OBSIR: OBJECT-BASED STEREO IMAGE RETRIEVAL

Xiangyang Xu, Wenjing Geng, Ran Ju, Yang Yang, Tongwei Ren and Gangshan Wu

State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing, China
xiangyang.xu@smail.nju.edu.cn, jenngeng@gmail.com, juran@smail.nju.edu.cn,
charlie.yang.nju@gmail.com, {rentw, gswu}@nju.edu.cn

ABSTRACT

Recent years, the stereo image has become an emerging media in the field of 3D technology, which leads to an urgent demand of stereo image retrieval. In this paper, we attempt to introduce a framework for object-based stereo image retrieval (OBSIR), which retrieves images containing the similar objects to the one captured in the query image by the user. The proposed approach consists of both online and offline procedures. In the offline procedure, we propose a salient object segmentation method making use of both color and depth to extract objects from each image. The extracted objects are then represented by multiple visual feature descriptors. In order to improve the image search efficiently, we construct an approximate nearest neighbor (ANN) index using cluster-based locality sensitive hashing (LSH). In the online stage, the user may supply the query object by selecting a region of interest (ROI) in the query image, or clicking one of the objects recommended by the salient object detector. For the image retrieval evaluation we build a new dataset containing over 10K stereo images. The experiments on this dataset show that the proposed method can effectively recommend the correct object and the final retrieval result is also better than other baseline methods.

Index Terms— Stereo image retrieval, object retrieval, salient object detection, query object recommendation

1. INTRODUCTION

Nowadays, numerous 3D devices such as stereo cameras and 3DTV have experienced an explosive growth in the industrial community and the stereo images have become an emerging media widely spread in people's daily life. With the sharp increasing of stereo image data, how to manage and access them efficiently turns out to be an urgent problem, which is just the same as digital images about 2 decades ago [1]. In the world of monoscopic images, content-based image retrieval (CBIR) enables us to access relevant images by image examples while object-based image retrieval (OBIR) methods [2, 3, 4] accomplish a search with a region of interest regarded as the desired object by user interaction.

Unfortunately, there are few research works on stereo image retrieval. In order to solve this urgent problem, we introduce a complete framework for object-based stereo image retrieval. First, a preprocessing including stereo rectification and stereo matching is adopted to produce the disparity map for each image which encodes the depth information. Second, the object segmentation procedure is performed by a salient object detector making use of depth information. Then, multiple visual features are extracted including the bag-of-visual-words (BoVW) and they are used to represent the objects. Finally, the feature vectors are indexed by a clustering-based LSH. In the online search phase, the user is first asked to upload an example image. To select the query object, the user may either drag a region of interest or pick up an object from the query recommendations. Based on LSH indexing, a list of stereo images is returned to the user efficiently. To evaluate the effectiveness of the proposed framework, we build a new stereo image dataset called "OBSIR dataset". In summary, our major contributions include:

- A novel framework for object-based stereo image retrieval;
- A salient object detection method that contributes to a better object segmentation and serves as a query recommendation in user interaction phase;
- A novel stereo image dataset designed for image retrieval evaluation that comes from three common sources, the websites, daily life photography and stereo movies.

To the best of our knowledge, we believe our work is the first attempt to explicitly establish a systematically framework for object-based stereo image retrieval, and also the first one to build up a stereo image dataset for the evaluation of stereo retrieval task.

The remaining of this paper is organized as follows. In Section 2, a brief review of the related work is introduced. The systemic overview and detailed approach is described in Section 3 and evaluated by a few experiments on the OBSIR dataset, as shown in Section 4. Finally, we conclude this paper with some remarks on the feature work in Section 5.

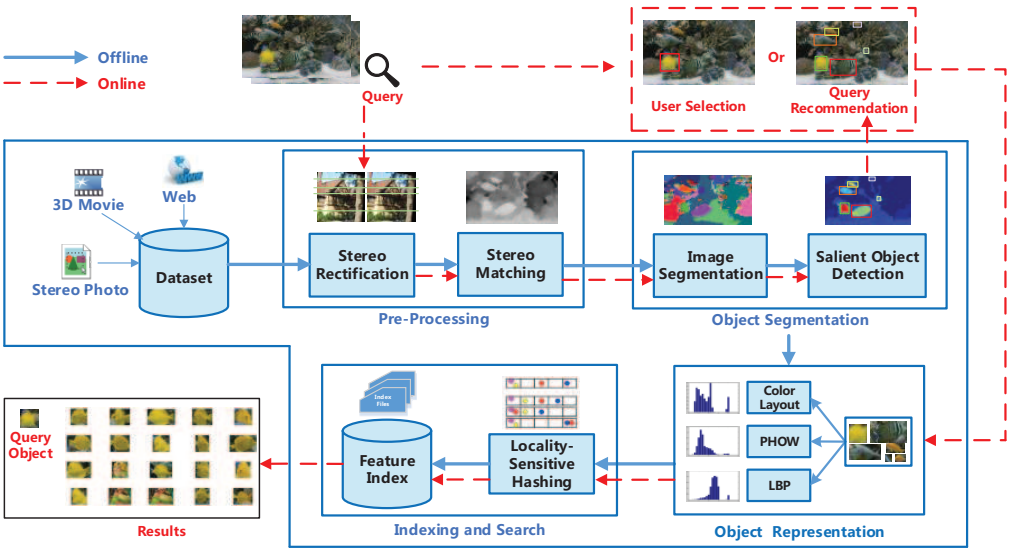

Fig. 1. Object-based stereo image retrieval framework

2. RELATED WORK

This paper serves as a variant of OBIR approach which adapts to stereo images. Meanwhile, the retrieval system also makes use of stereo matching, object segmentation and ANN-based search techniques. Therefore, in the rest of this section, we are going to review the research status on these mentioned domains.

Stereo matching. Stereo matching remains to be a challenging problem which catches lots of interests [5]. A lot of excellent methods have been proposed to solve this problem. In our application, we need the disparity maps to maintain essentially correct structure information and do not highly depend on accurate disparity values. For this purpose and considering of the computational efficiency, the ground control points based method [6] is an appropriate choice.

Salient object segmentation. The problem has been proposed early from 1990's, and a series of approaches [7, 8, 9] have been proposed from different aspects during the following years. For example, Jie Feng et al. [9] proposed a bounding-box-dependent method that the object saliency is calculated by comparing the content inside and outside the bounding box. And recently disparity information is introduced for saliency analysis [10], which also works on stereo images. Differed from these methods, our approach first perform an image segmentation and then follow with a saliency based ranking in order to extract multiple objects.

ANN-based search. Over the past years, numerous approaches [11, 2, 12] are proposed to avoid brute-force comparison when searching the k nearest neighborhood (kNN) for the query sample, they are so called Approximate Nearest Neighborhood search (ANN) methods. Generally, ANN methods can be roughly divided into two categories. One is tree-based methods such as optimized k-d tree [12], and the other is hashing-based methods. For example, Locality-Sensitive Hashing (LSH) [11] is one of the most popular ANN algorithms which can efficiently get the approximate results for queries. In order to reduce the number of the exhaustive similarity evaluations, we develop a clustering-based LSH indexing approach for the object-based stereo image retrieval method.

Object-based image retrieval. Object-based image retrieval is a well-studied problem in CBIR, for which the user usually captures the demand object through e.g., a bounding box. In [13] images are segmented into small regions, based on which the query object is modeled using Latent Semantic Analysis (LSA). In [3], each image is represented as a bag of visual words, and the author makes use of the language modeling to derive the ranking function. The visual words locating outside the object region is also taken into account as context.

3. OBJECT-BASED STEREO IMAGE RETRIEVAL

Given a stereo image, and the object region that the user specifies, we aim to search for the relevant images containing the similar object. The framework of our approach is shown in Fig. 1, which is composed of two pipelines: offline procedure and online processing. For the offline procedure, we first collect a large dataset of stereo images. Then a preprocessing including image resizing, duplicate removal, stereo image rectification and stereo matching is taken. In this step we get the disparity maps which encode the depth information. After that, we extract salient objects from the image. For each object, we extract visual features and use the BoVW model for representation. At last, the feature vectors are indexed by a clustering-based LSH. For the online procedure, a stereo image is uploaded by the user first and then passed through the preprocessing and object segmentation module sequentially. A few objects are generated then and displayed in the object box to give a recommendation to the user. The user may directly pick one object or drag a rectangular region of interest as the final query. Then the query region is passed

to the object representation module and converted to a fixed BoVW vector. The search is performed by the clustering-based LSH and at last, a list of objects is returned and displayed in the user interface. Object results are highlighted in their owner images to show the entire context of the objects.

3.1. Preprocessing

As the stereo images in our dataset are acquired from many different sources and vary greatly in the resolution, we first resize the images to a fixed height and remove exact duplicates from our dataset as they appear to be redundant in retrieval. Then we rectify all stereo images using [14] so that the structure displacement only occurs at the horizontal direction. Next we perform stereo matching using the ELAS algorithm [6] to obtain the disparity maps encoding the depth information. The basic idea is to search a set of robustly matched points as supports to reduce searching ambiguities. The seed supports make the algorithm perform robust for variational stereo images. Then we optimize the results using [15]. A fast version [16] is implemented to speed up computation. The disparity maps are stored together with the original stereo images for following processing.

3.2. Object segmentation

Inspired from [17], we follow the "split and merge" idea for object segmentation. The image is segmented into several regions where the intra-region similarity and the inter-region dissimilarity are expected to be maximized. These regions can be assumed as candidate objects. Then a saliency analysis based on color and depth is performed for ranking and filtering the objects.

Due to the noise of the color image and inaccuracy of the disparity map, we run the segmentation on superpixel level. SLIC [18] is employed to segment the image into superpixels. Then we construct an undirected graph $G = (V, E)$ where each node $v_i \in V$ indicates a superpixel and each edge $(v_i, v_j) \in E$ connects two neighboring superpixels v_i and v_j. The weight of an edge stands for the similarity between v_i and v_j:

$$\begin{aligned} Sim(v_i, v_j) &= \min(D_c(v_i, v_j), \lambda_s D_d(v_i, v_j)), \\ D_c(v_i, v_j) &= |\overline{I(v_i)} - \overline{I(v_j)}|, \\ D_d(v_i, v_j) &= \frac{|\overline{d(v_i)} - \overline{d(v_j)}|}{\min(\overline{d(v_i)}, \overline{d(v_j)})}, \end{aligned} \tag{1}$$

where $D_c(v_i, v_j)$ measures the absolute difference of the mean color $\overline{I(\cdot)}$ between v_i and v_j in Lab colorspace. $D_d(v_i, v_j)$ is the relative difference of mean disparity $\overline{d(\cdot)}$. λ_s is a weighted factor to balance the power between color and disparity. Then each pair of neighboring superpixels is merged if the weight of the linking edge is smaller than a threshold K_s and then a set of disjoint regions $O = \{o_1, o_2, \cdots, o_n\}$ are produced. We view the merge operation as a coarse object segmentation and follow with a three-step refinement:

- Small objects are removed. More formally, if $S(o_i) < S_d$ the region is deleted, where $S(o_i)$ indicates the ratio between the area of region o_i and the total area of the image. S_d is a threshold;
- Large objects are split. If $S(o_i) > S_s$, the merge step is redone by setting K_s to γK_s, where γ indicates a decay rate and S_s is a threshold;
- Sometimes an object is split into two or more parts due to occlusion. To solve this problem, for each object o_i we search a surrounding area in a range $[R_c, R_n]$ for nonadjacent objects $o_j \in \{O \setminus o_i\}$ and merge the two objects if $Sim'(o_i, o_j) < K_o$, where:

$$Sim'(o_i, o_j) = \max(D'_c(o_i, o_j), \lambda_o D'_d(o_i, o_j)) \tag{2}$$

D'_c is the intersection distance between the Lab color histograms of o_i and o_j, D'_d is the relative difference of mean disparity between object o_i and o_j. λ_o is a weighted factor.

Next we rank the objects according to their normalized saliency values. We first employ [19] to compute the color saliency Sal_c and [10] to compute the depth saliency Sal_d respectively. Then the final saliency is computed as:

$$Sal = G(x, y, \sigma_x, \sigma_y) * Sal_c Sal_d, \tag{3}$$

where $G(x, y, \sigma_x, \sigma_y)$ is a 2D Gaussian distribution function with σ_x and σ_y equal to half of the width and height of the image respectively. The purpose of this Gaussian window is to emphasize central objects. Then we rank the objects according to the normalized saliency value: $Sal_o(o_i) = Sal/S(o_i)$. We set two conditions to filter the objects:

- The normalized saliency value ranks top K_r in the object list of the image;
- The normalized saliency value is within a rate of S_e of the highest one.

At last each salient object is clipped using the bounding box of the object region for efficient feature extraction.

Query object recommendation. The salient object extraction is not only applied for object retrieval, but also regarded as an online service called query recommendation. Once the user uploaded a stereo image, he/she may either choose to select the query object using the rectangle tool or by clicking on one of the objects we recommended. The role of the latter is two-fold. First, it supplies more convenient interaction for users. Second, the other recommendations in the same image may arouse the user's interest and thus contribute for following retrievals.

3.3. Object representation

Adopted from several successful works on monoscopic image retrieval [20, 21, 22], three features are chosen to cover the visual information of the objects and the well-known BoVW model is used to represent the features. First, color layout from MPEG-7 standard is adopted to give a global description of the objects, encoding its color information. Second, local features from each object are extracted and a visual codebook is generated by clustering the feature vectors. Then each object is represented by a histogram of the visual words. The two local features pyramid histogram of visual

words (PHOW) [21] and local binary pattern (LBP) [23] have obtained promising results in image retrieval and thus are adopted in our work.

The color layout is computed over the bounding box of the given object to contain a few contexts and partitioned into 8*8 grids. Thus for an image with RGB channels the feature forms a vector of 192 dimensions totally. For PHOW, the vocabulary contains 10000 bins, trained from representative images using K-Means. For LBP, the uniform version of LBP [24] is adopted where the 8-bit LBP features are quantized to 58 patterns. We treat the 58 quantized patterns as LBP visual words, and the uniform LBP histogram are normalized. At last, the 3 representations are indexed separately and combined using linear regression as the final results.

3.4. Index construction

Exhaustive linear search is not practical in real image retrieval systems due to its inefficiency on the large scale of the datasets and the high dimensionality of visual features. To deal with this problem, a cluster-based locality-sensitive hashing is used in our system to speed up the query response. As conventional LSH needs a lot of hash tables to improve the search quality and thus results in a quite long evaluation list, we first perform a clustering of objects according to the possibly intrinsic clustering property of objects. In implementation, K-means is applied to give a coarse classification and then LSH [11] is built on the object clusters.

4. EXPERIMENTS

In this section, we demonstrate the conducted experiments of this work. After the introduction of the dataset and general experiment settings, we first illustrate an evaluation on the salient object detection, then we present the experiment of object retrieval.

4.1. OBSIR dataset

Since the public benchmark datasets like Middlebury stereo dataset [25] and KITTI vision benchmark [6] provide too limited amount of images to support a retrieval system, the experiments are performed on the OBSIR dataset which is built by ourselves from three common sources: stereo images downloaded from website, realistic photographs taken by a stereo camera and snapshots captured from 3D movies. First, we download 3000 stereo images from Flickr[1]. Then, we collect 2500 stereo photos taken from real world by a stereo camera. At last, we capture a number of snapshots from 25 famous stereo movies, such as Avatar, Monsters University and so on. In total, the OBSIR dataset contains 10513 stereo images.

4.2. Object segmentation evaluation

4.2.1. Experimental settings

The motivation of this experiment is to evaluate the correctness of the salient object detector. We select 500 images from the database and manually label the foreground objects by bounding boxes as ground truth. Each object guess is regarded as correct if the intersection between the predict object box and ground truth is larger than half of their union according to [26]. We compare our approach with three salient based methods: Salient object detection by composition (COMP)[9], Salient object detection by global contrast (RC)[19] and stereo-based saliency (SS)[10], standing for state-of-the-art. The parameters of our approach are set as $\{\lambda_s, K_s, S_d, S_s, \gamma, R_c, R_n, K_o, \lambda_o, K_r, S_e\}$ = $\{0.003,0.06,0.0005,0.5,0.8,10,50,0.12,10,5,0.8\}$. Since generally the number of objects of an image in our dataset does not exceed 5, we choose precision and recall at the truncate level 5 for evaluation.

4.2.2. Experimental results and analysis

Table 1 shows the comparison of our method and the other three baselines. Our method outperforms all the other methods in the term of both precision and recall. RC [19] and SS [10] are designed for single salient object extraction thus both get high precision but low recall once there are multiple objects in one image. It is worth noting that SS performs better than RC owing to the employment of depth information. The method COMP in [9] can detect multiple salient windows and thus achieves a higher recall than RC and SS.

In Fig. 2 we show a few examples of the results generated using the compared four methods, from which we can see that the proposed method can give a correct bounding box to each object no matter there is one or more objects on the image. On the contrary the method of SS bounds all the three objects in the second image using one box, and it blocks the user from selecting one of them. We can also observe that COMP produces a lot of fake bounding boxes with useless content inside. It suggests that saliency is less reliable without the help of depth information in stereo images.

Table 1. Salient object segmentation performance

	Ours	COMP [9]	RC [19]	SS [10]
Precision	0.703	0.304	0.592	0.614
Recall	0.571	0.527	0.259	0.268

Fig. 2. Salient object extraction examples

[1] http://www.flickr.com

4.3. Object-based stereo image retrieval

4.3.1. Experimental settings

To quantitatively evaluate the retrieval performance of our method, 100 stereo images are randomly chosen as queries. We adopt the proposed query object recommendation approach and pick up the correct objects from each image and finally there are 153 object queries in total.

In this experiment the proposed retrieval framework is compared with two baseline method which directly conducted on the left images and didn't make use of stereo information:

- CBIR, instead of using objects, this baseline directly indexes the features of the entire image. Consequently, we also use images as queries instead of objects;
- OBIR, this baseline adopts a 2D state-of-the-art salient object segmentation approach described in order to extract the objects in the images. In this paper, we adopt Jie Feng et al. salient object segmentation approach [9] to construct OBIR index.

Besides, we utilize three popular measurements in CBIR, including Precision-Recall curve, mean Average Precision (mAP) and Normalize discounted cumulative Gain (NDCG). AP is defined as the average precision at various recall levels, and mAP is defined as the average APs of of all queries. NDCG is defined as follows:

$$\text{NDCG@}k = \frac{\text{DCG@}k}{\text{IDCG@}k}, \tag{4}$$

where

$$\text{DCG@}k = \sum_{i=1}^{k} \frac{2^{r_i} - 1}{\log_2(i+1)}, \tag{5}$$

In Equation (5), r_i is the ground truth labeling relevance of i_{th} returned image. And in Equation (4), IDCG stands for ideal DCG which can be calculated like DCG.

Due to the manual workload, instead of labeling the ground truth before the evaluation, we propose to judge the result image list retrieved by the query , and we set the maximum truncate level of our experiment to 100, that is, each query will obtain 100 judgments, indicating the relevance of each result image. We characterized the returned results into 4 groups. Excellent and good represents positive results, while bad and fail are the negative ones.

4.3.2. Experimental results and analysis

Figure 3 illustrates the retrieval performance of the three compared approaches, from which we can see that the proposed retrieval framework outperforms the other two baselines in the term of all the three measurements. Within the three methods, CBIR performs the worst because only the image-to-image retrieval is applied. It tells that for object images, the visual features will be more effective if the objects are detected as presented individually. And we can also infer from the performance gap between OBIR that the proposed method achieves a more accurate object segmentation using depth information and leads to a better retrieval performance.

Figure 4 shows the top 6 retrieved images by the 5 selected object queries, which indicates that the proposed framework

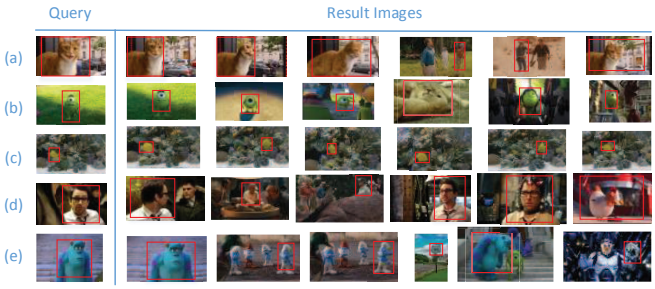

Fig. 4. Top returned examples of some queries from OBSIR

can effectively retrieve the demand images. However, the query of (e) obtains some rather disappointing results. We argue that it is because the descriptor of PHOW and LBP degrade a lot in the textureless areas and the color descriptor dominate the similarity and images with similar colors are retrieved. Besides, the saliency object extraction fails on the last image of query (d) and the 4th result of query (b) due to incorrect disparity maps.

5. CONCLUSION AND FUTURE WORK

This paper presented an object-based retrieval framework for stereo images. In the offline part, we first segmented the objects from each image by saliency. Then multiple visual features were extracted from each object region and indexed using LSH. In order to extract the correct objects, we proposed a novel salient object extraction method, making use of depth information. In the online part, we simplified the user interaction by recommending suspect object regions. It allowed the users to select among the recommended objects as well as drawing the entire bounding box. In the experiment of salient object segmentation, our approach captured more correct images than other baseline methods, and in the evaluation of retrieval, the proposed framework outperformed other retrieval frameworks based on monoscopic images.

We believe that the proposed framework can be regarded as a fundamental step of multiple applications, thus we envision our work in the future as follows. First, we are going to bring in some stereo geometric verification approaches to re-rank the search result, such that the retrieval performance can be further improved. Second, we are considering to encode the geometrical relationship between multiple objects of the same image, so as to make the framework operable for scene images. Finally, we are trying to develop some advanced query object recommendation method, making the querying process more effective and efficient.

Acknowledgements. This work is supported by the Natural Science Foundation of China (No.61021062 and No.61202320), Research Project of Excellent State Key Laboratory (No.61223003), Natural Science Foundation of Jiangsu Province (No.BK2012304) and National Special Fund (No.2011ZX05035-004-004HZ).

6. REFERENCES

[1] Shih-Fu Chang, "How far we've come: Impact of 20 years of multimedia information retrieval," *TOMCCAP*, vol. 9, no. 1s, pp. 42:1–42:4, October 2013.

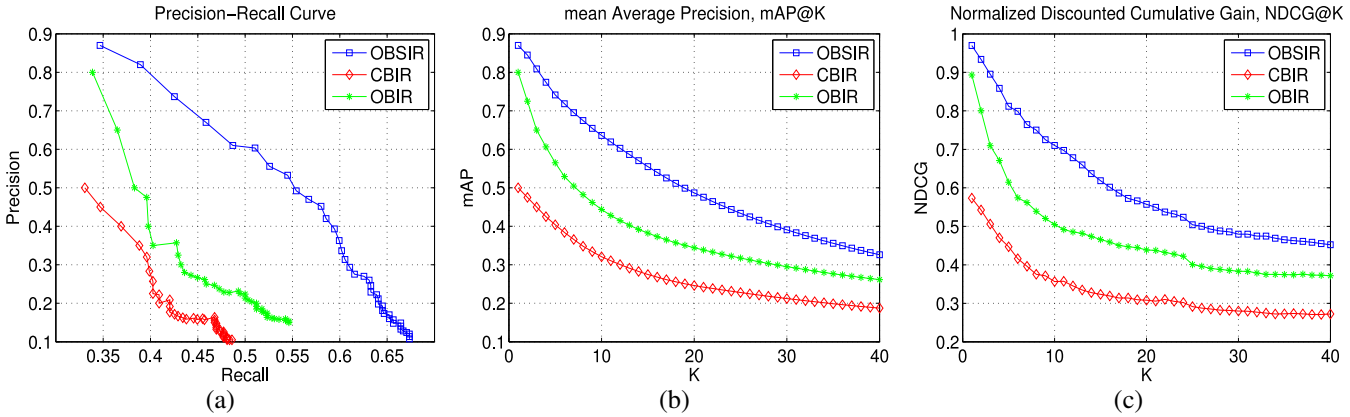

Fig. 3. Quantitative object-based image retrieval performance: (a) Precision-Recall Curve (b) mAP at different truncate levels (c) NDCG at different truncate levels

[2] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*. IEEE, 2007, pp. 1–8.

[3] Linjun Yang, Bo Geng, Yang Cai, Alan Hanjalic, and Xian-Sheng Hua, "Object retrieval using visual query context," *TMM*, vol. 13, no. 6, pp. 1295–1307, 2011.

[4] Xiaohui Shen, Zhe Lin, Jonathan Brandt, Shai Avidan, and Ying Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *CVPR*. IEEE, 2012, pp. 3013–3020.

[5] Daniel Scharstein and Richard Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *IJCV*, vol. 47, no. 1-3, pp. 7–42, 2002.

[6] Andreas Geiger, Martin Roser, and Raquel Urtasun, "Efficient large-scale stereo matching," in *ACCV*, pp. 25–38. 2011.

[7] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk, "Frequency-tuned salient region detection," in *CVPR*. IEEE, 2009, pp. 1597–1604.

[8] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal, "Context-aware saliency detection," *TPAMI*, vol. 34, no. 10, pp. 1915–1926, 2012.

[9] Jie Feng, Yichen Wei, Litian Tao, Chao Zhang, and Jian Sun, "Salient object detection by composition," in *ICCV*. IEEE, 2011, pp. 1028–1035.

[10] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu, "Leveraging stereopsis for saliency analysis," in *CVPR*. IEEE, 2012, pp. 454–461.

[11] Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in *SoCG*. ACM, 2004, pp. 253–262.

[12] You Jia, Jingdong Wang, Gang Zeng, Hongbin Zha, and Xian-Sheng Hua, "Optimizing kd-trees for scalable visual descriptor indexing," in *CVPR*. IEEE, 2010, pp. 3392–3399.

[13] Lukas Hohl, Fabrice Souvannavong, Bernard Merialdo, and Benoit Huet, "Enhancing latent semantic analysis video object retrieval with structural information.," in *ICIP*. IEEE, 2004, pp. 1609–1612.

[14] Richard Hartley and Andrew Zisserman, *Multiple view geometry in computer vision*, vol. 2, 2000.

[15] Qingxiong Yang, Ruigang Yang, James Davis, and David Nistér, "Spatial-depth super resolution for range images," in *CVPR*. IEEE, 2007, pp. 1–8.

[16] Kunal Narayan Chaudhury, Daniel Sage, and Michael Unser, "Fast o(1) bilateral filtering using trigonometric range kernels," *TIP*, vol. 20, no. 12, pp. 3376–3382, 2011.

[17] Pedro F Felzenszwalb and Daniel P Huttenlocher, "Efficient graph-based image segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, 2004.

[18] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *TPAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[19] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu, "Global contrast based salient region detection," in *CVPR*. IEEE, 2011, pp. 409–416.

[20] Fuxiang Lu, Xiaokang Yang, Rui Zhang, and Songyu Yu, "Image classification based on pyramid histogram of topics," in *ICME*. IEEE, 2009, pp. 398–401.

[21] Anna Bosch, Andrew Zisserman, and Xavier Muoz, "Image classification using random forests and ferns," in *ICCV*. IEEE, 2007, pp. 1–8.

[22] Xiaoli Yuan, Jing Yu, Zengchang Qin, and Tao Wan, "A sift-lbp image retrieval model based on bag-of features," in *ICIP*. IEEE, 2011, pp. 1061–1064.

[23] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *TPAMI*, vol. 24, no. 7, pp. 971–987, 2002.

[24] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," `http://www.vlfeat.org/`, 2008.

[25] Daniel Scharstein and Richard Szeliski, "Middlebury stereo vision page," `http://vision.middlebury.edu/stereo/`.

[26] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman, "The pascal visual object classes(voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.