DOI: 10.15439/2015F85
ACSIS, Vol. 5

# Visual Detection of Objects by Mobile Agents using CBVIR Techniques of Low Complexity

Andrzej Śluzek

Khalifa University, Abu Dhabi Campus
P.O. Box 127788, Abu Dhabi, UAE
Email: andrzej.sluzek@kustar.ac.ae

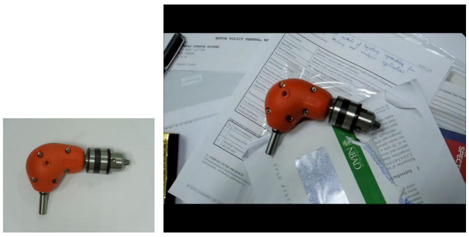
***Abstract*—Visual search for objects of interest in complex environment is an important (and still challenging) problem in mobile robotics. In particular, the usage of *content-based visual information retrieval* (CBVIR) methods, which are a natural choice for such tasks, is often constrained by the real-time requirements, and the mobility of searching agents is sometimes not sufficiently exploited in the search model. In this paper, a CBVIR-based scheme is proposed, which takes into account motion of the searching agents to achieve a low-cost and high-speed detection of objects of interest in cluttered scenes, with good overall performances. We combine standard CBVIR tools, i.e. MSER detector and SIFT descriptor (quantized into sufficiently large vocabularies) assuming additionally that objects become *objects of interest* only when approached closely enough by the mobile agent, i.e. when seen at an adequately large scale. Thus, an object of interest is considered detected only if a sufficient number of keypoints from the current video-frame are matched (including the corresponding matches of scales) to the keypoints from the database images of the object. Preliminary experiments on a limited-size dataset confirm performances of the scheme, although in the classical task of video-frame retrieval the scheme cannot compete with more sophisticated CBVIR algorithms. The scheme can prospectively become more flexible if combined with a range-finding device so that the approximate distances to the scene components within the currently inspected part of the image can be used to proportionally modify the scale correspondences.**

## I. Introduction and Background Works

Visual search for unpredictably located objects of interest (where such object are represented by their exemplary images) remains one of important tasks for mobile robotics. Understandably, CBVIR is a natural source of algorithms for such a task. Actually, one of the fundamental CBVIR concepts of *visual words* was originally proposed for videos [1]. Algorithms for sub-image retrieval (where the objective is to identify images containing fragments near-duplicate to the query image) are particularly important. Those algorithms are usually based on detecting keypoints and, subsequently, comparing their visual words. Generally, however, three major differences exist between real-time visual search by mobile agents and classical CBVIR tasks:

1) In CBVIR, infrequently arriving query images (submitted by the users) are matched against large/huge datasets, e.g. [2], [3], while in a camera-based visual search very small datasets (template images of objects of interest) are matched against large numbers of continuously arriving queries (video-frames acquired by a single camera or by simultaneously working multiple cameras attached to a mobile agent).
2) Because of (1), the computational costs of image preprocessing (i.e. keypoint detection and description, quantization into visual words, etc.) are in visual search as critical as the complexity of the actual image matching and retrieval. In classical CBVIR, the costs of image preprocessing are considered negligible.
3) In visual search, the objective is to detect all instances of interesting objects, where each instance is represented by a sequence of frames in the input video stream. However, not all such frames have to be perfectly identified. Thus, in the video search (unlike in standard CBVIR tasks) *recall* of individual frame retrieval can be compromised, but *precision* should be as high as possible.

As an illustrative example, Fig. 1 shows an object of interest and an exemplary video-frame returned by the search algorithm.



(a) (b)

Fig. 1. The object of interest (a) and an exemplary frame containing it (b).

In this paper, we propose and preliminarily evaluate a scheme which exploits the above characteristics of visual search (and mobility of the searching agent) to achieve at low costs a high performance object detection in cluttered scenes. As the basic components of this scheme, we use standard CBVIR tools, i.e. MSER detector [4] and SIFT descriptor [5] (in its RootSIFT variant [6]). Descriptors are quantized into a large vocabulary of one million words (to assure satisfactory *precision*). To avoid high costs of descriptor quantization into words, a simple quantization method based on the statistical properties of descriptors is used. Details of the image preprocessing phase are described in Section II.

978-83-60810-66-8/$25.00©2015, IEEE

Image matching is performed using the most straightforward criterion, i.e. the number of keypoint correspondences (e.g. [7], [5]) where a match is defined by identical visual words. However, we reject matches for which the ellipses of MSER keypoints are not in the correspondingly similar scales. Because the search is conducted by mobile agents, this requirement indicates that the agent finds an object at a specific distance (defined by the scale of the object images in the database). More details and explanations are given in Section III.

In Section IV, preliminary experimental results of the proposed approach are overviewed. In particular, performances of the method are compared to alternative solutions.

## II. Principles of Image Pre-processing

For a feature-based real-time visual search, efficient detection and description of keypoints is a critical factor. In particular, the number of keypoints should not be excessively large (and controllable in some sense). Thus, we use MSER detector which is affine invariant, has good performances (as reported in [8]), low complexity (which can be further reduced by using special techniques proposed for MSER detection in video sequences, e.g. [9]) and a few tuning parameters to control the numbers of detectable keypoints. Actually, recent implementations of MSER detectors in hardware, e.g. [10], [11], achieve a throughput approaching hundreds of frames per second, which indicates that MSER keypoint extraction is not a critical factor in a real-time image pre-processing, even if several cameras are simultaneously used.

Hardware implementations of SIFT descriptors (and detectors too) have been reported as well (most recently in [12]). Notably, development of a system-on-chip SIFT descriptor for affine-invariant keypoints is currently under way in our organization as well. Therefore, the feasibility of real-time SIFT description of detected keypoints can be considered documented. Even the Matlab implementation of MSER detector combined with a SIFT description module in C++ provide a throughput of 2-3 frames/sec (including all disk read/write overheads) which can be accepted as near real-time performances for slowly moving agents.

The final step of image pre-processing for CBVIR is *quantization* of descriptors into *visual words*. The size of vocabularies can be very diversified, typically ranging from a few thousand to a few million. Although small vocabularies provide better *recall* of keypoint matching, *precision* is generally unacceptably low. *Precision* obviously improves (at the expense of *recall*) with the growing size of vocabulary, but if the vocabulary becomes too large, the quantization intervals could be smaller that natural fluctuations of descriptors, and it might be difficult to find matches even in pairs of almost identical images.

Published results (e.g. [13], [3]) indicate that the recommended sizes of visual vocabularies are in the range of millions of words, especially if *precision* is more important than *recall* (which is the case in visual search by mobile agents, see Point (3) in Section I). Thus, RootSIFT descriptors are quantized into a vocabulary of 1M words (although tests have been conducted using smaller sizes as well - see Section IV). For such a large vocabulary, the standard quantization of descriptors into words by the (approximate) nearest neighbour approach could be a bottleneck in real-time processing of video frames. Instead, the descriptor space has been partitioned off-line into hypercubes of similar probabilities (the probability density was estimated using over 500 million keypoints from diversified images). Then, the descriptor quantization requires only a small number of additions and comparisons, and the processing time is negligibly small.

## III. Image Matching and Object Detection

In visual search, the objective is to identify video fragments (sequences of frames) containing the object(s) of interest, regardless the background visual contents. From CBVIR perspective, this is a problem of *partial near-duplicate* detection, for which a fully satisfactory solution has not been found yet. Nevertheless, most of the *state-of-the-art* methods seem to follow the same two-step approach. First, similarities between individual keypoints are established (using descriptors or visual words). Then, the geometric consistencies between groups of preliminarily matched keypoints are verified to detect clusters of similarly transformed keypoints (which are considered the near-duplicate fragments). Either more advanced algorithms, like the Hough transform, RANSAC, etc. are used (e.g. [14], [15], [16], [17]) to provide more credible results at higher computational costs, or simplified approaches (e.g. [18], [2], [3]) more suitable for large-scale applications are alternatively employed to verify the consistencies.

In the proposed scheme, we detect partial near-duplicates (presumably representing the objects of interest) in a way that merges the first step with a very simple variant of the geometric verification (where only the scale consistency of matching keypoints is verified). Altogether, the level of similarity between a query image (i.e. a video frame) and a database image of an object is defined by the number of keypoint correspondences, where two MSER keypoints $K_1$ (a query keypoint) and $K_2$ (a database keypoint) match if:

Definition 1.

1) The keypoints are described by the same visual word, i.e. $word(K_1) = word(K_2)$.
2) The keypoints have similar scales. Assuming that $M$ and $m$ indicate, correspondingly, the length of major and minor axes of the keypoint ellipses, the conditions for the scale consistency are:

$$0.8M(K_2) \leq M(K_1) \leq 1.2M(K_2), \tag{1}$$

$$0.8m(K_2) \leq m(K_1) \leq 1.2m(K_2). \tag{2}$$

The second requirement of the above definition can be justified as follows:

> *The visual scale of an object in a captured video obviously corresponds to the distance between this object and the camera. When a mobile agent explores its environment, it is expected to recognize*

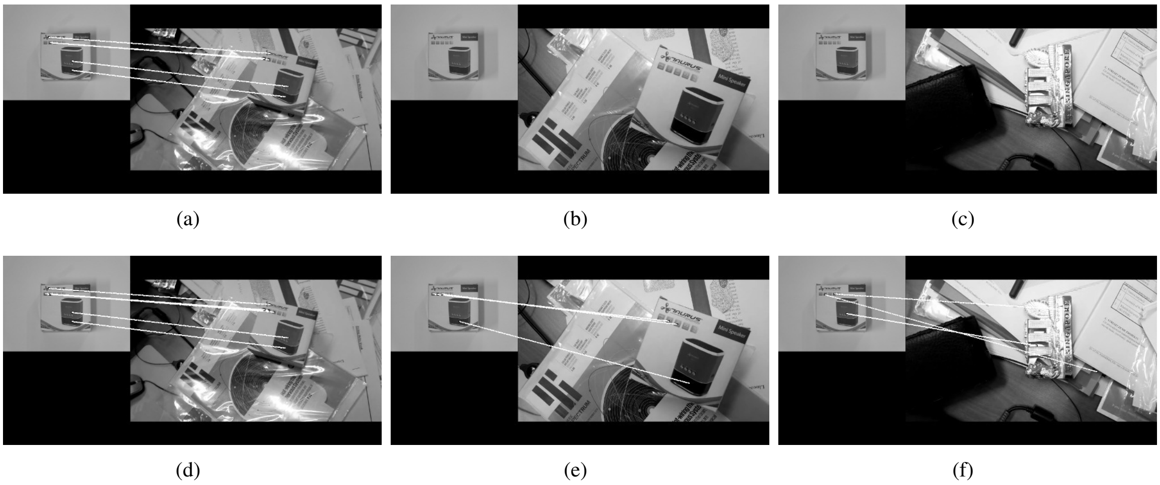


Fig. 2. Exemplary matches obtained by using the scale verification (a, b, c), and without such a verification (d, e, f). A vocabulary of 1M words is applied. In (b) the object is not detected because its scale is too large (but it is detected in (e) where the scale is not verified). In (c), scale verification prevents detection of a non-existing object (falsely detected in (f)).

> *objects of interest when they are approached at a sufficiently close distance, i.e. at least at a predefined threshold distance. Thus, the database should contain images of objects of interest in the* ***reference scales*** *approximately corresponding to such threshold distances. The images in larger scales are not needed because the objects should be detected earlier (at the threshold distance) while smaller scales represent objects too distant to be interesting for the agent. Therefore images in larger or smaller scales are not included in the database.*

The axes length tolerance in Eqs 1 and 2 is rather wide (and taken independently for major and minor axes) so that not only small scale deviations but also minor viewpoint changes (up to approx. $30^o$) are generally accepted by the matching algorithm.

A similar philosophy (although with much less efficient tools for keypoint detection and matching) was behind the results presented in [19].

Eventually, two images are considered partial near-duplicates (i.e. a part of the query frame matches the object of interest) if at least 4 pairs of keypoint correspondences are found according to Definition 1. This is the minimum number of matches needed for the verification of affine transformation consistency between images (three pairs to build the transformation, and the fourth one to verify it). Although currently only the scale consistency is applied, such a geometric verification might be used in the future for more advanced tasks (e.g. in determining the number of the same objects of interest in a single video frame).

This matching method is sufficiently fast for visual search tasks considered in this paper. If the images are pre-processed (i.e. MSER keypoints are extracted and assigned visual words, which are very fast operations as outlined in Section II) even the Matlab implementation provides a throughput of approx. 50-60 video frames of VGA resolution per second (i.e. the search can be conducted using 2-3 simultaneously working cameras).

Examples in Fig. 2 highlight the principles and specific characteristics of object detection using the proposed image matching technique.

## IV. Experimental verification

The proposed scheme has been preliminarily verified on a number of short (i.e. $30 - 60$ seconds) videos captured in heavily cluttered indoor environments. A small collection of objects of interest has be arbitrarily proposed (see Fig. 3).

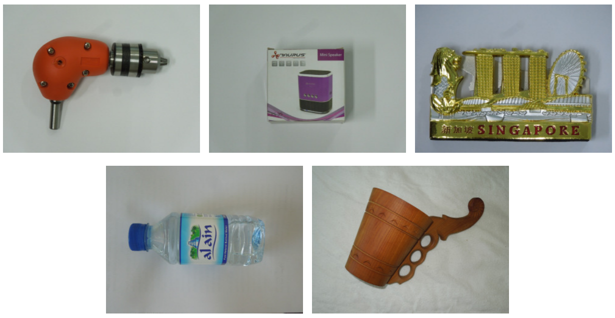


Fig. 3. Examples of objects of interest.

The objective is to identify all instances (but not necessarily all frames of the video) of the objects which are seen for some time at the reference or larger scale (i.e. more distant appearances of the objects are not counted). Fig. 4 shows a few frames from an exemplary beginning (when the object becomes sufficiently large) and from an exemplary end (when the object becomes too small and/or disappears from the field of view) of such instances. In all cases, the ground truth data, i.e. the initial and terminal frames of the instances, are established manually.

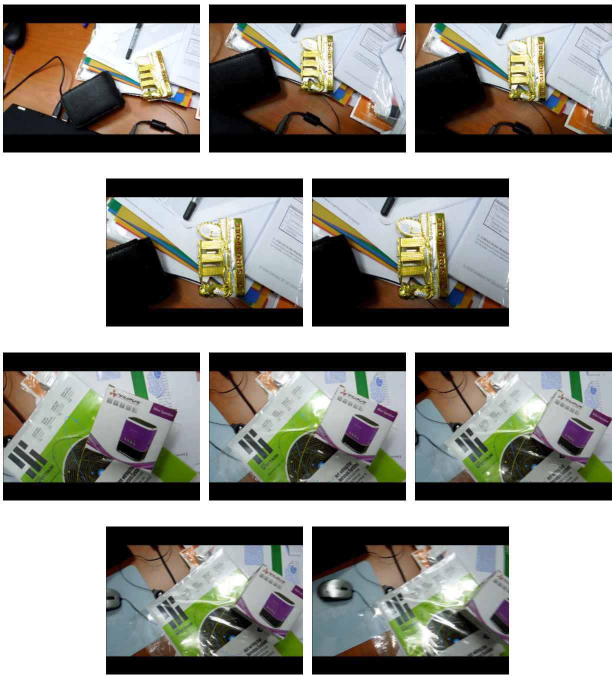


Fig. 4. Examples of frames from a typical initial part (two top rows) and a typical terminal part (two bottom rows) of a ground-truth instance of an object of interest.

The scheme's performances are evaluated by comparing *ground-truth instances* and so-called *active sequences* extracted by the scheme.

Definition 2.

> An *active sequence* is initiated whenever the algorithm identifies a frame matching (according to the specification in Section III) a database image of an object (e.g., Fig. 2A). Then, the active sequence continues until there are at least five consecutive frames which do no match the same object database images.

The value 5 has been established empirically; it corresponds to approx. $0.2$sec during which the object may be temporarily invisible (due to sudden flashes of light, temporary camera defocusing, etc.). However, when an active sequence is terminated, it does not mean the object is not visible anymore. Actually, the following cases are possible:

- The object is too close to the camera so that its scale is too large for a match with database images.
- The object becomes too distant (its scale is too small for a match) which means it is no more an object of interest.
- The object actually disappears from the field of view.

Regardless the reason for which an *active sequence* is terminated, the following requirements define a fully reliable object detection scheme:

(a) Each *active sequence* is fully enclosed within a *ground-truth instance*, i.e. non-existing objects are never detected.
(b) Each *ground-truth instance* overlaps at least one *active sequence*, i.e. each genuine instance of an object is detected at least by a single active sequence.

Using the above specifications, we straightforwardly define *precision* (*PA*) of active sequence extraction and *recall* (*RI*) of ground-truth instance detection in a visual search process by a mobile agent as follows:

$$PA = \frac{AS_{(a)}}{AS}, \tag{3}$$

where $AS_{(a)}$ is the number of active sequences satisfying the above Requirement (a), and $AS$ is the total number of extracted active sequences.

$$RI = \frac{GTI_{(b)}}{GTI}, \tag{4}$$

where $GTI_{(b)}$ is the number of ground-truth instances satisfying the above Requirement (b), and $GTI$ is the total number of ground-truth instances.

It was mentioned in Section I that in object detection by visual search *recall* of the individual frame retrieval can be compromised, but *precision* should be as high as possible. However, both *RI recall* and *PA precision* values should be at the highest possible levels to reliably detect object instances.

Performances (based on Eqs 3 and 4) of the scheme for the test dataset of videos and objects are summarized in the top row of Table I. To illustrate advantages of the proposed scheme over the other choices, we include in Table I the results for three alternative scenarios. First, the same vocabulary of 1M words is used but without the scale verification (the second row of Table I). Secondly, a much smaller vocabulary of 64k words is used (with the scale verification) instead of the proposed 1M vocabulary (the third row of Table I). The last scenario included in Table I will be discussed later.

Although each of the three schemes detects all instances of objects visible within the test dataset of videos, there are significant differences in the numbers of extracted active sequences, and (consequently) in the reliability of detection. When scale verification is ignored, or the size of vocabulary is significantly reduced, the number of active sequences grows disproportionally ($5 - 6$ times in our experiments) and *PA precision* falls dramatically. The explanations are similar for both cases. On one hand, credibility of individual keypoint correspondences is limited (even for larger vocabularies) if no means of geometric verification are used. On the other hand, if the vocabulary is small, the number of keypoint correspondences can be so large that even the scale verification is unable to delete all false positives. As a result, large numbers of false active sequences are extracted from the incoming stream of frames. Even though most of those incorrect active sequences are short ($1 - 2$ frames) they should not be ignored because there are some cases when the ground-truth instances are represented only by such short active sequences.

Fig. 5 shows examples of incorrect matches (some parts of Fig. 2 are also illustrative) including a rather unusual (since *PA precision* is equal to $98.8\%$) case of a false positive for matching with 1M words and scale verification.

It should be emphasized that high performances of the proposed low-complexity scheme are achievable for object

TABLE I
PERFORMANCES OF OBJECT DETECTION USING THE PROPOSED SCHEME AND THREE ALTERNATIVE SCHEMES.

| **Scheme** | Ground-truth instances | Active sequences | *RI recall* (Eq. 4) | *PA precision* (Eq. 3) |
|---|---|---|---|---|
| **1M words with scale verification** | 58 | 329 | 100.0% | **98.8%** |
| **1M words without scale verification** | 58 | 1886 | 100.0% | 20.6% |
| **64k words with scale verification** | 58 | 1618 | 100.0% | 15.6% |
| **the method from [16]** | 58 | 117 | 60.3% | 100.0% |

TABLE II
PERFORMANCES OF INDIVIDUAL FRAME RETRIEVAL. IF THE SCALE VERIFICATION IS USED, ONLY THE FRAMES CONTAINING THE OBJECT IN APPROX. THE REFERENCE SCALE ARE CONSIDERED THE GROUND TRUTH.

| **Method** | Ground-truth frames | Retrieved frames | *Recall* | *Precision* |
|---|---|---|---|---|
| **1M words with scale verification** | 5429 | 1342 | 22.1% | 89.4% |
| **1M words without scale verification** | 15479 | 19741 | 60.1% | 47.1% |
| **64k words with scale verification** | 5429 | 18078 | 73.3% | 22.0% |
| **the method from [16]** | 15479 | 7514 | 47.9% | 98.6% |

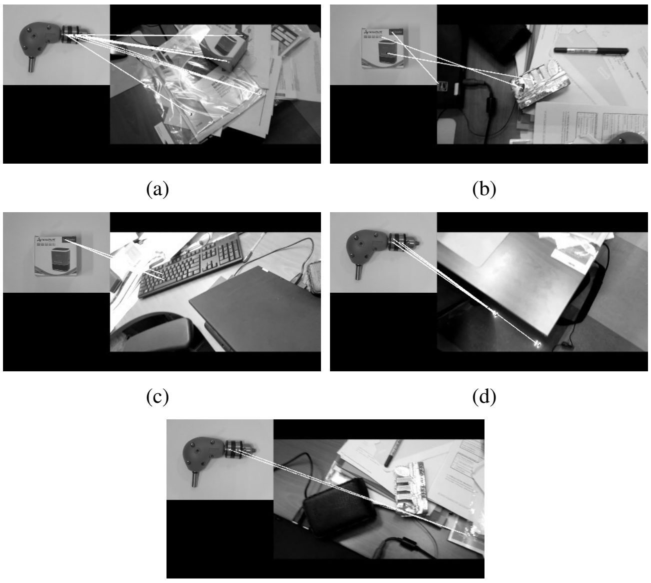


Fig. 5. Examples of false positive matches. In (a,b) 1M words without scale verification are used, while in (c,d) a vocabulary of 64k words is applied with scale verification. A very unusual case of a false positive for 1M words with scale verification is shown in (e).

detection task only. For a classical CBVIR problem of relevant frame retrieval, i.e. detection of ALL frames partially near-duplicate to the database images of objects of interest, the results are much poorer as shown in Table II presenting performances of various schemes in such a classical *relevant frame retrieval* task (using standard CBVIR definitions of *precision* and *recall* as the scores). Thus, as the last scenario, we included a high-performance (and high-complexity) image matching method proposed in [16] (the original executables of this method have been used). The last row of Table II clearly illustrates superiority of this advanced CBVIR method. The approach proposed in this paper satisfy only the requirement of high *precision*. Therefore, it is not surprising that methods similar to the proposed approach are rather seldom considered for typical CBVIR tasks.

Nevertheless, the algorithm of [16] performs poorer in the problem of object detection. As shown in the last row of Table I, its *RI recall* of instances detection is at unsatisfactory level of only 60%, which is much less than all other schemes presented in this table (which score 100%). Admittedly, it achieves 100% of *PA precision* but (as mentioned earlier) both parameters should be as high as possible in a reliable object detection scheme.

## V. CONCLUDING REMARKS AND FUTURE RECOMMENDATIONS

The paper proposes a CBVIR-based scheme for visual detection of predefined objects of interest in cluttered environments. The scheme seems to be an attractive option for low-cost mobile agents equipped with vision devices. The proposed scheme provides (as preliminarily confirmed by our limited-scale experiments) sufficiently high performances using only low-complexity CBVIR mechanisms (borrowed from a classical CBVIR problem of partial near-duplicate retrieval). Additionally, we assume that the encountered objects become *objects of interest* when seen from a sufficiently short distance. This threshold distance defines the scale at which the database template images of objects should be recorded. Then, a frame containing an object would be recognized as an interesting one if two conditions are met. First, it is similar to a database image of some object of interest (i.e. numerous keypoint correspondences defined by identical visual words from a sufficiently large vocabulary exist) and, secondly, a significant number of those matching keypoint pairs are in approximately the same scale (i.e. the object is approached at approximately the threshold distance).

The second requirement can be considered a limitating factor, especially if the threshold distance (or rather the scale corresponding to this distance) cannot be specified or it fluctuates, e.g. because of the camera zoom. However, in typical modern applications (in robotics in particular) the mobile agents are usually equipped with some kind of range-sensing devices which can provide the agent with the depth data, e.g. [20], [21]. Then, an estimate of the distance to the observed part of the environment can be used to correspondingly modify the reference scale to be applied in the scheme (so that an adaptive reference scale is used). The only modification needed is a minor change in Eqs 1 and 2, which should be rewritten as

$$0.8M\ (K_2) \leq S_F{\cdot}M(K_1) \leq 1.2M(K_2), \tag{5}$$

$$0.8m\ (K_2) \leq S_F{\cdot}m(K_1) \leq 1.2m(K_2), \tag{6}$$

where $S_F$ is the scale adaptation factor which can be estimated using the depth data from a range sensor and/or the current camera focus data.

## References

[1] [1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Conf. ICCV 2003,* vol. 2, Nice, 2003. doi: 10.1109/ICCV.2003.1238663 pp. 1470–1477. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2003.1238663

[2] H. Jegou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision,* vol. 87, no. 3, pp. 316–336, 2010. doi: 10.1007/s11263-009-0285-2. [Online]. Available: http://dx.doi.org/10.1007/s11263-009-0285-2

[3] H. Stew´ enius, S. Gunderson, and J. Pilet, "Size matters: Exhaustive geometric verification for image retrieval," in *Proc. ECCV 2012,* vol. II, Florence, 2012. doi: 10.1007/978-3-642-33709-3 48 pp. 674–687. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-33709-3 48

[4] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," *Image and Vision Computing,* vol. 22, pp. 761–767, 2004. doi: 10.1016/j.imavis.2004.02.006. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2004.02.006

[5] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* vol. 60, no. 2, pp. 91–110, 2004. doi: 10.1023/B:VISI.0000029664.99615.94. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[6] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Proc. IEEE Conf. CVPR 2012,* 2012. doi: 10.1109/CVPR.2012.6248018 pp. 2911–2918. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2012.6248018

[7] Y. Ke, R. Sukthankar, and L. Huston, "An efficient parts-based nearduplicate and sub-image retrieval system," in *Proc. ACM Multimedia Conf.,* 2004. doi: 10.1145/1027527.1027729 pp. 869–876. [Online]. Available: http://dx.doi.org/10.1145/1027527.1027729

[8] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *International Journal of Computer Vision,* vol. 65, pp. 43–72, 2005. doi: 10.1007/s11263-005-3848-x. [Online]. Available: http://dx.doi.org/10.1007/s11263-005-3848-x

[9] M. Donoser and H. Bischof, "Efficient maximally stable extremal region (mser) tracking," in *Proc. IEEE Conf. CVPR 2006,* 2006. doi: 10.1109/CVPR.2006.107 pp. 553–560. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.107

[10] F. Kristensen and W. MacLean, "Real-time extraction of maximally stable extremal regions on an fpga," in Proc. IEEE Symp. ISCAS 2007, 2007. doi: 10.1109/ISCAS.2007.378247 pp. 165–168. [Online]. Available: http://dx.doi.org/10.1109/ISCAS.2007.378247

[11] E. Salahat, H. Saleh, A. Sluzek, M. Al-Qutayri, B. Mohammed, and M. Ismail, "Architecture and method for real-time parallel detection and extraction of maximally stable extremal regions (msers)," *U.S. Patent Application* No. 14/482,629, 2014.

[12] J. Jiang, X. Li, and G. Zhang, "Sift hardware implementation for real-time image feature extraction," *IEEE Trans. Circuits Syst. Video Technol.,* vol. 24, no. 7, pp. 1209–1220, 2014. doi: 10.1109/TCSVT.2014.2302535. [Online]. Available: http://dx.doi.org/10.1109/TCSVT.2014.2302535

[13] D. Nist´ er and H. Stew´ enius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Conf. CVPR 2006,* vol. 2, 2006. doi: 10.1109/CVPR.2006.264 pp. 2161–2168. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.264

[14] O. Chum and J. Matas, "Matching with prosac - progressive sample consensus," in *Proc. IEEE Conf. CVPR 2005,* San Diego(CA), 2005. doi: 10.1109/CVPR.2005.221 pp. 220–226. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2005.221

[15] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Computer Vision,* vol. 2, 1999. doi: 10.1109/ICCV.1999.790410 pp. 1150–1157. [Online]. Available: http://dx.doi.org/10.1109/ICCV.1999.790410

[16] M. Paradowski and A. Śluzek, "Local keypoints and global affine geometry: Triangles and ellipses for image fragment matching," in *Innovations in Intelligent Image Analysis,* H. Kwasnicka and L. Jain, Eds. Springer-Verlag, 2011, vol. SCI339, pp. 195–224. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-17934-1 9

[17] Z. Wu, Q. Ke, M. Isard, and J. Sun, "Bundling features for large scale partial-duplicate web image search," in *Proc. IEEE Conf. CVPR 2009,* Miami Beach, 2009. doi: 10.1109/CVPR.2009.5206566 pp. 25–32. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2009.5206566

[18] O. Chum and J. Matas, "Large-scale discovery of spatially related images," *IEEE PAMI,* vol. 32, no. 2, pp. 371–377, 2010. doi: 10.1109/TPAMI.2009.166. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2009.166

[19] M. Islam and A. Śluzek, "Relative scale method to locate an object in cluttered environment," *Image and Vision Computing,* vol. 26, no. 2, pp. 259–274, 2008. doi: 10.1016/j.imavis.2007.06.001. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2007.06.001

[20] K. Shubina and J. Tsotsos, "Visual search for an object in a 3d environment using a mobile robot," *CVIU,* vol. 114, pp. 535–547, 2010. doi: 10.1016/j.cviu.2009.06.010. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2009.06.010

[21] K. Sjo, D. Lopez, C. Paul, P. Jensfelt, and D. Kragic, "Object search and localization for an indoor mobile robot," *J. Computing and Inf.Technology,* vol. CIT 217, no. 1, pp. 67–80, 2009. doi: 10.2498/cit. [Online]. Available: http://http://hrcak.srce.hr/index.php?show=toc&idbroj=3662