

Visual Vocabulary Learning and Its Application to 3D and Mobile Visual Search

Liujuan Cao, *Student Member, IEEE*

Abstract

In this technical report, we review related works and recent trends in visual vocabulary based web image search, object recognition, mobile visual search, and 3D object retrieval. Especial focuses would be also given for the recent trends in supervised/unsupervised vocabulary optimization, compact descriptor for visual search, as well as in multi-view based 3D object representation.

Index Terms

Visual search, object recognition, visual vocabulary, supervised learning, compact descriptor, 3D object retrieval.

I. INTRODUCTION

Recently, local feature representations [1][2][3] are very popular in computer vision research, with extensive applications in near duplicate visual search, mobile visual search, video copy detection, image annotation, and web-scale image retrieval. Generally speaking, the state-of-the-art visual search systems are built on the so-called visual vocabulary model, that is, (hierarchical) quantization of local feature spaces with inverted indexing speed up [4][5][6][7][8][9][10]. In this scenario, the local features [1][2] extracted from reference images are quantized into a set of visual words, each with an indexing file. Each reference image is then represented as a Bag-of-Words histogram and is inversely indexed into words that contain local features extracted from this image. This Bag-of-Words representation offers sufficient robustness against photographing variances in occlusions, viewpoints, illuminations, scales and backgrounds. Subsequently, the image search problem is reformulated from a document retrieval

Liujuan Cao is with the School of Computer Science, Harbin Engineering University, 150001, P. R. China, e-mail: caoliujuan@hrbeu.edu.cn.

perspective, from which perspective many successful techniques such as TF-IDF [11], pLSA [12] and LDA [13] can be directly used.

In general, both visual vocabulary and hashing based search techniques can be categorized into the approximate visual search techniques, which has been well exploited in the recent literature to handle the search deficiency in large image collections, *e.g.* Vocabulary Tree [5], Approximate K-means [6], Hamming Embedding [14], Locality Sensitive Hashing [15] and their variances [16][17][6][18][8][19]. In a typical setting, the visual vocabulary based search system works in a client-server paradigm: The client end (*e.g.* a mobile device or a web interface) sends a query image to the server. Or alternatively, in some recent mobile visual search systems [20][21][22][23][24][25], sending compact visual descriptors extracted from this query image can further reduce the wireless communication latency.

The server end searches similar reference images in its reference data set following three consecutive phases as:

- Extracting local features from the query image (if the server delivers compact visual descriptors instead of the query image, this step is skipped);
- Quantizing these local features into a Bag-of-Words histogram using the vocabulary;
- Ranking similar images in the inverted indexing files of all non-empty words, so as to avoid the linear scanning of all reference images in the similarity ranking.

In this technical report, we review related works and recent trends in visual vocabulary based web image search, object recognition, mobile visual search, and 3D object retrieval. Especial focuses would be also given for the recent trends in supervised/unsupervised vocabulary optimization, compact descriptor for visual search, as well as in multi-view based 3D object representation.

II. VISUAL VOCABULARY CONSTRUCTION

In this section, we first review the recent advances in visual vocabulary construction. Typically speaking, building visual vocabulary usually resorts to unsupervised vector quantization, which subdivides the local feature space into discrete regions each corresponds to a visual word. An image is represented as a Bag-of-Words (BoW) histogram, where each word bin counts how many local features of this image fall in the corresponding feature space partition of this word. To this end, many vector quantization schemes are proposed to build visual vocabulary, such as K-means [4], Hierarchical K-means (Vocabulary Tree) [5], Approximate K-means [6], and their variances [16][17][6][18][26][8][27][28]. Meanwhile, hashing local features into a discrete set of bins and indexed subsequently is an alternative choice, for which methods like Locality Sensitive Hashing (LSH) [29], Kernalized LSH [15], Spectral Hashing [30] and

its variances [31][19] are also exploited in the literature. The visual word uncertainty and ambiguity are also investigated in [14][6][32][33], using methods such as Hamming Embedding [14], Soft Assignments [6] and Kernelized Codebook [33]. Some other related directions include optimizing the initial inputs of visual vocabulary construction, such as learning a better local descriptor detector as in [34], coming up with a better similarity metric, such as learning an optimal hashing based distance matching as in [35] for human action search [36], incorporating Bayesian reasoning into the similarity calculation [37], as well as distribute the visual vocabulary model and its inverted indexing structure into multiple machines [38].

Stepping forward from unsupervised vector quantization, semantic or category labels are also exploited [39][40][41][9][42] to supervise the vocabulary construction, which learns the vocabulary to be more suitable for the subsequent classifier training, *e.g.*, the images in the same category are more likely to produce similar BoW histograms and vice versa. In terms of functionality, works in [40][41][39] exploits the learning-based codebook constructions. For instance, Mairal et al. [40] built a supervised vocabulary by learning discriminative and sparse coding models for object categorization. Lazebnik et al. [41] proposed to construct supervised codebooks by minimizing mutual information lost to index fully labeled data. Moosmann et al. [39] proposed an ERC-Forest to consider semantic labels as stopping tests in building supervised indexing trees. Another group of related works [43][44][45] refines (merges or splits) the initial codewords to build class(image)-specific vocabularies for categorization. Although working well for limited-number categories, these approaches cannot be scaled up to generalized scenarios with numerous and semantically correlative categories. Similar works can also be referred to the Learning Vector Quantization [46][47][48] in data compression, which adopted self-organizing maps [47] or regression lost minimization [48] to build codebook that minimizes training data distortions after compression. From the supervised learning point of view, works in topic decompositions (pLSA [49] or LDA [50][27]) can be also treated as supervised codebook refinement, which typically resorts to learning a topical-level representation instead of the word-level representation. It is worth to note that, by exploiting semantics learning in visual representation stage, this kind of works differs from works that adopt semantic learning to refine the subsequent recognition stages [51][52][53], *e.g.* classifiers based on machine translation [51] or semantic hierarchy [52]. On the other hand, optimizing the visual vocabulary can be also benefit the related tasks such as image annotation, relevance feedback learning, video location search and text detection, as shown in [54][55][56][57][58][59][60].

III. LANDMARK SEARCH AND RECOGNITION

A widely exploit scenario in visual vocabulary model comes from landmark search and recognition, where the near-duplicate changes in terms of different viewing angles, occlusions, appearance changes, as well as perspective changes fits the merits of visual vocabulary model well, with lots of applications such as mobile localization and advertisement recommendation [54]. Recently, scalable near-duplicate image retrieval [5,6,8,22,25,26] has been largely addressed by promising visual vocabulary models as well as inverted indexing techniques. The representative approaches for vocabulary construction include K Means clustering [4], Vocabulary Tree [5], and Approximate K-Means [61] et al. We organize this topics from the following two perspectives based upon the problem scale, named city-scale landmark search and worldwide landmark search:

A. City-Scale Landmark Search

Towards city-scale landmark search and recognition, Schindler et al. [7] presented a location recognition system through geo-tagged video streams with multiple path search in the vocabulary tree. Eade et al. [62] also adopted a vocabulary tree for real-time loop closing based on SIFT-like descriptors. Our previous works in [8] proposed a density-based metric learning to optimize the hierarchical structure of vocabulary tree [5] for street view location recognition. Yeh et al. [63] further adopted a hybrid color histogram to compensate the feature based ranking in mobile based location recognition applications. Cristani et al. [64] learnt a global-to-local image matching for location recognition. And their consecutive work in [65] identified landmark buildings based on image data, metadata, and other photos taken within a consecutive 15-minute window. In addition, Irschara et al. [66] further leverage structure-from-motion (SFM) to build 3D scene models for street views, combined with vocabulary tree for simultaneously scene modeling and location recognition. Xiao et al. [67] proposed to combine bag-of-features with simultaneous localization and mapping (SLAM) to further improve the recognition precision. The quantization issues in visual vocabulary are recently also well addressed to fit the city-scale landmark search scenario, such as the works in [8][10]. Incrementally vocabulary indexing is also explored in [26] to maintain a landmark search system in a time varying database.

B. Worldwide Landmark Search

Towards worldwide landmark search and recognition, the IM2GPS system [68] inferred possible location distributions of a given query by visual matching in a worldwide, geo-tagged landmark dataset. As a consecutive work, Kalogerakis et al. further [69] demonstrated how to combine single image matching

with sequential data to improve matching accuracy. Zheng et al. [70] developed a worldwide landmark recognition system, which used a predefined landmark list to query online image search engines to selected candidate images, followed by re-clustering and pruning to locate the final landmark location. Recent works also proposed to mine representative landmarks at a worldwide scale, such as the sparse representation based landmark mining approach in [71][72].

IV. 3D OBJECT RETRIEVAL AND RECOGNITION

Recently, extensive research efforts [?,73]–[75,84]–[86] have been dedicated to 3D object retrieval and recognition, and how to effectively and efficiently search 3D objects is an important topic in multimedia research. Early methods [77,87]–[90] mainly focus on model-based method, while the view-based methods [78]–[80,82,83,91]–[96] have attracted much attention nowadays. This is due to the fact that view-based methods [81,97,98] are with the highly discriminative property for object representation and visual analysis [99]–[105] also plays an important role in multimedia applications. For view-based 3D object retrieval, the visual words have been investigated for 3D object representation.

In these view-based 3D object retrieval methods, generally the SIFT features are extracted from all selected views, and a visual word dictionary is learnt. Then a bag-of-visual-words description is generated for 3D object representation. The matching between two 3D objects is conducted based on this representation.

The major advantages by using visual words description in 3D object retrieval and recognition are two-fold.

- 1) The visual words description is effective on image representation, which can be discriminative for the description of different classes of objects.
- 2) The visual words can be extracted easily, and it is robust to object scaling and rotating.

Furuya and Ohbuchi [106] first proposed to extract SIFT features for views of 3D objects. In this method, each 3D object is rendered into a group of depth images, and the SIFT features are extracted from these images. This method uses the bag-of-features approach to integrate the local features into a feature vector for each model. Then the matching of these two feature vectors determines the distance between the two 3D objects. Ohbuchi et al. [107] further proposed to employ Kullback-Leibler divergence (KLD) to calculate the distance between two bag-of-visual-feature based 3D objects. Osada et al. [108] employed the bag-of-visual-feature method to SHREC'08 CAD model track task. Ohbuchi and Shimizu [109] employed the semi-supervised manifold learning method for object class recognition. The proposed method projects the original feature space onto a lower dimensional manifold. Then the relevance feedback information

is employed to capture the semantic class information by using the manifold ranking algorithm. Though the bag-of-visual-feature description is effective on 3D object retrieval, the computational cost is high. Ohbuchi and Furuya [110] further accelerated the method by using a Graphics Processing Unit. A bag-of-region-words method [76] is introduced to extract visual features in the region level. This method first gridly selects points in each image and the local SIFT features are extracted for these points. Then each feature is encoded into a visual word with a pre-trained visual vocabulary. In this step, each view is split into a set of regions, and each region is represented by a bag-of-visual-words feature vector. All the obtained regions are further grouped into clusters based on the bag-of-visual-words feature, and one feature is chosen as the representative one from each cluster with corresponding weight. The Earth Movers Distance is used to measure the distance between two 3D objects.

Furuya and Ohbuchi [111] proposed to employ dense sampling to extract feature points in the views of 3D objects, and then the SIFT feature is extracted from each point. These visual features are further clustered into groups to generate the visual vocabulary. Then the feature histogram is generated to calculate the 3D object distance. This method has been further extended [111,112] to deal with large scale data. A distance metric learning method [113] is proposed to learn the distance metric for matching of 3D models. Endoh et al. [114] introduced to conduct learning on the manifold structure of 3D models by using clustering-based training sample reduction. Kawamura et al. [115] further employed the geometrical feature to improve the feature-based method.

V. MOBILE VISUAL SEARCH

A. *The Need of Compact Descriptor*

There are many evidences support the usage of compact descriptors for mobile visual search and augmented reality applications:

- Firstly, it remains a long way to provide a stable and high-speed (3G) wireless coverage everywhere, especially for those touristic landmarks that are often far away from urban areas or for developing countries, e.g., Lhasa, Tibet in our experiments. So it is unrealistic to guarantee the bandwidth is good enough to reliably and fast send a query photo. In particular, the recently established MPEG Ad Hoc Group **CDVS** is bring together the academia and industry practitioners to explore the next MPEG standard of **Compact Descriptor for Visual Search**.
- Secondly, from the server perspective, the network capability of receiving a batch of entire photo queries is by no doubt limited for a more powerful cloud platform that may handle intensive search at the server end. From the industry practice, a clear fact is that receiving multiple query photos is

much more challenging than receiving texts in the state-of-the-art search engines. More importantly, with compact upstream queries, more bandwidth can be saved up to downstream return the actually valuable searched information (in rich forms of text, images and video). That is one of the reasons why many internet service providers often set a smaller uplink bandwidth to save bandwidth for fast browsing.

- Finally, sending large amount of data via 3G wireless definitely causes serious battery energy consumption. Empirical evidence shows that compressing the query photo into a compact signature and sending the signature through the mobile is much more power saving.

In summary, the promising research efforts in compact visual descriptors are bringing great benefits in lightening the battery consumption, the cost of bandwidth and memory, which undoubtedly contribute to efficient and effective visual query delivery in mobile visual search, especially in the scenarios of video rate reality augmentation.

B. The State of The Arts

The ever growing computational power motivates the research efforts to extract visual descriptors directly on a mobile device [20,21,23,116]–[121]. Instead of sending an entire photo, sending such descriptors are compact enough to enable the low bit rate search. Comparing with the previous works in low dimensional local descriptors such as PCA-SIFT [122], GLOH [2], SURF [123], and MSR descriptors [124], works in [20,21,116]–[118] target at intensive compactness as well as efficient extraction in a standard mobile end. They are expected to work well in mobile visual search scenarios.

Coming with the ever growing computational power in the mobile devices, recent works have proposed to directly extract compact visual descriptors on the mobile devices [20,21,116]–[118]. Instead of sending the entire query, such descriptors are transmitted to enable a low bit rate search. Comparing with the previous works in low dimensional local descriptors such as PCA-SIFT [122], GLOH [2], SURF [123], and MSR descriptors [124], works in [20,21,116]–[118] target at very extreme compression rates as well as efficient online extraction in the mobile end. Consequently, recent works in [20,21,116]–[118] have focused on more compact descriptors specialized for the mobile visual search:

Towards compact local visual descriptors, Chandrasekhar et al. proposed a Compressed Histogram of Gradient (CHoG) [21], which are further compressed by both Huffman Tree and Gagic Tree to reduce the size of each descriptor to approximate 50 bits. Works in [117] employ Karhunen-Loeve transform to compress the SIFT descriptor, producing approximate 2 bits per SIFT dimension (128 dimensions in total). Tsai et al. [125] proposed to transmit the spatial layouts of interest points to improve the precision

of feature matching. Comparing with sending an entire query photo, sending above compact descriptors are much more efficient [126]. For instance, CHOg typically outputs only 50 bits per local feature. When 1,000 interest points are extracted per query (following the popular detector setting [3]), the data amount to transmit is only 8KB, much less than the entire query photo (typically over 20KB with JPEG compression).

Chen et al. [20] stepped forward to send the bag-of-features histogram [20,116] instead, which encodes the position difference of non-zero bins to yield approximate 2KB per query photo using a one million vocabulary. It largely outperforms directly sending the compact local descriptors (more than 5KB in reported works). Their successive work in [116] further compressed the inverted indexing structure of visual vocabulary [5] with arithmetic coding to reduce the memory and storage cost to maintain the visual search system in server(s). A recent group of representative works come from the endeavors of Ji et al. in compression the visual vocabulary based descriptor representation directly on the mobile end [22][27][23][24][25][121].

Beyond the context of mobile visual search, compact image signatures are recently investigated in [128][18][129][42]. For instance, Jegou et al. proposed a product quantization scheme [129] to learn a compact image descriptor that approximates the square distance of original Bag-of-Words histograms. The same authors also proposed a miniBOF feature [130] by packing the bag-of-features. Their recent work in [18] further aggregated local descriptors with PCA and locality sensitive hashing to produce a compact descriptor of approximate 32 bits in length.. Weiss et al. [128] used spectral hashing to compress GIST descriptor [131] into tens of bits. Wang et al. [132] proposed a locality-constrained linear coding (LLC) scheme over the Bag-of-Words histogram to improve the spatial pyramid matching. In multi-view coding, Yeo et al. [127] proposed a rate-efficient correspondence learning scheme to randomly project descriptors to build a minHashing code.

REFERENCES

- [1] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2): 91-110, 2004. 1
- [2] K. Mikolajczyk, B. Leibe, and B. Schiele. Local features for object class recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2005. 1, 7
- [3] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1/2): 43-72, 2006. 1, 8
- [4] J. Sivic and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. *International Conference of Computer Vision*, 2003. 1, 2, 4

- [5] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *IEEE Conference on Computer Vision and Pattern Recognition*, 2006. 1, 2, 4, 8
- [6] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabulary and fast spatial matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 1, 2, 3, 4
- [7] G. Schindler and M. Brown. City-scale location recognition. *IEEE Conference on Computer Vision and Pattern Recognition*, 2007. 1, 4
- [8] Ji R., Xie X., Yao H., and Ma W.-Y. (2009). Mining city landmarks from blogs by graph modeling. *ACM Multimedia* (pp. 105-114). 1, 2, 4
- [9] R. Ji, H. Yao, X. Sun, B. Zhong, and W. Gao. Towards Semantic Embedding In Visual Vocabulary. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2010. 1, 3
- [10] R. Ji, H. Yao, X. Xie, and Q. Tian. Vocabulary Hierarchy Optimization and Transfer for Scalable Image Search. *IEEE Multimedia Magazine*, 18(3), 66-77, 2011. 1, 4
- [11] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 1988. 2
- [12] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning Journal*, 2001. 2
- [13] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation, *Journal of Machine Learning Research*, 3(4/5): 993-1022, 2003. 2
- [14] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. *European Conference of Computer Vision*, 2008. 2, 3
- [15] B. Kulis and K. Grauman. Kernelized locality sensitive hashing for scalable image search. *International Conference of Computer Vision*, 2009. 2
- [16] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. *International Conference of Computer Vision*, 2005. 2
- [17] J. Yang, Y.-G. Jiang, A. Hauptmann, and C.-W. Ngo. Evaluating bag-of-words representations in scene classification. *ACM Conference on Multimedia Information Retrieval*, 2007. 2
- [18] H. Jegou, M. Douze, C. Schmid, and P. Perez. Aggregating local descriptors into a compact image representation. *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 2, 8
- [19] D. Liu, S. Yan, R. Ji, X.-S. Hua, and H.-J. Zhang. Image Retrieval with Query-Adaptive Hashing. *ACM Transactions on Multimedia Computing, Communications and Applications*, In Press. 2, 3
- [20] D. Chen, S. Tsai, and V. Chandrasekhar. Tree histogram coding for mobile image matching. *Data Compression Conference*, 2009. 2, 7, 8
- [21] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod. CHoG: Compressed histogram of gradients a low bit-rate feature descriptor. *IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 2, 7
- [22] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, and W. Gao. Location Discriminative Vocabulary Coding for Mobile Landmark Search. *International Journal of Computer Vision*, In Press, 2012. 2, 4, 8
- [23] R. Ji, L.-Y. Duan, J. Chen, H. Yao, and W. Gao. Learning Compact Visual Descriptor for Low Bit Rate Mobile Landmark Search. *International Joint Conference of Artificial Intelligence*, 2456-2463, 2011. 2, 7, 8
- [24] R. Ji, L.-Y. Duan, J. Chen, H. Yao, Y. Rui, S.-F. Chang, and W. Gao. Towards Low Bit Rate Mobile Visual Search with Multiple Channel Coding. *ACM Multimedia*, 2011. 2, 8

- [25] R. Ji, L.-Y. Duan, J. Chen, H. Yao, and W. Gao. When Codeword Frequency Meets Geographical Location. *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2400-2403, 2011. 2, 4, 8
- [26] R. Ji, X. Xie, H. Yao, and W.-Y. Ma. Vocabulary Tree Incremental Indexing for Scalable Scene Recognition. *IEEE International Conference on Multimedia and Expo*, 2008. 2, 4
- [27] R. Ji, J. Chen, L.-Y. Duan, and W. Gao. Towards Compact Topical Descriptor. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012. 2, 3, 8
- [28] L. Cao, R. Ji, Y. Gao, Y. Yang, and Q. Tian. Weakly Supervised Coding with Geometric Consistency Pooling. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012. 2
- [29] B. Kulis and K. Grauman. Kernelized locality-sensitive hashing for scalable image search. *International Conference of Computer Vision*, 1999. 2
- [30] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. *Advances in Neural Information Processing Systems*, 2008. 2
- [31] W. Liu, J. Wang, R. Ji, Y.-G. Jiang, and S.-F. Chang. Supervised Hashing with Kernels. *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012. 3
- [32] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization. *ACM Conference of Content based Image and Video Retrieval*, 2007. 3
- [33] J. Gemert, C. Veenman, A. Smeulders, and J. Geusebroek. Visual word ambiguity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(7): 1271-1283, 2009. 3
- [34] R. Ji, H. Yao, Q. Tian, P. Xu, X. Sun, and X. Liu. Context-Aware Semi-Local Feature Detector. *ACM Transactions on Intelligent Information Systems*, 3(3):44-71, 2012. 3
- [35] R. Ji, H. Yao, and X. Sun. Actor-Independence Action Retrieval Using Spatiotemporal Vocabulary with DTW Matching. *Pattern Recognition*, 44: 624-638, 2011. 3
- [36] R. Ji, X. Sun, H. Yao, P. Xu, and X. Liu. Attention-Driven Action Retrieval using 3D-SIFT with Dynamic Time Warping. *ACM Multimedia*, 2008. 3
- [37] R. Ji, X. Lang, H. Yao, and Z. Zhang. Semantic Supervised Region Retrieval Using Keyword Integrated Bayesian Reasoning. *International Journal on Innovative Computing, Information and Control*, 2008. 3
- [38] R. Ji, L.-Y. Duan, H. Yao, L. Xie, Y. Rui and W. Gao. Learning to Distribute Vocabulary Indexing for Scalable Visual Search. *IEEE Transactions on Multimedia*, In Press, 2012. 3
- [39] F. Moosmann, B. Triggs, and F. Jurie. Fast discriminative visual codebooks using randomized clustering forests. *Advances in Neural Information Processing Systems*, 2006. 3
- [40] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Supervised Dictionary Learning. *Advances in Neural Information Processing Systems*, 2008. 3
- [41] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2009. 3
- [42] R. Ji, H. Yao, W. Liu, X. Sun, and Q. Tian. Task Dependent Visual Codebook Compression. *IEEE Transactions on Image Processing*, In Press, 2012. 3, 8
- [43] F. Perronnin, C. Dance, G. Csurka, and M. Bressan. Adapted vocabularies for generic visual categorization. *ECCV*, 2006. 3
- [44] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive review. *IJCV*, 2007. 3

- [45] Jingen Liu, Yang Yang, and Mubarak Shah. Learning Semantic Visual Vocabularies Using Diffusion Distance. *CVPR*, 2009. 3
- [46] T. Kohonen. Learning vector quantization for pattern recognition. *Tech. Rep. TKK-F-A601, Helsinki Institute of Technology*, 1986. 3
- [47] T. Kohonen. Self-Organizing Maps, 3rd edition, *Springer-Verlag*, 2000. 3
- [48] A. Rao, D. Miller, K. Rose, and A. Gersho. A generalized VQ method for combined compression and estimation. *ICASSP*, 1996. 3
- [49] F.-F. Li and P. Pietro. A Bayesian hierarchical model for learning natural scene categories. *ICCV*, 2007. 3
- [50] A. Bosch, A. Zisserman, and X. Munoz. Scene classification using a hybrid generative/discriminative approach. *PAMI*, 2008. 3
- [51] P. Duygulu, K. Barnard, JFG de Freitas, and DA Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *ECCV*, 2002. 3
- [52] M. Marszalek and C. Schmid. Semantic Hierarchies for Visual Object Recognition. *CVPR*, 2007. 3
- [53] C. Liu, J. Yuen, A. Torralba. Dense Scene Alignment using SIFT Flow for Object Recognition. *CVPR*, 2009. 3
- [54] D. Liu, M. Scott, R. Ji, H. Yao, and X. Xie. Geolocation Sensitive Image Based Advertisement Platform. *ACM Multimedia*, 2009. 3, 4
- [55] X. Liu, H. Yao, and R. Ji. Exploring Statistical Properties for Semantic Annotation: Sparse Distributed and Convergent Assumptions for Keywords. *IEEE International Conference on Acoustic, Speech, and Signal Processing*, 2009. 3
- [56] R. Ji, P. Xu, H. Yao, J. Wang, and Z. Zhang. Real-Time Annotation by Manifold-based Biased Fisher Discriminant Analysis. *Visual Communication and Image Processing*, 2008. 3
- [57] R. Ji, H. Yao, J. Wang, Z. Zhang, and P. Xu. Clustering Based Cascade SVM Assemble for Adaptive Relevance Feedback Learning. *IEEE International Conference on Multimedia and Expo*, 2008. 3
- [58] R. Ji, Pengfei Xu, Hongxun Yao, Zhen Zhang, Xiaoshuai Sun, and Tianqiang Liu. Directional Correlation Analysis of Local Haar Binary Pattern (LHBP) for Text Detection. *IEEE International Conference on Multimedia and Expo*, 2008. 3
- [59] X. Sun, R. Ji, H. Yao, T. Liu, P. Xu, and X. Liu. Place Retrieval with Graph-View Model. *ACM Conference on Multimedia Information Retrieval*, 2008. 3
- [60] X. Liu, R. Ji, H. Yao, P. Xu, X. Sun, and T. Liu. Isomorphic Manifold Learning for Image Retrieval and Keyword Annotation. *ACM Conference on Multimedia Information Retrieval*, 2008. 3
- [61] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabulary and fast spatial matching. *CVPR*, 1-8, 2007. 4
- [62] E.-D. Eade and T.-W. Drummond. Unified loop closing and recovery for real time monocular SLAM. *BMVC*, 2008. 4
- [63] T. Yeh, J. Lee, and T. Darell. Adaptive vocabulary forest for dynamic indexing and category learning. *CVPR*, 1-8, 2007. 4
- [64] M. Cristani, A. Perina, U. Castellani, and V. Murino. Geolocated image analysis using latent representations. *CVPR*, 1-8, 2008. 4
- [65] D. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. *WWW*, 761-770, 2009. 4
- [66] A. Irschara, C. Zach, J. Frahm, H. Bischof. From structure-from-motion point clouds to fast location recognition. *CVPR*, 2599-2606, 2009. 4
- [67] J.-X. Xiao, J.-N. Chen, D.-Y. Yeung, and L. Quan. Structuring visual words in 3D for arbitrary-view object localization. *ECCV*, 725-737, 2008. 4

- [68] J. Hays and A. Efros. IMG2GPS: estimating geographic information from a single image. *CVPR*, 2008. 4
- [69] E. Kalogerakis, O. Vesselova, J. Hays, A. Efros, and A. Hertzmann. Image sequence geolocation with human travel priors. *CVPR*, 1-8, 2009. 4
- [70] Y.T Zheng., M. Zhao, Y. Song, and H. Adam. Tour the world: building a web-scale landmark recognition engine. *CVPR*, 1085-1092, 2009. 4
- [71] R. Ji, Y. Gao, B. Zhong, H. Yao, and Q. Tian. Mining City Landmarks by Modeling Reconstruction Sparsity. *ACM Transactions on Multimedia Computing, Communications and Applications*, 7S(1), 31-52, 2011. 5
- [72] R. Ji, X. Xie, H. Yao, and W.-Y. Ma. Mining City Landmarks by Graph Modeling. *ACM Multimedia*, 2009. 5
- [73] E. Paquet, A. Murching, T. Naveen, A. Tabatabai, and M. Rioux. Description of shape information for 2-D and 3-D objects. *Signal Process. Image Commun.*, vol. 16, pp. 103-122, 2000. 5
- [74] Y. Gao, M. Wang, J. Shen, Q. Dai, N. Zhang. Intelligent Query: Open Another Door to 3D Object Retrieval. *ACM Conference on Multimedia*, pp. 1711-1714, 2010. 5
- [75] Y. Gao, Y. Yang, Q. Dai, N. Zhang. Representative Views Re-Ranking for 3D Model Retrieval with Multi-Bipartite Graph Reinforcement Model. *ACM Conference on Multimedia*, pp. 947-950, 2010. 5
- [76] Y. Gao, Y. Yang, Q. Dai, N. Zhang. 3D Object Retrieval with Bag-of-Region-Words. *ACM Conference on Multimedia*, pp. 955-958, 2010. 6
- [77] Y. Gao, Q. Dai, N. Zhang. 3D Model Comparison using Spatial Structure Circular Descriptor. *Pattern Recognition*, vol.43, no.3, pp. 1142-1151, 2010. 5
- [78] Y. Gao, J. Tang, H. Li, Q. Dai, N. Zhang. View-based 3D Model Retrieval with Probabilistic Graph Model. *Neurocomputing*, vol.73, no.10-12, pp. 1900-1905, 2010. 5
- [79] Y. Gao, Q. Dai, M. Wang, N. Zhang. 3D Model Retrieval using Weighted Bipartite Graph Matching. *Signal Processing: Image Communication*, vol.26, no.1, pp. 39-47, 2011. 5
- [80] Y. Gao, J. Tang, R. Hong, S. Yan, Q. Dai, N. Zhang, T.-S. Chua. Camera Constraint-Free View-Based 3D Object Retrieval. *IEEE Transactions on Image Processing*, vol.21, no.4, pp. 2269-2281, 2012. 5
- [81] Y. Gao, M. Wang, Z. Zha, J. Shen, X. Li, X. Wu. Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search. *IEEE Transactions on Image Processing*. In press. 5
- [82] Y. Gao, M. Wang, D. Tao, R. Ji, Q. Dai. 3D Object Retrieval and Recognition with Hypergraph Analysis. *IEEE Transactions on Image Processing*. In press. 5
- [83] Y. Gao, M. Wang; R. Ji; Z. Zha; J. Shen. K-Partite Graph Reinforcement and Its Application in Multimedia Information Retrieval. *Information Science*, vol.194, no.1, pp. 224-239, 2012. 5
- [84] Q. Xiao, N. Zhang, F. Li, Y. Gao. Object detection based on combination of local and spatial information. *Journal of Systems Engineering and Electronics* vol. 22, no. 4, August 2011, pp.715-720. 5
- [85] B. Leibe and B. Schiele. Analyzing appearance and contour based methods for object categorization. *IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2003. 5
- [86] Q. Xiao, H. Wang, F. Li, Y. Gao. 3D object retrieval based on a graph model descriptor *Neurocomputing*, vol. 74. no.17, pp. 3486-3493, 2011. 5
- [87] C. Ip, D. Lapadat, L. Soeger, and W. C. Regli. Using shape distributions to compare solid models. *ACM Symp. Solid Model. Appl.*, 2002, pp. 273-280. 5
- [88] B. Leng and Z. Qin. A powerful relevance feedback mechanism for content-based 3-D model retrieval. *Multimedia Tools Appl.*, vol. 40, no. 1, pp. 135-150, Oct. 2008. 5

- [89] T. Funkhouser, P. Min, M. Kazhdan, J. Chen, A. Halderman, D. Dobkin, and D. Jacobs. A search engine for 3-D models. *ACM Trans. Graph.*, vol. 22, no. 1, pp. 83-105, Jan. 2003. 5
- [90] J. Tangelder and R. Veltkamp. Polyhedral model retrieval using weighted point sets. *Int. J. Image Graph.*, vol. 3, no. 1, pp. 209-229, 2003. 5
- [91] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. V. Gool. Towards multi-view object class detection. in *Proc. Comput. Vis. Pattern Recognit.*, 2006, pp. 1589-1596. 5
- [92] S. Savarese and L. Fei-Fei. 3D generic object categorization, localization and pose estimation. in *Proc. IEEE Int. Conf. Comput. Vis.*, Riode Janeiro, Brazil, Oct. 2007. 5
- [93] Y. Ohkita, Y. Ohishi, T. Furuya, R. Ohbuchi. Non-rigid 3D Model Retrieval Using Set of Local Statistical Features *IEEE ICME 2012 Workshop on Hot Topics in 3D Multimedia (Hot3D)*, 2012. 5
- [94] M. Sun, H. Su, S. Savarese, and L. Fei-Fei. A multi-view probabilistic model for 3-D object classes. in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 1247-1254. 5
- [95] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories. *Proc. IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, Oct. 2009, pp. 213-220. 5
- [96] Y. Gao, M. Wang, Z. Zha, Q. Tian, Q. Dai, N. Zhang. Less is More: Efficient 3D Object Retrieval with Query View Selection. *IEEE Transactions on Multimedia*, vol.11, no.5, pp.1007-1018, 2011. 5
- [97] T. F. Ansary, M. Daoudi, and J. P. Vandeboere. A Bayesian 3-D search engine using adaptive views clustering. *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 78-88, Jan. 2007. 5
- [98] Y. Gao, M. Wang, S. Yan, J. Shen, D. Tao. Tag-Based Social Image Search with Visual-Text Joint Hypergraph Learning. *ACM Conference on Multimedia.*, 2011. 5
- [99] Y. Gao, W.-B. Wang, J.-H. Yong, H.-J. Gu. Dynamic Video Summarization using Two-Level Redundancy Detection. *Multimedia Tools and Applications*, vol.42, no.2, 233-250, 2009 5
- [100] Y. Gao, J. Tang, X. Xie. Key frame vector and its application to shot retrieval. *Proceedings of the 1st international workshop on Interactive multimedia for consumer electronics*, pp. 27-34, 2009
- [101] Y. Gao, Q. Dai. Clip-based Video Summarization and Ranking. *ACM International Conference on Image and Video Retrieval*, pp. 135-140, 2008. 5
- [102] Y. Gao, W.-B. Wang, J.-H. Yong. A Video Summarization Tool using Two-Level Redundancy Detection for Personal Video Recorder. *IEEE Transactions on Consumer Electronic*, vol.54, no.2, 521-526, 2008 5
- [103] Y. Gao, Q. Dai. Shot-based Similarity Measure for Content-based Video Summarization. *IEEE International Conference on Image Processing*, 2008 5
- [104] T. Wang, Y. Gao, P. Wang, E. Li, W. Hu, Y. Zhang, J. Yong. Video Summarization by Redundancy Removing and Content Ranking. *ACM Conference on Multimedia*, pp. 577-580, 2007 5
- [105] T. Wang, Y. Gao, J. Li, P. Wang, X. Tong, W. Hu, Y. Zhang, J. Li. 3D object retrieval based on a graph model descriptor *International workshop on TRECVID video summarization*, pp. 79-83, 2007. 5
- [106] T. Furuya and R. Ohbuchi. Dense sampling and fast encoding for 3-D model retrieval using bag-of-visual features *ACM Int. Conf. Image Video Retrieval*, 2008. 5
5
- [107] R. Ohbuchi, K. Osada, T. Furuya, T. Banno. Salient local visual features for shape-based 3D model retrieval *IEEE Shape Modeling International*, 2008. 5

- [108] K. Osada, T. Furuya, R. Ohbuchi. SHREC'08 Entry: Local 2D Visual Features for CAD Model Retrieval *IEEE Shape Modeling International*, 2008. 5
- [109] R. Ohbuchi, T. Shimizu. Ranking on Semantic Manifold for Semantic 3D Model Retrieval *ACM MIR*, 2008. 5
- [110] R. Ohbuchi, T. Furuya. Accelerating Bag-of-Features SIFT Algorithm for 3D Model Retrieval *the SAMT 2008 Workshop on Semantic 3D Media*, Koblenz, Germany, 2008. 6
- [111] R. Ohbuchi and T. Furuya. Scale-Weighted Dense Bag of Visual Features for 3D Model Retrieval from a Partial View 3D Model *IEEE ICCV 2009 workshop on Search in 3D and Video*, 2009. 6
- [112] R. Ohbuchi, M. Tezuka, T. Furuya, T. Oyobe. Squeezing Bag-of-Features for Scalable and Semantic 3D Model Retrieval *International Workshop on Content-Based Multimedia Indexing*, 2010. 6
- [113] R. Ohbuchi, T. Furuya. Distance Metric Learning and Feature Combination for Shape-Based 3D Model Retrieval *ACM workshop on 3D object retrieval*, 2010. 6
- [114] M. Endoh, T. Yanagimachi, R. Ohbuchi. Efficient manifold learning for 3D model retrieval by using clustering-based training sample reduction *ICASSP*, 2012. 6
- [115] S. Kawamura, K. Usui, T. Furuya and R. Ohbuchi. Local Geometrical Feature with Positional Context for Shape-based 3D Model Retrieval *Eurographics 2012 Workshop on 3D Object Retrieval*, 2012. 6
- [116] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod. Inverted index compression for scalable image matching. *DCC*, 525-252, 2010. 7, 8
- [117] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, J. Singh, and B. Girod. Transform coding of image feature descriptors. *VCIP*, doi:10.1117/12.805982, 2009. 7
- [118] M. Makar, C. Chang, D. Chen, S. Tsai, and B. Girod. Compression of image patches for local feature extraction. *ICASSP*, 821-824, 2009. 7
- [119] R. Ji, H. Yao, Z. Zhang, P. Xu, and X. Liu. Exploit Visual Textual Fusion for Semantic Supervised Region Retrieval. *ACM Multimedia Systems Journal*, 15(4): 201-219, 2009. 7
- [120] R. Ji, L.-Y. Duan, J. Chen, S. Yang, H. Yao, T. Huang, and W. Gao. Learning the Trip Suggestion from Landmark Photos on the Web. *IEEE International Conference on Image Processing*, 2011. 7
- [121] R. Ji, L.-Y. Duan, J. Chen, S. Yang, T. Huang, H. Yao, and W. Gao. PKUBench: A Context Rich Mobile Visual Search Benchmark. *IEEE International Conference on Image Processing*, 2011. 7, 8
- [122] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *CVPR*, II-506 - II-513, 2004. 7
- [123] H. Bay, T. Tuytelaars, and L. Van Gool. SURF: Speed up robust features. *ECCV*, 450-459, 2006. 7
- [124] G. Hua, M. Brown, and S. Winder. Discriminant embedding for local image descriptors. *ICCV*, 1-8, 2007. 7
- [125] S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar. Location coding for mobile image retrieval. *MobileMedia*, 2010. 7
- [126] V. Chandrasekhar, D. Chen, A. Lin, G. Takacs, S. Tsai, N. Cheung, Y. Reznik, R. Grzeszczuk, and B. Girod. Comparison of local feature descriptors for mobile visual search. *ICIP*, 3885-3888, 2010. 7
- [127] C. Yeo, P. Ahammad, and K. Ramchandran. Rate-efficient visual correspondences using random projections. *ICIP*, 217-220, 2008. 8
- [128] Y. Weiss, A. Torralba, and R. Fergus. Spectral Hashing. *NIPS*, 1753-1760, 2009. 8
- [129] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *PAMI*, (99), 1-1, 2010. 8
- [130] H. Jegou, M. Douze, and C. Schmid. Packing bag-of-features. *ICCV*, 1-8, 2009. 8
- [131] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large databases for recognition. *CVPR*, 1-8, 2008. 8

- [132] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. *CVPR*, 3360-3367, 2010. [8](#)