latter are efficient in both distance computing and memory storing while the latter is well adapted to the inverted index structure.

B. Social Media Mining with CBIR

Different from the traditional unstructured Web media, the emerging social media in recent years have been characterized by community based personalized content creation, sharing, and interaction. There are many successful prominent platforms of social media, such as Facebook, Twitter, Wikipedia, LinkedIn, Pinterest, etc. The social media is enriched with tremendous information which dynamically reflects the social and cultural background and trend of the community. Besides, it also reveals the personal affection and behavior characteristics. As an important media of the user-created content, the visual data can be used as an entry point with the content-based image retrieval technique to uncover and understand the underlying community structure. It would be beneficial to understand the behavior of individual users and conduct recommendation of products and services to users. Moreover, it is feasible to analyze the sentiment of crowd for supervision and forewarning.

C. Cross-modal Retrieval

In the above discussion of this survey, I focus on the visual content for image retrieval. However, besides the visual features, there are other very useful clues, such as the textual information around images in Web pages, the click log of users when using the search engines, the speech information in videos, etc. Those multi-modal clues are complementary to each to collaboratively identify the visual content of images and videos. Therefore, it would be beneficial to explore cross-modal retrieval and fuse those multi-modal features with different models. With multimodal representation, there are still many open search topics in terms of collaborative quantization, indexing, search reranking, etc.

VIII. Conclusion

In this paper, I have investigated the advance on content based image retrieval in recent years. I focus on the four key modules of the general framework, i.e., query formation, image representation, image indexing and retrieval scoring. For each component, I have briefly discussed the key problems and categorized a variety of representative strategies and methods. Further, I have summarized some potential directions that may boost the advance of content based image retrieval in the near future.

REFERENCES

[1] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Transactions on circuits and systems for video technology*, vol. 8, no. 5, pp. 644–655, 1998.

[2] A. Alzubi, A. Amira, and N. Ramzan, "Semantic content-based image retrieval: A comprehensive study," *Journal of Visual Communication and Image Representation*, vol. 32, pp. 20–54, 2015.

[3] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. G. Snoek, and A. D. Bimbo, "Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval," *ACM Computing Surveys (CSUR)*, vol. 49, no. 1, p. 14, 2016.

[4] Z. Lin, G. Ding, M. Hu, and J. Wang, "Semantics-preserving hashing for cross-view retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3864–3872.

[5] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[6] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2, no. 1, pp. 1–19, 2006.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.

[9] S. Zhang, M. Yang, T. Cour, K. Yu, and D. N. Metaxas, "Query specific fusion for image retrieval," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 660–673.

[10] R. Arandjelović and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2911–2918.

[11] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang, "Mobile product search with bag of hash bits and boundary reranking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 3005–3012.

[12] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 809–816.

[13] W. Zhou, H. Li, Y. Lu, and Q. Tian, "Large scale image search with geometric coding," in *Proceedings of the 19th ACM international conference on Multimedia*. ACM, 2011, pp. 1349–1352.

[14] L. Zhang and Y. Rui, "Image searchfrom thousands to billions in 20 years," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 9, no. 1s, p. 36, 2013.

[15] X. Tang, K. Liu, J. Cui, F. Wen, and X. Wang, "Intentsearch: Capturing user intention for one-click internet image search," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 7, pp. 1342–1353, 2012.

[16] B. Moghaddam, Q. Tian, N. Lesh, C. Shen, and T. S. Huang, "Visualization and user-modeling for browsing personal photo libraries," *International Journal of Computer Vision*, vol. 56, no. 1-2, pp. 109–130, 2004.

[17] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys (Csur)*, vol. 40, no. 2, p. 5, 2008.

[18] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer vision and pattern recognition, 2006 IEEE computer society conference on*, vol. 2. Ieee, 2006, pp. 2161–2168.

[19] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Mindfinder: interactive sketch-based image search on millions of images," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1605–1608.

[20] C. Xiao, C. Wang, L. Zhang, and L. Zhang, "Sketch-based image retrieval via shape words," in *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 2015, pp. 571–574.

[21] J. Wang and X.-S. Hua, "Interactive image search by color map," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 1, p. 12, 2011.

[22] H. Xu, J. Wang, X.-S. Hua, and S. Li, "Image search by concept map," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 275–282.

[23] ——, "Interactive image search by 2d semantic map," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1321–1324.

[24] T. Lan, W. Yang, Y. Wang, and G. Mori, "Image retrieval with structured object queries using latent ranking svm," in *European conference on computer vision*. Springer, 2012, pp. 129–142.

[25] R. Baeza-Yates, B. Ribeiro-Neto, *et al.*, *Modern information retrieval*. ACM press New York, 1999, vol. 463.

[26] Z. Lin and J. Brandt, "A local bag-of-features model for large-scale object retrieval," in *European conference on Computer vision*. Springer, 2010, pp. 294–308.

[27] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International journal of computer vision*, vol. 87, no. 3, pp. 316–336, 2010.

[28] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 25–32.

[29] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Lp-norm idf for large scale image search," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 1626–1633.

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, *et al.*, "Going deeper with convolutions." Cvpr, 2015.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.