REPLICABLE AND MEASURABLE
ROBOTICS RESEARCH

GALILEO — IMAGE LICENSED BY INGRAM PUBLISHING,
LAMP—©ISTOCKPHOTO.COM/JONNY JIM

Where Are We After Five Editions?

Robot Vision Challenge, a Competition that Evaluates Solutions for the Visual Place Classification Problem

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

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 subobjectives 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 data sets, benchmarking techniques, and a wide experience in the place classification research that is reflected in this article.

Digital Object Identifier 10.1109/MRA.2015.2460931
Date of publication: 10 December 2015

The Challenge Over Time

The Robot Vision challenge started as part of the ImageCLEF lab in 2009. The challenge has since been organized five times, initially devoted to image-based place categorization, how to determine

IMAGE LICENSED BY GRAPHIC STOCK

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 [1]–[3] show that it is possible to recognize both manufactured and natural objects, such as cars, cows, or offices 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 red, green, blue-depth (RGB-D) data and to perform both place and object categorization. Due to the environmental variations, such as lighting or moving furniture around, 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 multimodal sensor data and a richer challenge in terms of categorization of places and objects.

Creating a Categorization Competition

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

The use of generalist images prevents us from using PASCAL VOC proposals to solve the problem of robot localization.

The results 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 and on the dynamic properties of real-world environments. This aspect is particularly pronounced in the case of visual and multimodal place classification, where the data of interest capture the general appearance of large-scale environments. The appearance of places varies in time because of illumination changes (day and night, artificial light turned 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 and simplification of the experimental process and, as a result, would provide a boost for progress in the field of robotics.

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

The Robot Vision challenge was motivated by those principles and the need to provide the community with a similar benchmark for the task of visual and multimodal 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 the case of place classification, we assume that the data are captured by a mobile robot platform. This defines a specific visual domain in which occlusions and noninformative samples are typical and data consist of a 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 the short-term and long-term dynamic changes (e.g., the presence of people, rooms being redecorated, and so on). This resulted in a series of data sets, benchmarks, and a challenge that became a unique and important event for the robotics community.

Related Challenges

The use of competitions in robotics has encouraged the proposal of solutions to some of the most well-known problems over the past two decades [13]. The RoboCup competition can be considered the most representative challenge [14], as it includes several leagues where a large set of tasks is evaluated, such as robot design and construction, navigation, localization, mapping, perception, decision making, and human–robot interaction. The requested infrastructure is the main drawback of the RoboCup. Namely, a research group aiming to participate at a RoboCup league should include a medium/large number of members and robotic platforms, but also an appropriate environment for the deployment of the competition scenario.

Unlike the RoboCup, challenges based on data sets are approachable for most researchers. This encourages participation and promotes heterogeneous proposals from multidisciplinary groups. The PASCAL Visual Object Classes (VOCs) challenge [5] was first proposed in 2005, and it introduces the object detection and recognition problem relying on the use of data sets. This competition evaluates pure computer vision proposals for detecting and recognizing objects in images obtained from Flickr. Despite the intrinsic relationship between object recognition and robotics, the use of generalist images prevents us from using PASCAL VOC proposals to solve the problem of robot localization. The ImageNET Large Scale Visual Recognition Challenge (ILSVRC) [12] can be seen as the natural successor of PASCAL VOC, which ended in 2012.

The Reconstruction Meets Recognition Challenge (RMRC) [15] started in 2013 and is similar to the Robot Vision challenge. First, this challenge evaluates proposals for two robotic problems as segmentation and detection. Moreover, these tasks are presented for indoor environments that have been imaged using RGB-D sensors (using images from the New York University [8] data set). The main difference with respect to the Robot Vision task is the annotation scheme: RGB-D images are labeled at pixel/point level with object categories. In addition, the RMRC RGB-D images were not recorded using a temporal continuity, which is an important issue to keep in mind when trying to solve a localization task.

We can also find an ongoing challenge proposal with a strong relationship with the Robot Vision task. This is the Large-Scale Scene Understanding Challenge (LSUN) (http://lsun.cs.princeton.edu), which holds a scene classification task. In this task, perspective images should be classified with the scene category using ten different options. Some of the scene categories used in the LSUN challenge have been already used in the Robot Vision task, like the conference room or the kitchen.

Table 1. The task evolution.

		First Edition	Second Edition	Third Edition	Fourth Edition	Fifth 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	—	—	—	—
Objective	Two tasks	X	X	X	X	—
	Unknown classes	—	X	X	—	—
	Kidnappings	—	—	—	X	—
	Object detection	—	—	—	—	X

Task Evolution

Despite the fact that the robot vision challenge was initially 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 1.

The first edition of the competition [16] included room annotations, but also poses annotations. Concretely, each image was annotated with two different types of information: 1) the label of the room where the image was acquired from and 2) the specific $<x, y, \theta>$ 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 [17], [18], pose annotations were removed and monocular images were replaced by stereo ones, allowing participants to exploit the three dimensional (3-D) configuration of the environment. The fourth edition of the challenge [19] 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 [20] included unprocessed 3-D information in the form of point cloud files (PCD format [21]).

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 with a test image, participants can rely on the information obtained using the 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.

The organizers proposed a baseline method for both the feature extraction and the classification steps.

Data Sets

Since 2009, several data sets have been created for the Robot Vision competition. The first data set used in the challenge was the KTH-IDOL2 database [22]. This data set 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 $< x, y, \theta >$. As previously mentioned, although the pose information was provided in the training data, participants

Table 2. The number of classes and training, validation, and testing instances.

Task	Number	Number of Images		
Edition	of Classes	Training	Validation	Test
First	5	2,899	2,789	1,690
Second	9	12,684	4,783	5,102
Third	10	4,782	2,069	2,741
Fourth	9	7,112	0	6,468
Fifth	10	5,263	1,869	3,515

All the editions used a score that computed the performance of the participant submission.

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 the COsy Localization Database [23]. This data set was generated using a pair of high-quality cameras for stereo vision inside the same environment, similar to the KTH-IDOL2 data set. The fourth edition of the challenge used images from the unreleased VIDA data set [19]. This data set 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 Visual and Depth Robot Indoor Localization with Objects (ViDRILO) information data set [24]. This data set includes images of the environment and point cloud files (in PCD format) [20].

Important variations in the data provided to the participants in each of the five editions of the competition also exist. 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 2, we report the number of classes, as well as the number of training, validation, and test images in each edition of the competition. It should be noted that the second and third editions included an unknown class not imaged in the training/validation sequences.

All classes were named according to the common name of the place/room (e.g., kitchen, corridor) or its expected functionality (printer area, video conference room). The complete list of classes and their inclusion in each edition of the challenge are reported in Table 3. As can be seen, the number of classes increased from five to ten since 2009. The first three editions of the competition used different class labels for rooms that actually shared a common semantics, but with spatial differences (e.g., a meeting room and a large meeting room). Since the fourth edition, class labels only represent semantic categories (no spatial characteristics) and are expected to lead into a standard semantic labeling system.

Table 3. The scene classes.

		Task Edition				
Scene Class	Key	First	Second	Third	Fourth	Fifth
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-person 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	–	–	–
Student's 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's office	PO	–	–	–	X	X
Video conference room	VC	–	–	–	X	X
Technical room	TR	–	–	–	X	X
Hall	HA	–	–	–	–	X
Secretary	SC	–	–	–	–	X
Warehouse	WH	–	–	–	–	X

Figure 1 shows an exemplar image for each class in all the challenge editions. In this figure, we have used the class keys defined in Table 3. This table shows how some classes have been maintained throughout most of the editions of the challenge (e.g., corridor, elevator area, and student office). Moreover, there are some classes with different names but representing similar places (e.g., professor's office, one-person office, small office, and small office 2). As has been introduced, the evolution in the class naming has been carried out with generalization purposes. For illustration purposes, Figure 2 shows exemplar images from these classes for three different editions of the competition.

Participation

The robot vision challenge has received considerable attention from the research community since its release. We observed a similar scenario during all these years: a large number of research groups registered, but a small percentage of them

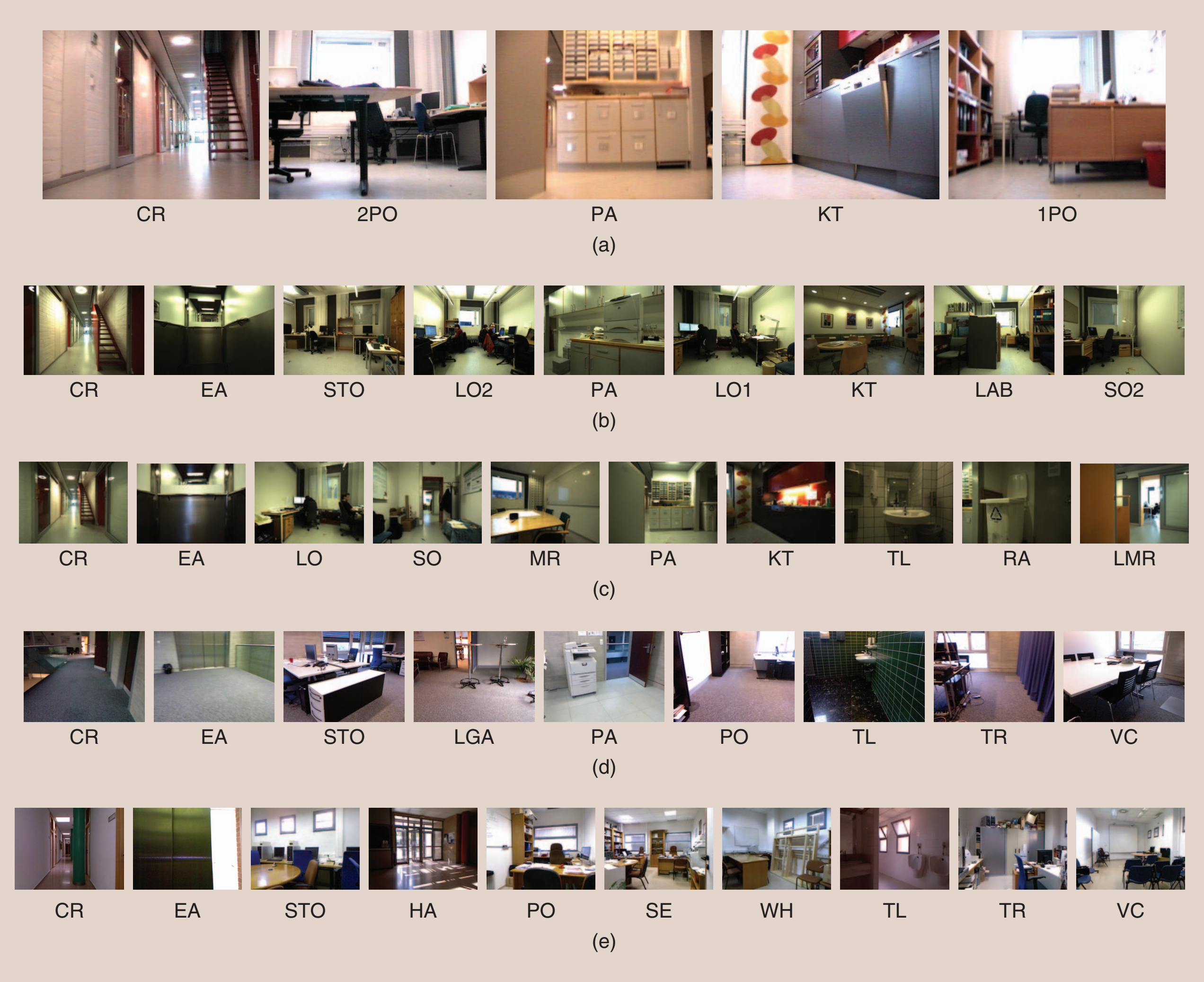

Figure 1. The exemplar images for every class and edition combination. Classes are named using the keys defined in Table 3. (a) RobotVision at ImageCLEF 2009. (b) RobotVision at ImageCLEF 2010 ICPR. (c) RobotVision at ImageCLEF 2010. (d) RobotVision at ImageCLEF 2012. (e) RobotVision at ImageCLEF 2013.

submitted results and wrote a working note with their proposal details (see Table 4).

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 classifying test images, resulting in nonaltering the score. The maximum reachable scores for the mandatory task of each edition were 1,690, 5,102, 2,741, 2,445, and 7,030, respectively. Regarding the optional task, the maximum scores were 1,690, 5,102, 2,741, and 4,079, respectively for the first to fourth editions. The fifth edition of the task had no optional task. All the results are shown in Tables 5 and 6 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 shown in Figures 3 and 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 objects' distribution in the scene. 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 five out of six groups obtaining results higher than the 60% of the maximum score.

Pose annotations were removed and monocular images were replaced by stereo ones.

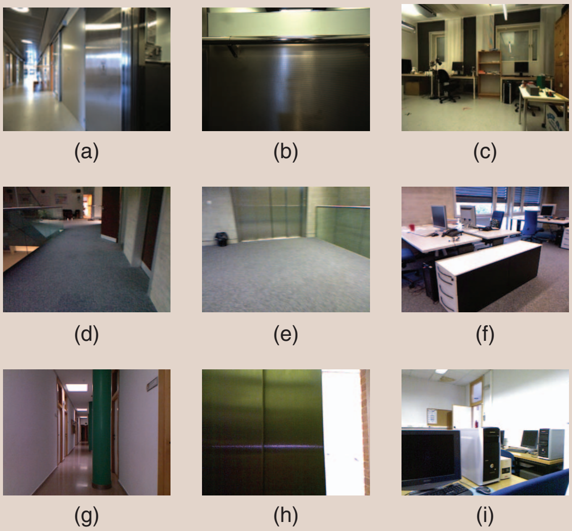

Figure 2. The class evolution. Images for (a), (d), and (g) classes corridor, (b), (e), and (h) elevator area, and (c), (f), and (i) student office.

Participant Proposals and Remarkable Results

First Edition

Twenty-nine different groups registered for the first edition of the Robot Vision challenge [16], organized in 2009. Among them, seven groups submitted at least one run to the competition, with four groups competing both the mandatory and the optional task. For the mandatory task, a wide range of techniques was proposed for the image representation and classification steps. The best result, 793 points out of 1,690, was obtained by the Idiap group using a multicue discriminative approach [25]. The visual cues considered by this group included two global descriptors: 1) composite receptive field histogram (CRFH) [26] and 2) principal component analysis of census transform [27]; as well as two local descriptors: 1) scale invariant feature transform (SIFT) [27] and 2) speeded-up robust features (SURF) [28]. A support vector machine (SVM) [29] was trained for each visual cue, and a high-level cue integration scheme, the discriminative accumulation scheme (DAS) [30], 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 Sistemas Inteligentes y Mineria de Datos (SIMD) group [31] 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.

Table 4. The participation.

	Task Edition				
Participation	**First**	**Second**	**Third**	**Fourth**	**Fifth**
Registered groups	19	28	71	43	39
Participant groups	7	8	7	8	6
Working notes submitted	5	3	3	4	2

Table 5. The mandatory task results.

	Task Edition				
Results	**First**	**Second**	**Third**	**Fourth**	**Fifth**
Maxmimum score	1,690	5,102	2,741	2,445	7,030
Score first group	793	3,824	677	2,071	6,033
Score second group	787	3,674	662	1,817	5,722
Score third group	784	3,372.5	638	1,348	5,004.750
Score fourth group	544	3,344	253	1,225	4,638.250
Score fifth group	511	3,293	62	1,028	4,497.875
Score sixth group	456	3,272	−20	551	−487
Score seventh group	–	2,922.5	−77	462	–
Score eighth group	–	2,283.5	–	−70	–

Second Edition

The 2010@ICPR edition [32] had 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 with the stereo images obtained from two cameras mounted on the robot in this edition. Among the proposals for the mandatory task, the approach adopted by the computer vision and geometry (CVG) group [33] stood out for its full usage of the stereo images to reconstruct the 3-D geometry of the rooms. Using the reconstructed geometry, the authors could extract viewpoint invariant features [34], while canonical SIFT features were also extracted from monocular images. Using this approach, the CVG group achieved the best score of 3,824 points out of 5,102. The optional task was once again won by the SIMD group [35], this time by computing SIFT similarities between test frames and a set of training candidate frames, which was selected by means of clustering techniques.

Third Edition

The third edition of the Robot Vision challenge [36] was attended by seven groups, three of which particpated in both the mandatory and optional tasks. For this edition, participants were asked to classify images recorded on a floor different from the one used to acquire the training

Table 6. The optional task results.

Results	Task Edition First	Second	Third	Fourth
Maxmimum score	1,690	5,102	2,741	4,079
Score first group	916.5	3,881	2,052	3,930
Score second group	890.5	3,783.5	62	3,859
Score third group	884.5	3,453.5	−67	3,169
Score fourth group	853	2,822	–	1,041

images. Consequently, this edition of the competition required the algorithms to show higher generalization capabilities. For the second time, the CVG group won the mandatory task with an approach combining a weighted *k*-NN search using global features, with a geometric verification step [37]. Surprisingly, this winning approach made use of only a single holistic image representation (GIST) [38]. 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 R_2 distance to the query image. 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 2,741. For the optional task, the approach proposed by the Idiap group [39] proved to be the most effective. The proposed multi-cue system combined up to three different visual descriptors (namely, pyramid histogram of oriented gradients (PHOG) [40], CRFH [26], and principal local binary pattern [41]), in a discriminative multiple kernel SVM. A door detector was implemented to determine the transition from one room to another, while a stability estimation algorithm was used to evaluate the stability of the classification process. Using this approach, the Idiap group obtained the winning score of 2,052 points out of 2,741.

Fourth Edition

As mentioned above, the 2012 edition [19] introduced range images obtained with a Microsoft Kinect sensor. Up to 43 groups were recorded in the task, but only eight submitted their results. The organizers proposed a baseline method for both the feature extraction and the classification steps [42]. Regarding feature extraction, the pyramid PHOG [40] (for visual images) and the normal aligned radial feature [43] (for depth images) were proposed. Concerning classification and cue integration, the Batch Strongly Convex Multi Online Kernel Learning (OBSCURE) [44] method was selected. Because of this proposal, the results of the baseline submission were 462 out of 2,445 for the mandatory task and 1,041 out of 4,079 for the optional task. The group from the Universidad Tecnologica Nacional, Cordoba, Argentina (CIII UTN FRC) [45] was the winner for both tasks with a score of 2,071 and 3,930, respectively. This group made use of the depth information (in fact, they were only group that used it). For low-level features, they used the SIFT descriptor [46], reducing the dimensionality of this descriptor with principal components analysis (PCA) and encoding the information in Fisher Vectors image signatures [47] for both color and depth images. Finally, the classification phase relied on SVMs. The group from the Alexandru Ioan Cuza University, Iasi, Romania (UAIC2012) [48] presented an interesting approach for the mandatory task, which used a combination of hue, saturation, and value and RGB color histograms in conjunction with SIFT descriptors (following a bag-of-word approach). They achieved 1,348 points. The group from the Ural Federal University, Yekaterinburg, Russian Federation (USU room 409) [49] proposed the use of a growing Kohonen network. They obtained 1,225 points in the mandatory task using statistical pixel values like expected value, variance, and standard deviation as inputs.

The Robot Vision task has served for sharing techniques and knowledge between worldwide researchers.

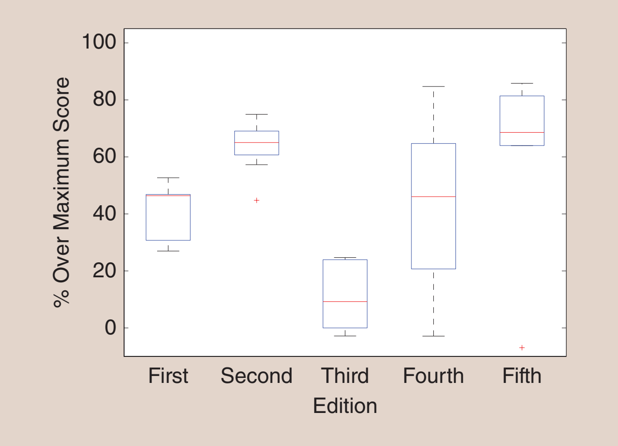

Figure 3. The results out of the maximum score for the mandatory task.

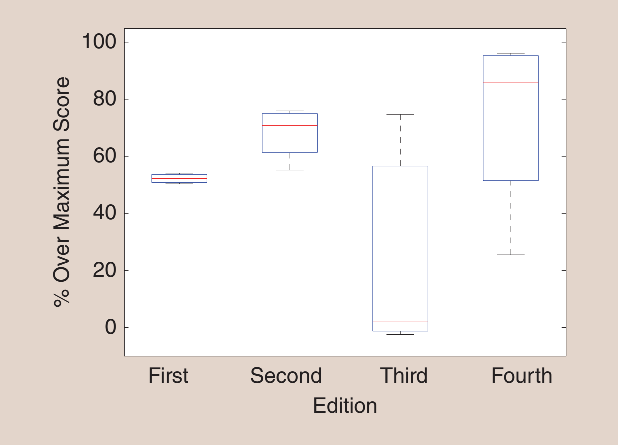

Figure 4. The results out of the maximum score for the optional task.

Fifth Edition

The fifth edition [20] encouraged participants to use 3-D 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 5,004.75 out of 7,030 points. The highest result, 6,033.5 points, was obtained by the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China (MIAR ICT) [50]. They proposed the use of kernel descriptors [51] 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 appearance of the object (or lack thereof) is used to classify the room. The group from the University of Sfax National School of Engineers, Tunisia (REGIM) [52] achieved a score of 4,638.25 points, slightly lower than the baseline score. Visual images were processed using Pyramid Histogram of Visual Words [53], and linear SVMs were selected as the classification model.

We plan to manage both room classification and object recognition problems jointly.

As a general remark, we can point out 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.

Analysis of the Results

In addition to the generation of solutions to the problem provided by each edition, the Robot Vision task has served for sharing techniques and knowledge between worldwide researchers. This experience has helped several robotic laboratories generate their own place classifiers and also develop novel approaches that have been successfully deployed in different environments. Here, we review some of these proposals.

The University of Glasgow only participated in the first edition of the challenge ranking second for the optional task [54], but their novel matching approach was used to develop a robust localization system [55]. The PicSOM content-based image retrieval system [56] from the University of Helsinki was evaluated in the second edition of the challenge. Despite their low ranking (fifth [57]), the experience gathered from the Robot Vision challenge allowed them to develop a concept detection system relying on the use of kernel maps [58]. Members from the CVG group, winner of the second and third editions, also presented an extension of their proposals, namely, the visual mapping and localization problem [59], for outdoor environments. Another remarkable proposal drawn from the Robot Vision competition is the generalist use of Fisher Vectors for image classification, as presented by the winner of the fourth edition of the task in [60].

The use of kernel descriptors, as part of the MIAR ICT participation in the 2013 edition, can be considered the most outstanding technique presented in the Robot Vision challenge [50]. This technique generates rich path-level features from pixel attributes, which are then used to generate image representations suitable for further classifications. Kernel descriptors have been previously used for scene labeling and classification [61], but the temporal continuity of the sequences of images has not been exploited in previous proposals. The MIAR ICT team approach smooths the score of their algorithm using a smoothing window, which increases its accuracy by 8.43% when validated against the validation sequence.

Conclusions

In this article we presented an overview of the five consecutive editions of the Robot Vision challenge at ImageCLEF. First, we described the challenge and why we consider it a relevant and important competition for the robotics science community in particular, and to the research community in computer vision in general. We also described some related competitions and 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, intrascene object information, the amount of training and validation data, and baseline useful software have been used across editions. As a result of these changes, different data sets have been released and publicly provided as a testbed for the visual scene classification problem. Finally, we 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 argue that the Robot Vision challenge has provided valuable resources including data sets, benchmarking techniques, and state-of-the-art solutions to the visual place classification problem. Moreover, the challenge has contributed to the generation of a semantic localization researcher community.

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 the list of objects recognized in a scene as inputs. Also, the use of online development environments will be considered.

Acknowledgments

This work was supported by grant DPI2013-40534-R of the Ministerio de Economia y Competitividad of the Spanish

Government and by Consejería de Educación, Cultura y Deportes of the Junta de Comunidades de Castilla-La Mancha regional government through project PPII-2014-015-P. Jesus Martínez-Gómez is also funded by the JCCM grant POST2014/8171.

References

[1] B. Leibe, A. Leonardis, and B. Schiele, "Robust object detection with interleaved categorization and segmentation," *Int. J. Comput. Vis.*, vol. 77, nos. 1–3, pp. 259–289, 2008.

[2] A. Pronobis, O. M. Mozos, B. Caputo, and P. Jensfelt, "Multi-modal semantic place classification," *Int. J. Robot. Res.*, vol. 29, nos. 2–3, pp. 298–320, 2009.

[3] G. Griffin and P. Perona, "Learning and using taxonomies for fast visual categorization," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2008, pp. 1–8.

[4] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Inst. Technol., Pasadena, CA, Tech. Rep. 7694, 2007.

[5] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.

[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. IEEE Conf. Computer Vision Pattern Recognition*, 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 *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2010, pp. 3485–3492.

[8] P. K. N. Silberman, D. Hoiem, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. European Conf. Computer Vision*, 2012, pp. 746–760.

[9] A. Howard and N. Roy. (2003). The robotics data set repository (Radish). [Online]. Available: http://radish.sourceforge.net/

[10] E. Nebot. The Sydney Victoria Park Dataset. [Online]. Available: 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. Int. Conf. Intelligent Robot Systems*, 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," *Int. J. Comput. Vis.*, Apr. 2015, to be published.

[13] S. Behnke, "Robot competitions-ideal benchmarks for robotics research," in *Proc. IROS-2006 Workshop Benchmarks Robotics Research*, 2006, pp. 1–5.

[14] H. Kitano, M. Asada, Y. Kuniyoshi, I. Noda, and E. Osawa, "RoboCup: The robot world cup initiative," in *Proc. 1st Int. Conf. Autonomous Agents*, ACM, 1997, pp. 340–347.

[15] R. Urtasun, R. Fergus, D. Hoiem, A. Torralba, A. Geiger, P. Lenz, N. Silberman, J. Xiao, and S. Fidler. (2013). Reconstruction meets recognition challenge. [Online]. Available: http://ttic.uchicago.edu/ rurtasun/rmrc/

[16] A. Pronobis, L. Xing, and B. Caputo, "Overview of the CLEF 2009 robot vision track," in *Multilingual Information Access Evaluation II. Multimedia Experiments* (Lecture Notes in Computer Science, vol. 6242). Berlin Heidelberg, Germany: Springer, 2010, pp. 110–119.

[17] 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*. Berlin Heidelberg, Germany: Springer, 2010, pp. 171–179.

[18] A. Pronobis, M. Fornoni, H. I. Christensen, and B. Caputo, "The robot vision track at ImageCLEF 2010," in *Proc. CLEF Notebook Papers/LABs/Workshops*, 2010, pp. 1–6.

[19] J. Martínez-Gómez, I. García-Varea, and B. Caputo, "Overview of the ImageCLEF 2012 robot vision task," in *Proc. Working Notes ImageCLEF Laboratory*, 2012, pp. 1–10.

[20] J. Martínez-Gómez, I. García-Varea, M. Cazorla, and B. Caputo, "Overview of the ImageCLEF 2013 robot vision task," in *Proc. Working Notes ImageCLEF Laboratory*, 2013, pp. 1–12.

[21] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (PCL)," in *Proc. IEEE Int. Conf. Robotics and Automation*, 2011, pp. 1–4.

[22] J. Luo, A. Pronobis, B. Caputo, and P. Jensfelt, "The kth-IDOL2 database," KTH, CAS/CVAP, Tech. Rep. 304, 2006.

[23] A. Pronobis and B. Caputo, "Cold: The cosy localization database," *Int. J. Robot. Res.*, vol. 28, no. 5, pp. 588–594, 2009.

[24] J. Martinez-Gomez, M. Cazorla, I. Garcia-Varea, and V. Morell, "ViDRILO: The visual and depth robot indoor localization with objects information dataset," *Int. J. Robot. Res.*, 2015, to be published.

[25] L. Xing and A. Pronobis, "Multi-cue discriminative place recognition," in *Multilingual Information Access Evaluation II. Multimedia Experiments* (Lecture Notes in Computer Science, vol. 6242). Berlin Heidelberg, Germany: Springer, 2010, pp. 315–323.

[26] O. Linde and T. Lindeberg, "Object recognition using composed receptive field histograms of higher dimensionality," in *Proc. 17th Int. Conf. Pattern Recognition*, Cambridge, U.K., 2004, pp. 1–6.

[27] J. Wu and J. M. Rehg, "Where am I: place instance and category recognition using spatial PACT," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Anchorage, AK, 2008, pp. 1–8.

[28] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Speeded-up robust features (SURF)," *Comput. Vis. Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[29] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge, U.K., Cambridge Univ. Press, 2000.

[30] M. Nilsback and B. Caputo, "Cue integration through discriminative accumulation," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2004, vol. 2, pp. 578–585.

[31] 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 CLEF. Workshop Co-Located with 13th European Conf. Digital Libraries*, Greece, 2009, pp. 1–14.

[32] A. Pronobis, H. I. Christensen, and B. Caputo, "Overview of the ImageCLEF@ICPR 2010 robot vision track," in *Proc. Recognizing Patterns Signals, Speech, Images Videos—ICPR Contests*, Istanbul, Turkey, 2010, pp. 171–179.

[33] F. Fraundorfer, C. Wu, and M. Pollefeys, "Methods for combined monocular and stereo mobile robot localization," in *Proc. Recognizing Patterns Signals, Speech, Images Videos—ICPR Contests*, Istanbul, Turkey, 2010, pp. 180–189.

[34] C. Wu, B. Clipp, X. Li, J. Frahm, and M. Pollefeys, "3d model matching with viewpoint-invariant patches (VIP)," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, Anchorage, AK, 2008, pp. 1–8.

[35] 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* (Lecture Notes in Computer Science, vol. 6388). Berlin Heidelberg, Germany: Springer, 2010, pp. 200–209.

[36] A. Pronobis, M. Fornoni, H. I. Christensen, and B. Caputo, "The robot vision track at ImageCLEF 2010," in *Proc. CLEF LABs Workshops, Notebook Papers*, Padua, Italy, 2010.

[37] O. Saurer, F. Fraundorfer, and M. Pollefeys, "Visual localization using global visual features and vanishing points," in *Proc. CLEF LABs Workshops, Notebook Papers*, Padua, Italy, 2010.

[38] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[39] M. Fornoni, J. Martínez-Gómez, and B. Caputo, "A multi cue discriminative approach to semantic place classification," in *Proc. CLEF LABs Workshops, Notebook Papers*, Padua, Italy, 2010.

[40] A. Bosch, A. Zisserman, and X. Munoz, "Representing shape with a spatial pyramid kernel," in *Proc. 6th ACM Int. Conf. Image Video Retrieval*, New York, 2007, pp. 401–408.

[41] T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Computer Vision-ECCV 2000*. Berlin Heidelberg, Germany: Springer, 2000, pp. 404–420.

[42] J. Martínez-Gómez, I. García-Varea, and B. Caputo, "Baseline multi-modal place classifier for the 2012 robot vision task," in *Proc. Working Notes ImageCLEF Laboratory*, 2012, pp. 1–10.

[43] B. Steder, R. B. Rusu, K. Konolige, and W. Burgard, "NARF: 3D range image features for object recognition," in *Proc. Workshop Defining Solving Realistic Perception Problems Personal Robotics IEEE/RSJ Int. Conf. Intelligent Robots Systems*, Taipei, Taiwan, 2010, pp. 1–2.

[44] F. Orabona, L. Jie, and B. Caputo, "Online-batch strongly convex multi kernel learning," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2010, pp. 787–794.

[45] J. Redolfi and J. Sánchez, "Leveraging robust signatures for mobile robot semantic localization," in *Proc. Working Notes ImageCLEF Laboratory*, 2012, pp. 1–11.

[46] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[47] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Int. Conf. Computer Vision Pattern Recognition*, 2007, pp. 1–8.

[48] E. Boros, A.-L. Gînsca, and A. Iftene, "UAIC participation at robot vision@ 2012-an updated vision," in *Proc. Working Notes ImageCLEF Laboratory*, 2012, pp. 1–12.

[49] K. Zhagorina and A. Buslavyev, "Computer analysis of visual image similarity," in *Proc. Working Notes ImageCLEF Laboratory*, 2012, pp. 1–8.

[50] R. Xu, S. Jiang, X. Song, S. Wang, Y. Xie, F. Wang, and X. Lv, "MIAR ICT participation at robot vision 2013," in *Proc. Working Notes ImageCLEF Laboratory*, 2013, pp. 1–11.

[51] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in *Proc. Advances Neural Information Processing Systems*, 2010, pp. 244–252.

[52] A. Ksibi, B. Ammar, A. B. Ammar, C. B. Amar, and A. M. Alimi, "Regimrobvid: Objects and scenes detection for robot vision 2013," in *Proc. Working Notes ImageCLEF Laboratory*, 2013, pp. 1–6.

[53] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in *Proc. IEEE 11th Int. Conf. Computer Vision*, 2007, pp. 1–8.

[54] Y. Feng, M. Halvey, and J. M. Jose, "University of glasgow at ImageCLEF 2009 robot vision task: A rule based approach," in *Multilingual Information Access Evaluation II. Multimedia Experiments*. Berlin Heidelberg, Germany: Springer, 2010, pp. 295–298.

[55] Y. Feng, J. Ren, J. Jiang, M. Halvey, and J. M. Jose, "Effective venue image retrieval using robust feature extraction and model constrained matching for mobile robot localization," *Mach. Vis. Applicat.*, vol. 23, no. 5, pp. 1011–1027, 2012.

[56] J. Laaksonen, M. Koskela, S. Laakso, and E. Oja, "PicSOM—Content-based image retrieval with self-organizing maps," *Pattern Recognit. Lett.*, vol. 21, nos. 13–14, pp. 1199–1207, 2000.

[57] M. Sjöberg, M. Koskela, V. Viitaniemi, and J. Laaksonen, "PicSOM experiments in ImageCLEF robot vision," in *Recognizing Patterns in Signals, Speech, Images and Videos*. Berlin Heidelberg, Germany: Springer, 2010, pp. 190–199.

[58] M. Sjöberg, M. Koskela, S. Ishikawa, and J. Laaksonen, "Large-scale visual concept detection with explicit kernel maps and power mean SVM," in *Proc. 3rd ACM Conf. Int. Multimedia Retrieval*, ACM, 2013, pp. 239–246.

[59] M. Pollefeys, J.-M. Frahm, F. Fraundorfer, C. Zach, C. Wu, B. Clipp, and D. Gallup, "Towards large-scale visual mapping and localization," in *Robotics Research*. Berlin Heidelberg, Germany: Springer, 2011, pp. 535–555.

[60] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, 2013.

[61] X. Ren, L. Bo, and D. Fox, "RGB-(D) scene labeling: Features and algorithms," in *Proc. IEEE Conf. Computer Vision Pattern Recognition*, 2012, pp. 2759–2766.

Jesus Martínez-Gómez, University of Castilla-La Mancha, Spain. E-mail: jesus_martinez@dsi.uclm.es.

Barbara Caputo, University of Rome La Sapienza, Italy. E-mail: bcaputo@idiap.ch.

Miguel Cazorla, University of Alicante, Spain. E-mail: miguel@dccia.ua.es.

Henrik I. Christensen, Georgia Institute of Technology, Atlanta, United States. E-mail: hic@cc.gatech.edu; hichristensen@gmail.com.

Marco Fornoni, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland. E-mail: Marco.FORNONI@idiap.ch.

Ismael García-Varea, University of Castilla-La Mancha, Spain. E-mail: Ismael.Garcia@uclm.es.

Andrzej Pronobis, University of Washington, Seattle, United States. E-mail: pronobis@cs.washington.edu.