DEPTH-BASED LOCAL FEATURE SELECTION FOR MOBILE VISUAL SEARCH

Zhaoliang Liu, Ling-Yu Duan, Jie Chen, Tiejun Huang

Institute of Digital Media, School of EE & CS, Peking University, Beijing, 100871, China
{zhaoliang_liu, lingyu, cjie, tjhuang}@pku.edu.cn

ABSTRACT

Selecting local features is crucial in generating robust compact descriptors for mobile visual search. The state-of-the-art MPEG Compact Descriptors for Visual Search (CDVS) standard has utilized the intrinsic characteristics (e.g., scale, orientation, peak, center distance, etc.) of interest points to select salient local features for selective aggregation and compression of local feature descriptors at different bit rates. In particular, the statistics of center distance was considered as an important attribute to select features in mobile visual search, which heavily relies on the assumption of a centralized object in a 2-dimensional query image. However, the ad-hoc assumption would probably fail to delineate query objects in a cluttered scene. In this paper, we propose to incorporate the depth cue to select local features. As most mobile phones are not yet equipped with depth sensor, we recover the disparity of local features through an auxiliary image to fast estimate the depth of a query image. The experiments have shown that, the incorporation of depth cue into feature selection can significantly improve the retrieval performance of the state-of-the-art CDVS compact descriptors at lower bit rates. For example, the mAP is improved from 84.5% to 88.6% at 512 bytes.

Index Terms— Mobile visual search, Local feature selection, Depth estimation, Interest points

1. INTRODUCTION

Mobile visual search [1, 2] aims to discover reference images containing the same target objects depicted by a query image, while the reference image database is hosted at remote server and query images are captured by mobile device. MPEG Compact Descriptors for Visual Search (CDVS) standard [3] provides the state-of-the-art solution for mobile visual search. The CDVS compact visual descriptor consists of an aggregated global descriptor [2, 4] and compressed local descriptors [2, 5]. CDVS supports typical client-server architectures. Visual descriptors are extracted and compressed on the mobile client, where retrieval is performed on the server using the transmitted compact descriptor as the query. In the environment of wireless transmission, CDVS aims to improve retrieval performance while reducing the data size of a compact descriptor.

Based on the image content, interest point detection can result in several hundred to several thousand local features. For small feature data size (512 bytes to 4KB), it is not feasible to include all local features. So selecting a subset of feature descriptors becomes critical [2, 6]. There are also other advantages of feature selection. Local feature descriptors are aggregated to form a global feature descriptor. Incorporating noisy local features would significantly degrade the discriminative power of the global descriptor [2, 7].

Many research efforts have been devoted to the selection of local feature. [8, 9, 10] select features that appear frequently in images containing the same object as stable features, which can only be utilized in pre-processing of image database or query expansion. [11, 12] predict useful features by training a classifier for visual descriptors. In particular, [6] estimates those features that are most likely to be matched correctly by intrinsic characteristics of interest points, which is adopted in CDVS. In this work, we further improve the performance of interest point characteristics based lightweight feature selection, which is favorable for mobile devices.

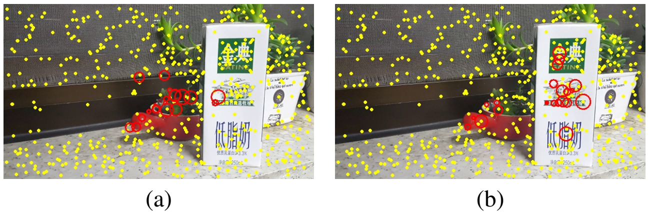

Fig. 1. Exemplar results of local feature selection. (a) local feature selection result of [6]; (b) local feature selection result of the proposed method. The yellow dots show the discard local features; the red circles show the selected local features, the larger circles indicate the larger probability to be matched correctly.

Although the center distance has been proved to be a useful complement to local feature characteristics by SIFT (e.g. scale, orientation, scale, etc) in [6], this assumption of centralized objects would fail to distinguish target objects from cluttered mobile scenes (as shown in Fig 1). In essence, as these characteristics are detected in 2D pixels, the feature selection model in [6] just relies on still image based 2D location statistics, rather than 3D location statistics, to delineate the model of target objects.

An important empirical finding is that no matter how greatly the viewpoint or target object location varies, people usually put the target object in a scene at a certain depth different from cluttered or non-cluttered background. Thus, depth is an effective cue for integrating the 3D model of a target object into feature selection.

There are two key problems in depth-based local feature selection. The first one is how to recover the depth information. Although some low-cost depth sensors, like Kinect, appear in the market, most smart mobile phones are not equipped with a depth sensor. It is essential to figure out a low-complexity approach to inferring depth cue of query images without any physical depth sensor. The second one is how to properly incorporate depth information into local feature selection. With inaccurate focal length and/or unknown camera movement, the direct disparity comparison between images is infeasible. A measurable and comparable depth measure across all images is needed. What's more, the depth measure may not be conditional independent with the scale given a feature match, which would break the conditional independent assumption of different characteristics in lightweight feature selection [6].

In this paper, we propose a novel depth-based local feature se-

978-1-4673-9961-6/16/$31.00 ©2016 IEEE

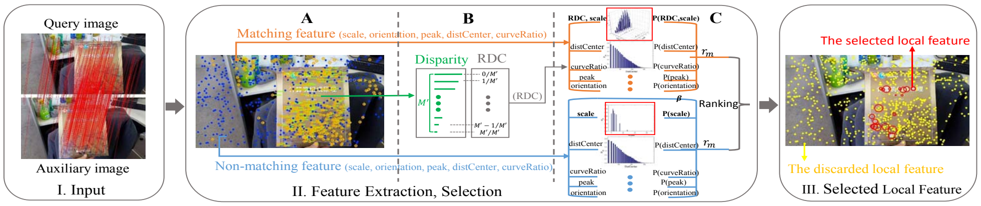

Fig. 2. Overview of our approach. (I) The input includes a query images and an auxiliary image. (II) The local features are extracted and matched between the input images. A, the disparity is recovered; B, the RDC is calculated; C, r_m are calculated for matching descriptors and non-matching descriptor separately. (III) The local features with higher r_m (marked by a larger red circle) are selected.

lection for mobile search. The proposed approach can yield more discriminative local feature set to generate compact descriptors even for query images with foreground distractor or background clutter. The contributions of this paper are threefold. Firstly, we introduce the depth information into local feature selection, and adopt a low-complexity approach to recover depth cue of local features. Secondly, we come up with a novel rank-based depth measure which is comparable across all images. Furthermore, we propose an effective relevance measure to select local features based on the correlation analysis between depth cue and scale. Thirdly, we built up a benchmark of indoors object images (sampled from mobile videos), which have validated the effectiveness of lightweight depth estimation in selecting features and the significant impact on retrieval performance, especially at small descriptor sizes.

2. LOCAL FEATURE SELECTION

2.1. Problem Fomulation

Let f_m denote a local feature. A relevance measure r_m is assigned to each local feature, indicating the probability of being matched correctly. The objective of local feature selection is to maximize the sum of relevance measure of selected features:

$$\max_{h_1,h_2,\ldots h_M} \sum_{m=1}^{M} h_m r_m \quad s.t. \quad \sum_{m=1}^{M} h_m = L. \tag{1}$$

where M is the number of local features in an image; L is the expected number of selected features; $h_m = 1$ for selected, 0 for discarded.

Eq (1) can be solved by ranking and selecting the features by relevance values. The key issue is to figure out the relevance measure of each local feature by modeling the probability of successful matching with regard to different intrinsic characteristics.

2.2. Local Feature Selection using Intrinsic Characteristics

Let $c_m = (c_{1,m}, \ldots, c_{i,m}, \ldots, c_{\mathbb{I},m})$ denote intrinsic characteristics c_i of local feature f_m. Typically, a local feature selection approach attempts to learn r_m statistically on five characteristics of interest points: the scale of the interest point s, the orientation o, the peak response value of the LoG p, the distance from the interest point to the image center dc, and the ratio of the squared trace to the determinant of the Hessian cr. Furthermore, [14] adds a self-matching score characteristic to increase the discriminative power.

It is worthy to note that, in [6], by assuming that different characteristics are conditionally independent given a feature match, r_m can be simplified as

$$r_m = \prod_{i=1}^{\mathbb{I}} P_{c_i}(c_{i,m}). \tag{2}$$

where $P_{c_i}(\cdot)$ is the conditional probability of being matched correctly for a local feature given the value of the characteristics c_i.

In practice, at the training stage, correctly matching local features are selected by pairwise image matching including SIFT matching, ratio test, and geometric consistency check [15]. Each characteristic $c_{i,m}$ of local feature f_m is quantized to k_i bins $\{c_i^1, c_i^2, \ldots, c_i^{k_i}\}$ by K-Means. The conditional probability of being matched correctly $P_{c_i^j}$ for local features in bin c_i^j is estimated by Maximum Likelihood method. In the test stage, for each local feature, $P_{c_i}(c_{i,m}) = \{P_{c_i^j} | c_{i,m} \in c_i^j\}$. r_m is calculated by Eq (2).

3. DEPTH BASED LOCAL FEATURE SELECTION

As illustrated in Fig 2, the depth based local feature selection consists of: (1) recovering the disparity of local features in a query image though an auxiliary image captured from a different viewpoint, (2) generating the relative depth characteristics from the disparity, and (3) modeling the relevancy measure by combining depth and other intrinsic characteristics of interest points.

3.1. Recovering the Disparity of Local Features

There are different ways of recovering 3D structure from multi-view images, such as SFM [16], SLAM [17] and stereo matching [13]. But their main problem is with heavy computational complexity. Unlike stereo matching, the depth information is needed for local features only, rather than the pixel-wise dense depth map. Thus, we adopt a low-complexity approach to recover depth information of local features in query image I^Q though an auxiliary image I^A captured from a different viewpoint.

Firstly, we perform local features matching between I^Q and I^A. Then, though image rectification [18] by well matched local features, each pair of matching features have the same y-coordinate. So, the disparity $disp_m$ of matching local features (f_m^Q, f_m^A) is

$$disp_m = d_x(f_m^Q, f_m^A). \tag{3}$$

where $d_x(\cdot,\cdot)$ is the Euclidean distance measure of matching local features in x-direction of the rectified images. Note that the disparity is calculated only for matching local features between I^Q, I^A.

3.2. Relative Depth Characteristic

The relationship between the disparity $disp_m$ and the depth dep_m of local feature pair (f_m^Q, f_m^A) [19] is

$$dep_m = \frac{F \cdot B}{disp_m} \tag{4}$$

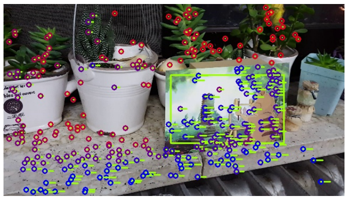

Fig. 3. The circles show local features, the green arrow line show the disparity of each local feature. The color of circles ranges from blue to red with the depth increase of local features. The target object is marked by a green bounding box.

where F is the focal length of camera, B is the baseline of I^Q, I^A, namely the length of the line connecting camera centers of I^Q, I^A.

With unknown or inaccurate focal length and/or baseline, the direct depth comparison between images is infeasible. The disparity normalization is then needed. As the depth is inversely proportional to the disparity and the depth range of different images may vary greatly, it is difficult to apply a uniform normalization rule. In this work, we propose a new rank-based normalization method to describe the relative depth characteristics (RDC). Suppose that the matching local features in a query image are sorted in a descending order by the disparity, and the rank order of local feature f_m^Q with $disp_m$ is denoted by $rank_m$. The RDC $c_{rdc,m}^Q$ of f_m^Q is

$$c_{rdc,m}^Q = \frac{rank_m}{M'}, \tag{5}$$

where M' is the total number of matching local features between I^Q, I^A.

As RDC measures the rank of local features falling into a certain depth range, the comparison of rank order cross different images is meaningful even if their depth ranges vary dramatically.

3.3. The Combo Relevance Measure

To measure the final relevance, a naive way is to combine RDC with all intrinsic characteristics, say six multipliers in Eq (2). However, RDC may not be conditionally independent with other characteristics. As the scale is more influenced by the original size as well as the depth of an object, we simply assume that scale is the only attribute that is high probably correlated with the RDC (i.e., not conditionally independent).

We perform the correlation analysis using chi-square test [20] in the training set to confirm the dependence relationship.

H_0: the scale and the RDC is independent given a feature match;

H_1: the scale and the RDC is dependent given a feature match.

Suppose that the attribute values of scale and RDC are quantized into I and J bins respectively, and the capacity of "match label" of a local feature (matched or not) is K. Thus the degree of freedom is $K(I-1)(J-1)$.

At the α level of significance, H_0 is injected if

$$\chi^2 = \sum_{k=1}^{K}\sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{i,j,k} - \frac{\sum_{j=1}^{J} n_{i,j,k} \sum_{i=1}^{I} n_{i,j,k}}{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{i,j,k}}\right)^2}{\frac{\sum_{j=1}^{J} n_{i,j,k} \sum_{i=1}^{I} n_{i,j,k}}{\sum_{i=1}^{I}\sum_{j=1}^{J} n_{i,j,k}}} > \chi_\alpha^2(K(I-1)(J-1)), \tag{6}$$

where $n_{i,j,k}$ is the number of local features that fall into the i^{th} bin of scale, the j^{th} bin of RDC for matched ($k=1$) or not ($k=2$).

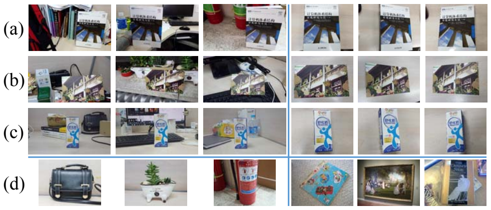

Fig. 4. Example images of the object dataset. Rows (a)∼(c): the left 3 columns are query images, the right 3 columns are clean reference images; Row (d): the left 3 images are distractor objects, the right 3 images are the distractors from CDVS benchmark.

For $\alpha = 0.05$, we empirically set $I = 15, J = 8, K = 2$, and count $n_{i,j,k}$ of all images in the training set of our self-built dataset. We calculate the statistics $\chi^2 = 9051.73 > \chi_{0.05}^2(196) = 229.66$. So H_0 is injected, and the two attributes of scale and RDC are conditional dependent with each other given a feature match.

Therefore, for a matching local feature f_m^Q, rather than simple multiplying, the combo relevance measure r_m^Q is updated by the joint probability of the scale and the RDC:

$$r_m^Q = P_{c_{rdc,s}}(c_{rdc,m}^Q, c_{s,m}^Q) P_{c_o}(c_{o,m}^Q) P_{c_p}(c_{p,m}^Q) P_{c_{dc}}(c_{dc,m}^Q) P_{c_{cr}}(c_{cr,m}^Q) \tag{7}$$

; otherwise, for a non-matching local feature, the RDC is unavailable, and the combo relevance measure is calculated by five intrinsic characteristics and a regularization parameter $\beta \in [0,1]$:

$$r_m^Q = \beta P_{c_s}(c_{s,m}^Q) P_{c_o}(c_{o,m}^Q) P_{c_p}(c_{p,m}^Q) P_{c_{dc}}(c_{dc,m}^Q) P_{c_{cr}}(c_{cr,m}^Q). \tag{8}$$

Over the selected local features in query images, we directly apply CDVS pipeline to generate compact descriptors by local features aggregation and compression [2]. In the subsequent performance evaluation, the state-of-the-art CDVS descriptor is employed to test the impact of depth based feature selection on the retrieval performance.

4. EXPERIMENTS

4.1. Dataset

Since the public visual search benchmarks such as INRIA Holidays [21] and Stanford MVS [22] do not provide auxiliary images to derive RDC, we built up a dataset of indoors object images with a mobile camera (Sumsung S6). To make the dataset more challenging, we capture target objects (books, common objects, postcards) in cluttered scenes. Sample images are shown in Fig 4. The dataset is available at http://pan.baidu.com/s/1gdSoYOj.

Query set. For each target object, 5∼6 videos with average duration of 10s are taken in different cluttered scenes. Key frames are selected every 50 frames to form the query image set. For each query image, the RDC is recovered based on the subsequent selected frame as the auxiliary image for depth estimation (e.g., the $50i+50^{th}$ frame acts the auxiliary image of the the $50i^{th}$ frame, $i = 0, 1...$). There will be not enough matched features between query and auxiliary for large time interval, so we empirically take 50. In total, we select 1009 frames to form the query image set for retrieval test.

Reference set. The number of reference images is 108. For each target object, 3 clean reference images are captured. To study the performance, we merge reference images with about 1 million distractor images. Firstly, we collect 108 distractor images from 36

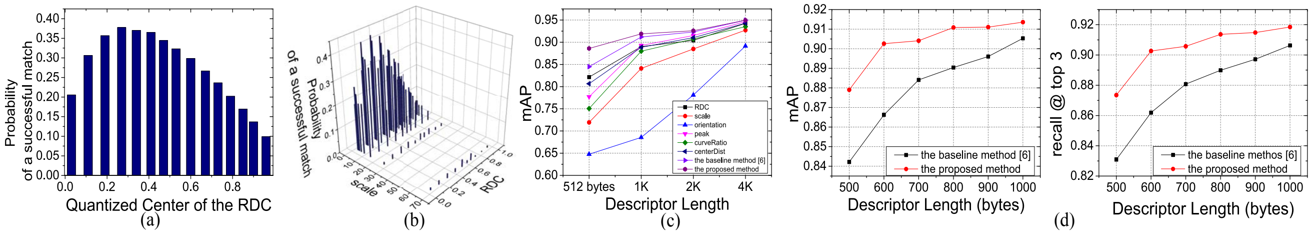

Fig. 5. (a) The statistical model of the RDC; (b) The joint statistical model of RDC and scale; (c) Retrieval performance comparison at 4 pre-defined CDVS descriptor lengths; (d) Retrieval performance comparison at very low bit rates.

distractor objects in the cluttered scenes of query images. Secondly, the Common Object (exactly the UKBench dataset [23]), Graphics and Paintings dataset in MPEG CDVS Benchmark [24] and Flicker1M dataset [21] are added as distractor images.

The retrieval performance is measured by mAP and recall @ top 3 since there are 3 reference images for each query image.

4.2. Experiment Setup

In order to learn the statistical models of characteristics, the dataset is divided into training set and test set. The training set contains 461 query images, while the test set includes 548 query images. The statistical model of 5 characteristics applied in CDVS [6] and the RDC, are learnt on the training set.

After selecting the salient local features, the CDVS compact descriptor of each query image is produced [2]. The subsequent retrieval process is performed in the framework of CDVS as well.

In order to show the impact of feature selection at different data sizes of descriptor, the experiments are performed at 4 pre-defined descriptor lengths in CDVS: 512 bytes, 1K, 2K and 4K. Moreover, we set 6 short descriptor lengths of 500∼1000 bytes to validate the effects of the proposed feature selection method at very low bit rates.

4.3. Experiment Results

Probability model of RDC. In this part, we show the trained statistical model of the RDC. The independent statistical model of RDC is shown in Fig 5(a), while the joint statistical model of RDC and scale is shown in Fig 5(b). The joint statistical model of RDC and scale is adopted in this work to measure the relevance of matching local features. The consistent distributions in Fig 5(a,b) may justify the role of RDC characteristic in feature selection. In practice, The middle range of RDC (from 0.2 to 0.5) usually corresponds to the target object, within which a higher matching probability results. The smaller RDC values are usually caused by noisy foreground objects (like desk texture, hands, etc.), while the bigger RDC values are often caused by background scenes. Note that for smaller or bigger RDC values, the probability of good matching is much lower.

Effect of parameters. The parameters in the proposed method include the bin size of each attribute and the regularization parameter β. The bin sizes of s, o, p, dc, cr are set to 8,32,16,32,16 as in CDVS [2]. Table 1 illustrates the effects of bin sizes of the RDC and β (performed without Flicker1M). With a larger bin size, fine quantization will produce slightly better retrieval performance. β is a regulation item for all non-matching local features. Without depth information, the non-matching local features do not get the joint probability of scale and RDC. So β is added into the model so that the non-matching features can be ranked together with matching local features. If $\beta = 0$, r_m for a non-matching local feature is 0, which means the selected local features are all matching features. $\beta = 1$ means that the r_m of non-matching local features is determined by 5 characteristics, while 6 for matching local features. From Table 1, the highest retrieval performance is with moderate $\beta = 0.3$.

mAP(%) bin size β	10	15	20	25	30
1.0	89.31	89.97	89.70	89.69	89.79
0.8	90.24	89.91	89.90	90.27	90.22
0.6	90.68	90.68	90.81	90.85	91.04
0.4	90.80	90.87	90.93	90.20	91.04
0.3	91.09	91.00	91.05	91.34	**91.43**
0.2	91.00	90.84	90.81	91.08	91.24
0.0	90.22	89.81	89.87	90.10	90.39

Table 1. Effects of parameters RDC bin size and β @ 512 bytes.

Retrieval performance comparison. The retrieval performance is compared over a variety of feature selection settings: using only one characteristic (RDC, scale, orientation, peak, curveRatio, distCenter), the baseline method [6] and the proposed combo feature selection method. The RDC is proved to be one of the most effective characteristics at 4 pre-defined descriptor lengths in CDVS (as seen in Fig 5(c)). At low bit rates, the smaller the descriptor length is, the more gain can be obtained by incorporating RDC (as seen in Fig 5(d)). The reason is, the proposed approach may generate better ranking of local features than [6] by employing the depth strengthened relevance measure, so that those high quality features of target objects can still be selected out even when a smaller number of local features are allowed. At the lowest 512 bytes set up in CDVS, the mAP of the proposed method is 88.6% while the mAP of [6] is 84.5%.

5. CONCLUSION

By incorporating depth cue into the local feature selection, the proposed method can yield more discriminative local feature set for query images in cluttered scenes and significantly improve the retrieval performance. The proper use of depth cue is promising in improving image retrieval and recognition performance in mobile search scenarios.

6. ACKNOWLEDGEMENT

This work was supported by the National Hightech R&D Program of China (863 Program):2015AA016302, and Chinese Natural Science Foundation: 61271311,61390515,61421062.

7. REFERENCES

[1] B. Girod, V. Chandrasekhar, D. Chen, N. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, and R. Vedantham, “Mobile visual search,” *IEEE Signal Processing Magazine*, vol. 28, pp. 61–76, 2011.

[2] L. Duan, V. Chandrasekhar, J. Chen, J. Lin, Z. Wang, T. Huang, B. Girod, and W. Gao, “Overview of the mpeg-cdvs standard,” *IEEE Transactions on Image Processing (TIP)*, vol. 25, no. 1, pp. 179–194, 2016.

[3] S. Paschalakis et al., “Information technology - multimedia content descriptor interface - part 13: Compact descriptors for visual search,” *ISO/IEC 15938-13:2015*.

[4] J. Lin, L. Duan, Y. Huang, S. Luo, T. Huang, and W. Gao, “Rate-adaptive compact fisher codes for mobile visual search,” *IEEE Signal Processing Letters (SPL)*, vol. 21, no. 2, pp. 195–198, 2014.

[5] S. Paschalakis, K. Wnukowicz, M. Bober, et al., “Cdvs ce2: Local descriptor compression proposal,” *M25929, ISO/IEC JTC1/SC29/WG11, Sweden: Jul*, 2012.

[6] G. Francini, S. Lepsøy, and M. Balestri, “Selection of local features for visual search,” *Signal Processing: Image Communication*, vol. 28, no. 4, pp. 311–322, 2013.

[7] J. Lin, L. Duan, T. Huang, and W. Gao, “Robust fisher codes for large scale image retrieval,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2013, pp. 1513–1517.

[8] J. Knopp, J. Sivic, and T. Pajdla, “Avoiding confusing features in place recognition,” in *European Conference on Computer Vision (ECCV)*, pp. 748–761. Springer, 2010.

[9] G. Tolias and H. Jégou, “Visual query expansion with or without geometry: refining local descriptors by feature aggregation,” *Pattern Recognition*, vol. 47, no. 10, pp. 3466–3476, 2014.

[10] P. Turcot and D.G. Lowe, “Better matching with fewer features: The selection of useful features in large database recognition problems,” in *IEEE International Conference on Computer Vision Workshop (ICCV Workshop)*, Sept 2009, pp. 2109–2116.

[11] W. Hartmann, M. Havlena, and K. Schindler, “Predicting matchability,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2014, pp. 9–16.

[12] K.H. Jin, E. Dunn, and J. Frahm, “Predicting good features for image geo-localization using per-bundle vlad,” in *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1170–1178.

[13] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *International Journal of Computer Vision (IJCV)*, vol. 47, no. 1-3, pp. 7–42, 2002.

[14] X. Xin, Z. Li, Z. Ma, and A.K. Katsaggelos, “Robust feature selection with self-matching score,” in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 2013, pp. 4363–4366.

[15] S. Lepsøy, G. Francini, G. Cordara, D. Gusmão, and P.P. Buarque, “Statistical modelling of outliers for fast visual search,” in *IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2011, pp. 1–6.

[16] S. Agarwal, N. Snavely, I. Simon, S.M. Seitz, and R. Szeliski, “Building rome in a day,” in *IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2009, pp. 72–79.

[17] R. Mur-Artal, J.M.M. Montiel, and J.D. Tardos, “Orb-slam: A versatile and accurate monocular slam system,” *IEEE Transactions on Robotics (TRO)*, vol. 31, no. 5, pp. 1147–1163, Oct 2015.

[18] R.I. Hartley, “Theory and practice of projective rectification,” *International Journal of Computer Vision (IJCV)*, vol. 35, no. 2, pp. 115–127, 1999.

[19] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[20] A. Agresti and M. Kateri, *Categorical data analysis*, Springer, 2011.

[21] H. Jegou, M. Douze, and C. Schmid, “Hamming embedding and weak geometric consistency for large scale image search,” in *European Conference on Computer Vision (ECCV)*, pp. 304–317. Springer, 2008.

[22] V. Chandrasekhar, D. Chen, S. Tsai, N. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod, “The stanford mobile visual search data set,” in *ACM Multimedia Systems Conference (MMSys)*. 2011, MMSys ’11, pp. 117–122, ACM.

[23] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2006, vol. 2, pp. 2161–2168.

[24] “Mpeg cdvs (compact descriptors for visual search) dataset,” `http://pacific.tilab.com/Dataset-20120210/`.