The Robot Vision Challenge. Where are we after 5 editions?

Jesús Martínez-Gómez, Barbara Caputo, Miguel Cazorla, Henrik I. Christensen, Marco Fornoni,
Ismael García-Varea and Andrzej Pronobis

Abstract—This article describes the Robot Vision challenge, a competition that evaluates solutions for the visual place classification problem. Since its origin, this challenge has been proposed as a common benchmark where worldwide proposals are measured using a common overall score. Each new edition of the competition introduced novelties, both for the type of input data and sub-objectives of the challenge. All the techniques used by the participants have been gathered up and published to make it accessible for future developments. The legacy of the Robot Vision challenge includes datasets, benchmarking techniques, and a wide experience in the place classification research that is reflected in this article.

Index Terms—competition, benchmarking, semantic localization, robotics, place classification.

I. INTRODUCTION

THE robot vision challenge started as part of the ImageCLEF lab in 2009. The challenge has since then been organized 5 times. The initial challenge was devoted to image based place categorization, namely how to determine from a single image which room the robot is in. This would be equivalent to localization in a topological map. One of the motivations for this was the significant progress on image categorization. Recent results by 2009 demonstrated [1], [2], [3] that it is possible to recognize both manufactured and natural objects such as cars, cows, offices, etc. in natural images. This motivated the challenge of using place recognition for robot tasks. Initially the task was focused on the use of intensity images for place categorization; participants were asked to process each image individually. Over time the challenge has been expanded to include RGB-D data but also to perform both place and object categorization. Due to environmental variations such as lighting, moving furniture around, etc. a holistic analysis of an image may not generate robust results. A set of objects may point to a particular room category i.e., pots in the kitchen and book shelves in the office. Given progress on object categorization and a need for further robustness, there was a need to provide multi-modal sensor data and a richer challenge in terms of categorization of places and objects.

Jesús Martínez-Gómez and Ismael García-Varea are with the Computing Systems Department, University of Castilla-La Mancha, Spain.

Barbara Caputo is with the Department of Computer, Control and Management Engineering, University of Rome La Sapienza, Italy,

Miguel Cazorla is with the Institute for Computing Research, University of Alicante, Spain

Henrik I. Christensen is with Institute for Robotics and Intelligent Machines at Georgia Institute of Technology, Atlanta, GA, USA

Marco Fornoni is with École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

Andrzej Pronobis is with the University of Washington, USA

In this paper we describe the main motivation for the competition in Section II. The evolution of the challenge and some notable results are discussed in Section III. Over the years we have received suggestions and surprising results from the participants, which are described in Section IV and finally an outlook and conclusions are provided in Section V.

II. WHY TO CREATE A PLACE CATEGORIZATION COMPETITION?

Performing repeatable experiments which produce quantitative, comparable results is a major challenge in robotics. First, running experiments often require expensive hardware. Historically, such hardware has been almost always custom built and standardized, and complete robot platforms started to emerge only recently. Furthermore, executing experiments involving real robots is often very time consuming and can be a major engineering challenge. As a result, a large chunk of robotics research have been evaluated in simulation or on a very limited scale.

The result of robotics experiments depend greatly on the sensory data captured by the robot operating in its environment. Such data are inherently unstable over time and depend on the actions taken by the robot, but also on the dynamic properties of real-world environments. This aspect is particularly pronounced in case of visual and multi-modal place classification, where the data of interest capture general appearance of large-scale environments. The appearance of places varies in time because of illumination changes (day and night, artificial light on and off) and because of human activities (furniture moved around, objects being taken out of drawers, and so on). All this calls for standardized benchmarks and databases allowing for fair comparisons, simplification of the experimental process, and as a result, that would boost progress in the field.

Databases and standard benchmarks have long been exploited in the computer vision community, especially for the tasks of object recognition and categorization [4], [5], [6] as well as scene understanding [7], [8]. Also in robotics, research on Simultaneous Localization and Mapping (SLAM) heavily exploits publicly available datasets [9], [10], [11]. An important component when providing the community with a standard dataset is to also provide a standard evaluation procedure. As a result, several research-oriented challenges and competitions emerged around publicly available datasets [5], [12].

The Robot Vision challenge was motivated by those principles and the need to provide the community with similar

benchmark for the task of visual and multi-modal place classification. Our aim was to address the distinct characteristics of the problem that were not reflected by the standard computer vision benchmarks or the typical robotics SLAM evaluations. In case of place classification, we assume that the data is captured by a mobile robot platform. This defines a specific visual domain in which occlusions and non-informative samples are typical and data consists of continuous stream of heavily dependent samples. Moreover, additional information from other robotics sensors, such as laser range scanners, might be available.

Since the early editions of the robot vision challenge, our focus was on addressing those specific characteristics of the place classification problem. Equally important was the robustness to real-world variations in typical human environments. This included illumination and weather conditions as well as short-term and long-term dynamic changes (e.g. presence of people, rooms being re-decorated, etc.). This resulted in a series of datasets, benchmarks and a challenge that became a unique and important event for the robotics community.

III. TASK EVOLUTION

Despite the robot vision challenge was early planned as a visual place recognition competition, other additional tasks have been included since its birth. Moreover, the information provided to the participants has also changed from one edition to the other. A summary of this evolution can be seen in Table I.

The first edition of the competition [13] included room annotations, but also pose annotations. Concretely, each image was annotated with two different types of information: a) the label of the room where the image was acquire from and b) the specific $\langle x, y, \theta \rangle$ pose of the robot. Although pose annotations were included in the training data, participants were encouraged not to use this information in their final proposals.

In the second and third editions [14], [15], pose annotations were removed and monocular images were replaced by stereo ones, allowing participants to exploit the 3D configuration of the environment. The fourth edition of the challenge [16] included two different cues: visual information and depth information. In this edition, the depth information was provided in the form of depth images. Finally, the fifth edition [17] included unprocessed 3D information in the form point cloud files (PCD format [18]).

With respect to the objectives of the competition, two main tasks have been proposed since the first edition. Both tasks focus on visual place recognition, but they differ in the source of information. For the first task, participants have to provide information about the location of the robot separately for each test image. On the other hand, in the second task the temporal continuity of the sequence can be used to classify images: when presented a test image participants can rely on the information obtained using previous images, making this task closer to real-world robotic localization scenarios. The fifth edition of the challenge also introduced an object recognition task. Visual place recognition and object recognition can be

considered as two subproblems of semantic localization, where each location is described in terms of its semantic contents.

A. Datasets

Since 2009, several datasets have been created for the Robot Vision competition. The first dataset used in the challenge was The KTH-IDOL2 database [19]. This dataset was acquired using a mobile robot platform in the indoor environment of The Computer Vision and Active Perception laboratory (CVAP) at The Royal Institute of Technology (KTH) in Stockholm, Sweden. Each training image was annotated with the topological location of the robot and its pose $\langle x, y, \theta \rangle$. As previously mentioned, although the pose information was provided in the training data, participants were encouraged not to make use of this information in their final submission. The two editions of the competition that took place in 2010 were based on COLD-Stockholm, an extension of COLD: The COsy Localization Database [20]. This dataset was generated using a pair of high-quality cameras for stereo vision inside the same environment, as for the KTH-IDOL2 dataset. The fourth edition of the challenge used images from the unreleased VIDA dataset [16]. This dataset includes perspective and range images acquired with a Kinect camera at the Idiap Research Institute in Martigny, Switzerland. Depth information was provided in the form of depth images, with color codes used to represent different distances. Finally, the fifth edition of the competition used images from the unreleased dataset ViDRiLO: The Visual and Depth Robot Indoor Localization with Objects information Dataset. This dataset includes images of the environment and point cloud files (in PCD format) [17].

There exist quite important variations in the data provided to the participants in each of the five editions of the competition. Since the visual place recognition task was treated as a classification problem, two key factors must be analyzed: 1.- the type and number of classes, and 2.- the number of training, validation and test images. In Table II we report the number of classes, as well as the number of training, validation and test images in each edition of the competition. Let us remember that second and third editions included an unknown class not imaged in the training/validation sequences.

Table II
NUMBER OF CLASSES AND TRAINING, VALIDATION, AND TESTING INSTANCES

Task Edition	Number of Classes	Number of Images		
		Training	Validation	Test
1st	5	2899	2789	1690
2nd	9	12684	4783	5102
3rd	10	4782	2069	2741
4th	9	7112	0	6468
5th	10	5263	1869	3515

All classes were named according to the common name of the place/room (e.g. kitchen, corridor) or its expected functionality (printer area, videoconference room). The complete list of classes and their inclusion in each edition of the challenge is reported in Table III. As it can be seen, the number of classes increased from 5 to 10 since 2009. The first three editions of the competition used different class

Table I
TASK EVOLUTION

		1st Edition	2nd Edition	3rd Edition	4th Edition	5th Edition
Sources	Monocular Images	X	-	-	X	X
	Stereo Images	-	X	X	-	-
	Depth Images	-	-	-	X	-
	Point Clouds	-	-	-	-	X
	Semantic Annotations	X	X	X	X	X
	Pose Annotations	X	-	-	-	-
Objectives	Two Tasks	X	X	X	X	-
	Unknown Classes	-	X	X	-	-
	Kidnappings	-	-	-	X	-
	Object Detection	-	-	-	-	X

labels for rooms that actually shared a common semantics, but with spatial differences (e.g. meeting room and large meeting room). Since the fourth edition, class labels represent just semantic categories (no spatial characteristics), and are expected to lead into a standard semantic labelling system.

Table III
SCENE CLASSES

Scene Class	Key	Task Edition				
		1st	2nd	3rd	4th	5th
Corridor	CR	X	X	X	X	X
Kitchen	KT	X	X	X	-	-
Printer Area	PA	X	X	X	X	-
One-Person Office	1PO	X	-	-	-	-
Two-Persons Office	2PO	X	-	-	-	-
Elevator Area	EA	-	X	X	X	X
Large Office 1	LO1	-	X	-	-	-
Large Office 2	LO2	-	X	-	-	-
Small Office 2	SO2	-	X	-	-	-
Students Office	STO	-	X	-	X	X
Lab	LAB	-	X	-	-	-
Large Office	LO	-	-	X	-	-
Meeting Room	MR	-	-	X	-	-
Recycle Area	RA	-	-	X	-	-
Small Office	SO	-	-	X	-	-
Toilet	TL	-	-	X	X	X
Large Meeting Room	LMR	-	-	X	-	-
Lounge Area	LGA	-	-	-	X	-
Professor Office	PO	-	-	-	X	X
Videoconference Room	VC	-	-	-	X	X
Technical Room	TR	-	-	-	X	X
Hall	HA	-	-	-	-	X
Secretary	SC	-	-	-	-	X
Warehouse	WH	-	-	-	-	X

Fig. 1 shows an exemplar for each class in all the challenge editions. In this figure, we have used the class keys defined in Tab. III. Such table shows how some classes have been maintained throughout most of the editions of the challenge (e.g. Corridor, Elevator Area and Students Office). Moreover, there are some classes with different names but representing similar places (e.g. Professor Office, 1-Person Office, Small Office and Small Office2). As it has been introduced, the evolution in the class naming has been carried out with generalization purposes. For illustration purposes Fig. 2 shows exemplar images from these classes for three different editions of the competition.

B. Participation

The robot vision challenge has received a considerable attention from the research community since its release. We

Figure 2. Class evolution: images for classes Corridor (left), Elevator Area (center) and Students Office (right) from the datasets used in the second (top row), fourth (middle row) and fifth edition (bottom row).

observed a similar scenario during all these years: a large number of research groups registered, but a small percentage of them submitted results and wrote a working note with their proposal details (see Table IV).

Table IV
PARTICIPATION

Participation	Task Edition				
	1st	2nd	3rd	4th	5th
Registered Groups	19	28	71	43	39
Participant Groups	7	8	7	8	6
Working Notes Submitted	5	3	3	4	2

With respect to the obtained results, all the editions used a score that computed the performance of the participant submission. This score was always based on positive values for test images correctly classified and negative values for misclassified ones. We also allowed the possibility of not classify a test images, resulting in a non effect on the score. The maximum reachable scores for the mandatory task of each edition were 1690, 5102, 2741, 2445, and 7030, respectively. Regarding the optional task, the maximum scores were 1690, 5102, 2741, and 4079, respectively for the first to fourth editions. The fifth edition of the task had no optional task. All the results are shown in Tables V and VI, for the mandatory and optional tasks of each edition respectively.

In order to graphically present the results obtained for all the editions of the challenge, we have created two boxplots

Figure 1. Exemplar images for every class and edition combination. Classes are named using the keys defined in Tab. III

Table V
MANDATORY TASK RESULTS

Results	Task Edition				
	1st	2nd	3rd	4th	5th
Maximum Score	1690.0	5102.0	2741	2445	7030.000
Score 1st Group	890.5	3824.0	677	2071	6033.500
Score 2nd Group	793.0	3674.0	662	1817	5722.500
Score 3rd Group	787.0	3372.5	638	1348	5004.750
Score 4th Group	784.0	3344.0	253	1225	4638.250
Score 5th Group	544.0	3293.0	62	1028	4497.875
Score 6th Group	511.0	3272.0	-20	551	-487.000
Score 7th Group	456.5	2922.5	-77	462	-
Score 8th Group	-	2283.5	-	-70	-

Table VI
OPTIONAL TASK RESULTS

Results	Task Edition			
	1st	2nd	3rd	4th
Maximum Score	1690	5102	2741	4079
Score 1st Group	916.5	3881.0	2052	3930
Score 2nd Group	884.5	3783.5	62	3859
Score 3rd Group	853.0	3453.5	-67	3169
Score 4th Group	-	2822.0	-	1041

shown in Fig. 3 and Fig. 4. We can observe how the first two editions of the competitions were quite balanced, both for the mandatory and optional subtasks. The experience achieved in

the first year was successfully used to obtain higher results in the second one. The third edition notoriously increased the difficulty of the competition, with low results in the mandatory subtask. This happened because the test sequence was acquired on a different floor of the building, which introduced important changes in the scene objects distribution. The changes introduced in the fourth edition were translated into larger variations between participant results. Finally, the fifth edition of the competition was properly managed by participants, with 5 out of 6 groups obtaining results higher than the 60% of the maximum score.

IV. PARTICIPANTS PROPOSALS AND REMARKABLE RESULTS

Twenty-nine different groups registered to the first edition of the Robot Vision Challenge [13], organized in 2009. Amongst them, seven groups submitted at least one run to the competition, with four groups competing on both the mandatory and the optional task. For mandatory task, a wide range of techniques were proposed for the image representation and classification steps. The best result, 890.5 points out of 1690, was obtained by the Idiap group using a multi-cue discriminative approach [21]. The visual cues considered by this group included two global descriptors: CRFH [22] and PACT [23]; as well as two local descriptors: SIFT [23] and SURF [24].

Figure 3. Results out of the maximum score for the mandatory task.

Figure 4. Results out of the maximum score for the optional task.

A Support Vector Machine (SVM) [25] was trained for each visual cue and a high-level cue integration scheme (DAS) [26] was used to combine the scores provided by the different SVMs. Interestingly, the DAS cue integration method assigned the highest weight to the SIFT features and a zero weight to the SURF features (i.e. the SURF features were actually not used in the final system). A threshold-based system was then used to refrain from making a decision. For the optional task the best result, 916.5 points, was obtained by the SIMD group [27] using a particle filter approach to estimate the position of the robot given the previous position. A set of candidate positions (particles) were sampled around the previous estimated position and separately evaluated using the similarity between the query image and the training images whose annotated positions (obtained using the annotated odometry information) were the closest to the considered particle.

The 2010@ICPR edition [28] had a participation similar to the first edition, with eight participating groups, out of which four competed for both the mandatory and the optional task. As mentioned above, participants were provided in this edition with the stereo images obtained from two cameras mounted on the robot. Amongst the several proposals for the mandatory task, the approach adopted by the CVG group [29] stood out for its full usage of the stereo images to reconstruct the 3D geometry of the rooms. Using the reconstructed geometry the authors could extract viewpoint invariant features (VIP) [30], while canonical SIFT features were also extracted

from monocular images. Using both VIP and SIFT features, each query image was then represented as a bag of visual words and matched with the closest training images. Finally, a geometrical verification step was performed on the top-10 matching images, and the matching images re-ranked according to their geometrical score. Using this approach the CVG group achieved the best score of 3824 points out of 5102. The optional task was once again won by the SIMD group [31], this time by computing SIFT similarities between test frames and a set of training candidate frames, which was selected by means of clustering techniques. In addition, a sort of temporal smoothing using prior assigned labels was used to classify very uncertain test frames.

The third edition of the Robot Vision challenge [32] was attended by seven groups, with three of them participating to both tasks: mandatory and optional. For this edition, participants were asked to classify images recorded on a floor different from the one used to acquire the training images. Consequently, this edition of the competition required the algorithms to show higher generalization capabilities. For second time, the mandatory task was won by the CVG group, with an approach combining a weighted k -NN search using global features, with a geometric verification step [33]. Surprisingly, this winning approach made use of only one single holistic image representation (GIST) [34]. Furthermore, the GIST representation of a query image was directly matched with the GIST representations of the training images, by sorting the training images according to their ℓ_2 -distance to the query image. To allow the system to refrain from making a wrong decision, the matching images were compared with a set of ambiguous images learned from the training data (e.g. images taken nearby door areas). If the majority of the matching images were not ambiguous, a geometric verification was performed by computing the angular distance between the vanishing points of the query image and of the matching images. An image was finally assigned to the class of the matching image at the lowest angular distance. This approach obtained a score of 677 points out of 2741. For the optional task, the approach proposed by the Idiap group [35] resulted to be the most effective. The proposed multi-cue system combined up to three different visual descriptors (namely: PHOG [36], CRFH and PLBP), in a discriminative multiple-kernel SVM. A door detector was implemented for discovering the transition from one room to another, while a stability estimation algorithm was used to evaluate the stability of the classification process. If the process was stable enough, the samples with a low margin were assigned to the class assigned to the last n samples encountered after crossing a door. Otherwise, they were assigned to the “Unknown” class. Using this approach, the Idiap group obtained the winning score of 2052 points out of 2741.

As mentioned above, the 2012 edition [16] introduced range images obtained with a Microsoft Kinect sensor. Up to 43 groups were recorded in the task, but only 8 submitted their results. The organizers proposed a baseline method for both the feature extraction and the classification steps [37]. Regarding feature extraction, the Pyramid Histogram of Oriented Gradients (PHOG) [36] (for visual images), and the Normal

Aligned Radial Feature (NARF) [38] (for depth images) were proposed. Concerning classification and cue integration, the Batch Strongly Convex Multi Online-Kernel Learning (OB-SCURE) [39] method was selected. Thanks to this proposal, the results of the baseline submission was 462 out of 2445 for the mandatory task and 1041 out of 4079 for the optional one. The group from the Universidad Tecnológica Nacional, Cordoba, Argentina (CIII UTN FRC) [40] was the winner for both tasks with a score of 2071 and 3930 respectively. This group made use of the depth information (in fact, the only group which use it). As low-level features they used the SIFT descriptor [41], reducing the dimensionality of this descriptor with Principal Components Analysis (PCA) and encoding the information in Fisher Vectors image signatures [42] for both color and depth images. Finally, the classification phase relied on SVMs. The group from the Alexandru Ioan Cuza University, Iasi, Romania (UAIC2012) [43] presented an interesting approach for the mandatory task, which used a combination of HSV and RGB color histograms in conjunction with SIFT descriptors (following a Bag-of-Word approach). They achieved 1348 points. The group from the Ural Federal University, Yekaterinburg, Russian Federation (USUroom409) [44] proposed the use of a growing Kohonen network. They obtained 1225 points in the mandatory task using as inputs some statistical pixel values like expected value, variance, or standard deviation.

The fifth edition [17] encouraged participants for using 3D information (point cloud files) with the inclusion of rooms completely imaged in dark. It also introduced the identification of the objects present in the scene, as well as a proposed classification method similar to that of the 2012 edition, used to present a baseline score of 5004.75 out of 7030 points. The highest result, 6033.5 points, was obtained by the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China (MIAR ICT) [45]. They proposed the use of Kernel descriptors [46] for both visual and depth information, while PCA was applied for dimensionality reduction. They used SVM classifiers and managed object recognition and room classification separately. Actually, both problems were expected to be handled together, but none of the participants presented a proposal where the objects appearance (or lack) is used to classify the room. The group from the University of Sfax National School of Engineers, Tunisia (REGIM) [47] achieved a score of 4638.25 points, slightly lower than the baseline score. Visual images were processed using Pyramid Histogram of Visual Words (PHOW) [48], and linear SVMs were selected as classification model.

As a general remark, we can conclude that most editions were won by those participants taking advantage of the introduced novelties. Namely, those proposals that ranked first in the second and third editions were based on the spatial geometry acquired from stereo images. The winner of the fourth edition was the only proposal using range information, and a similar scenario was found in the 2013 edition. Regarding the participant proposals, SIFT features, the Bag-of-Words approach, and SVM classifiers stood out among the rest of the techniques for their inclusion in most of the winning proposals.

V. CONCLUSIONS

In this paper, an overview of the five consecutive editions of the Robot Vision challenge at ImageCLEF has been presented. First, we have described the challenge and why we consider this challenge a relevant and important competition for the robotics science community in particular, and to the research community in computer vision in general. We have also described how the tasks proposed in each challenge have evolved from the first edition to the last one. The challenge has evolved over the different editions: different information sources, different scene categories, visual and depth images, intra-scene object information, amount of training and validation data, and baseline useful software, among others. As a result of these changes different datasets have been released and publicly provided as a testbeds for the visual scene classification problem. Finally, we have presented the more relevant proposals as well as a summary of the best results obtained in the five editions of the competition.

As a main conclusion we can argue that the Robot Vision challenge has provided some valuable resources including datasets, benchmarking techniques, and a state-of-the-art solutions to the visual place classification problem.

For future editions of the challenge we plan to manage both room classification and object recognition problems jointly. Subsequently, we will challenge participants to classify rooms using as input the list of objects recognized in a scene.

REFERENCES

- [1] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *International Journal of Computer Vision (IJCV)*, vol. 77, no. 1-3, pp. 259-289, 2008.
- [2] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *The International Journal of Robotics Research*, pp. 298-320, 2009.
- [3] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008, pp. 1-8.
- [4] G. Griffin, A. Holub, and P. Perona, "Caltech-256 Object Category Dataset," Caltech, Tech. Rep. 7694, 2007.
- [5] "The PASCAL Visual Object Classes challenge," available at: <http://www.pascal-network.org/challenges/VOC/>.
- [6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248-255.
- [7] J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba, "SUN Database: Large-scale scene recognition from abbey to zoo," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 3485-3492.
- [8] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgb-d images," in *Proc. of the European Conference on Computer Vision (ECCV)*, 2012, pp. 746-760.
- [9] A. Howard and N. Roy, "The Robotics Data Set Repository (Radish)," 2003, available at: <http://radish.sourceforge.net/>.
- [10] E. Nebot, "The Sydney Victoria Park dataset," available at: <http://www-personal.acfr.usyd.edu.au/nebot/dataset.htm>.
- [11] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *Proc. of the International Conference on Intelligent Robot Systems (IROS)*, 2012, pp. 573-580.
- [12] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," 2014.

- [13] A. Pronobis, L. Xing, and B. Caputo, "Overview of the clef 2009 robot vision track," in *Multilingual Information Access Evaluation II. Multimedia Experiments*, ser. Lecture Notes in Computer Science, C. Peters, B. Caputo, J. Gonzalo, G. J. Jones, J. Kalpathy-Cramer, H. Müller, and T. Tsirikla, Eds. Springer Berlin Heidelberg, 2010, vol. 6242, pp. 110–119.
- [14] A. Pronobis, H. I. Christensen, and B. Caputo, "Overview of the imageclef@ icpr 2010 robot vision track," in *Recognizing Patterns in Signals, Speech, Images and Videos*. Springer, 2010, pp. 171–179.
- [15] A. Pronobis, M. Fornoni, H. I. Christensen, and B. Caputo, "The Robot Vision Track at ImageCLEF 2010," in *CLEF Notebook Papers/LABs/Workshops*, 2010.
- [16] J. Martínez-Gómez, I. García-Varea, and B. Caputo, "Overview of the ImageCLEF 2012 Robot Vision Task," in *Working Notes of the ImageCLEF Laboratory*, 2012.
- [17] J. Martínez-Gómez, I. García-Varea, M. Cazorla, and B. Caputo, "Overview of the ImageCLEF 2013 Robot Vision Task," in *Working Notes of the ImageCLEF Laboratory*, 2013.
- [18] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2011, pp. 1–4.
- [19] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "The kth-idol2 database," KTH, CAS/CVAP, Tech. Rep. 304, 2006.
- [20] A. Pronobis and B. Caputo, "Cold: The cosy localization database," *The International Journal of Robotics Research*, vol. 28, no. 5, pp. 588–594, 2009.
- [21] L. Xing and A. Pronobis, "Multi-cue discriminative place recognition," in *Multilingual Information Access Evaluation II. Multimedia Experiments*, ser. Lecture Notes in Computer Science, C. Peters, B. Caputo, J. Gonzalo, G. J. Jones, J. Kalpathy-Cramer, H. Müller, and T. Tsirikla, Eds. Springer Berlin Heidelberg, 2010, vol. 6242, pp. 315–323.
- [22] O. Linde and T. Lindeberg, "Object recognition using composed receptive field histograms of higher dimensionality," in *17th International Conference on Pattern Recognition (ICPR)*, Cambridge, UK, 2004, pp. 1–6.
- [23] J. Wu and J. M. Rehg, "Where am I: place instance and category recognition using spatial PACT," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, USA, 2008, pp. 1–8.
- [24] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-up robust features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.
- [25] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.
- [26] M. Nilsback and B. Caputo, "Cue integration through discriminative accumulation," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2004, pp. 578–585.
- [27] J. Martínez-Gómez, A. Jiménez-Picazo, and I. García-Varea, "A particle-filter based self-localization method using invariant features as visual information," in *Working Notes for CLEF. Workshop co-located with the 13th European Conference on Digital Libraries (ECDL)*, Corfù, Greece, 2009.
- [28] A. Pronobis, H. I. Christensen, and B. Caputo, "Overview of the imageclef@icpr 2010 robot vision track," in *Recognizing Patterns in Signals, Speech, Images and Videos - ICPR Contests*, Istanbul, Turkey, 2010, pp. 171–179.
- [29] F. Fraundorfer, C. Wu, and M. Pollefeys, "Methods for combined monocular and stereo mobile robot localization," in *Recognizing Patterns in Signals, Speech, Images and Videos - ICPR Contests*, Istanbul, Turkey, 2010, pp. 180–189.
- [30] C. Wu, B. Clipp, X. Li, J. Frahm, and M. Pollefeys, "3d model matching with viewpoint-invariant patches (VIP)," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, USA, 2008, pp. 1–8.
- [31] J. Martínez-Gómez, A. Jiménez-Picazo, J. Gámez, and I. García-Varea, "Combining image invariant features and clustering techniques for visual place classification," in *Recognizing Patterns in Signals, Speech, Images and Videos*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2010, vol. 6388, pp. 200–209.
- [32] A. Pronobis, M. Fornoni, H. I. Christensen, and B. Caputo, "The robot vision track at imageclef 2010," in *CLEF LABs and Workshops, Notebook Papers*, Padua, Italy, 2010.
- [33] O. Saurer, F. Fraundorfer, and M. Pollefeys, "Visual localization using global visual features and vanishing points," in *CLEF LABs and Workshops, Notebook Papers*, Padua, Italy, 2010.
- [34] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [35] M. Fornoni, J. Martínez-Gómez, and B. Caputo, "A multi cue discriminative approach to semantic place classification," in *CLEF LABs and Workshops, Notebook Papers*, Padua, Italy, 2010.
- [36] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CIVR)*, New York, NY, USA, 2007, pp. 401–408.
- [37] J. Martínez-Gómez, I. García-Varea, and B. Caputo, "Baseline multi-modal place classifier for the 2012 robot vision task," in *Working Notes of the ImageCLEF Laboratory*, 2012.
- [38] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "Narf: 3d range image features for object recognition," in *Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, 2010.
- [39] F. Orabona, L. Jie, and B. Caputo, "Online-batch strongly convex multi kernel learning," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010, pp. 787–794.
- [40] J. Redolfi and J. Sánchez, "Leveraging robust signatures for mobile robot semantic localization," in *Working Notes of the ImageCLEF Laboratory*, 2012.
- [41] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [42] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.
- [43] E. Boros, A.-L. Gînsca, and A. Iftene, "Uaic participation at robot vision@ 2012-an updated vision," in *Working Notes of the ImageCLEF Laboratory*, 2012.
- [44] K. Zhagorina and A. Buslavyev, "Computer analysis of visual image similarity," in *Working Notes of the ImageCLEF Laboratory*, 2012.
- [45] R. Xu, S. Jiang, X. Song, S. Wang, Y. Xie, F. Wang, and X. Lv, "Miar ict participation at robot vision 2013," in *Working Notes of the ImageCLEF Laboratory*, 2013.
- [46] L. Bo, X. Ren, and D. Fox, "Kernel Descriptors for Visual Recognition," in *Advances in Neural Information Processing Systems*, 2010, pp. 244–252.
- [47] A. Ksibi, B. Ammar, A. B. Ammar, C. B. Amar, and A. M. Alimi, "Regimrobvid: Objects and scenes detection for robot vision 2013," in *Working Notes of the ImageCLEF Laboratory*, 2013.
- [48] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Proc. of the IEEE 11th International Conference on Computer Vision (ICCV)*, 2007, pp. 1–8.