Overview of the ImageCLEF 2012 Robot Vision Task

Jesus Martinez-Gomez[1], Ismael Garcia-Varea[1], and Barbara Caputo[2] **

[1] University of Castilla-La Mancha
Albacete, Spain
[2] Idiap Research Institute, Centre Du Parc, Rue Marconi 19
P.O. Box 592, CH-1920 Martigny, Switzerland
[1] {Jesus.Martinez, Ismael.Garcia} @uclm.es
[2]bcaputo@idiap.ch

Abstract. This article describes the RobotVision@ImageCLEF 2012 challenge, which addresses the problem of multimodal place classification. Participants of the challenge were asked to classify rooms on the basis of image sequences captured by cameras mounted on a mobile robot. The proposals of the participants had to answer the question "where are you?" (I am in the elevator, in the toilet, etc) when presented with a test sequence, acquired within the same building and floor but with different lighting conditions than the training sequence. The 2012 edition of the challenge introduced the use of depth images in addition to visual images. Moreover, several techniques for feature extraction and cue integration were also proposed. As in previous editions, two different tasks were proposed: task 1 and task 2. In task 1 (mandatory) participants were asked to classify the frames separately, while the temporal continuity of the image sequence could only be exploited in task 2 (optional). Eight different groups participated to the 2012 edition of the Robot Vision challenge. The winner in both tasks was the Centro de Investigación en Informática para la Ingeniería (CIII), from the Universidad Tecnológica Nacional, Argentina (CIII UTN FRC). This participant obtained an overall score of 2071 (84.70% of the maximum score) in task 1 and 3930 (96.35% of the maximum score) in task 2.

1 Introduction

The ImageCLEF 2012 Robot Vision challenge has been the fourth edition of a competition [10] that started in 2009 within the ImageCLEF [1] as part of the Cross Lange Evaluation Forum (CLEF) Initiative [2]. Since its origin, the Robot

** This work was supported by the SNSF project MULTI (B. C.), and by the European Social Fund (FEDER), the Spanish Ministerio de Ciencia e Innovacion (MICINN), and the Spanish "Junta de Comunidades de Castilla-La Mancha" (MIPRCV Consolider Ingenio 2010 CSD2007-00018, TIN2010-20900-C04-03, PBI08-0210-7127 and PPII11-0309-6935 projects, J. M.-G. and I. G.-V.)

[1] http://imageclef.org/
[2] http://www.clef-initiative.eu//

Vision task has been addressing the problem of place classification for mobile robot localization.

The 2009@ImageCLEF edition of the task [9], with 7 participating groups, defined some details that have been maintained for all the following editions. Participants were given training data consisting of sequences of frames recorded in indoor environments. These training frames were labelled with the name of the rooms they were acquired from. The task consisted on building a system capable to classify test frames using as class the name of the rooms previously seen. Moreover, the system could refrain from making a decision in the case of lack of confidence. Two different subtasks were then proposed: obligatory and optional. The difference between both subtasks was that the temporal continuity of the test sequence could only be exploited in the optional task. The score for each participant submission was computed as the sum of the frames that were correctly labelled minus a penalty that was applied to the frames that were misclassified. No penalties were applied for frames not classified.

In 2010, two editions of the challenge took place. The second edition of the task, 2010@ICPR [7] was held in conjunction with ICPR 2010. 9 groups participated to this edition, which introduced the use of stereo images and two types of different training sequences (easy and hard) that had to be used separately. The 2010@ImageCLEF edition [8], with 7 participating groups, was focused on generalization: several areas could belong to the same semantic category.

Several changes have been proposed for the ImageCLEF 2012 Robot Vision task. Firstly, stereo images have been replaced by images acquired using two types of camera: a perspective camera for visual images and a depth camera (the Microsoft Kinect sensor) for range images. Therefore, each frame consists of two types of images and the challenge is focused on the problem of multimodal place classification. In addition to the use of depth images, the optional task contains kidnappings and no unknown rooms appear in the test sequences. Moreover, several techniques for features extraction and cue integration have been proposed to the participants.

We received a total of 23 runs from 8 different groups. 18 runs were submitted to the task 1 (mandatory) and 5 to the task 2 (optional). The best result in both tasks was obtained by the Centro de Investigación en Informática para la Ingeniería (CIII), from the Universidad Tecnológica Nacional, Argentina (CIII UTN FRC).

The rest of the paper details the challenge and is organized as follows: Section 2 describes the 2012 ImageCLEF edition of the RobotVision task. Section 3 presents all the participants groups, while the results are reported in Section 4. Finally, in Section 5, conclusions are drawn and future work is outlined.

2 The RobotVision Task

This section describes the details concerning the setup of the ImageCLEF 2012 Robot Vision task. Section 2 gives a description of the task. Section 2.1 describes all the sequences of frames provided for training and test while the two subtasks are explained in Section 2.2. The performance evaluation is detailed in Section 2.3 and finally, Section 2.4 describes the information provided by the organizers.

2.1 Description

The fourth edition of the Robot Vision challenge was focused on the problem of multi-modal place classification. Participants were asked to classify functional areas on the basis of image sequences, captured by a perspective camera and a Kinect mounted on a mobile robot (see Fig. 1) within an office environment.

Fig. 1. Mobile robot platform used for data acquisition.

Participants had available visual images and range images that could be used to generate 3D point cloud files. The difference between visual images, range images and 3D point cloud files can be observed in Figure 2. Training and test sequences were acquired within the same building and floor but with some variations in the lighting conditions or the acquisition procedure (clockwise and counter clockwise).

Two different tasks were considered in the Robot Vision challenge: task 1 and task 2. For both tasks, participants should be able to answer the question "where are you?" when presented with a test sequence imaging a room category already seen during training. The difference between both tasks was the presence

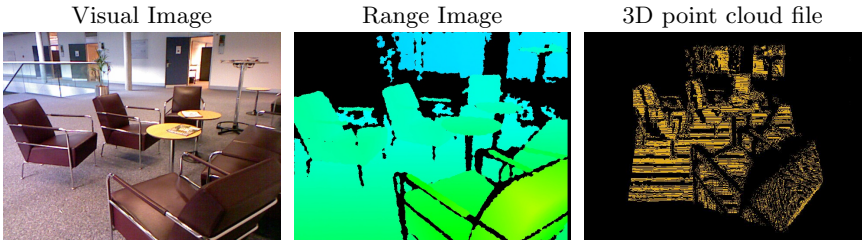

Fig. 2. Visual, depth and 3D point cloud files.

(or lack) of kidnappings in the final test sequence, and the availability on the use of the temporal continuity of the sequence.

The kidnapping (only task 2) is affected by the robot changing room. Room changes in sequences without kidnappings were usually represented by a small number of images showing a smooth transition. On the other side, room changes with kidnappings were represented by a drastic change for frames, as can be observed in Figure 3.

2.2 The Data

Training and validation sequences consisted of a set of the Robot Vision VIDA dataset. VIDA is a dataset with images acquired within an indoor environment using a robot platform in the IDIAP research building. This dataset contains sequences of several rooms belonging to different room categories such as "Corridor" or "Toilet". Sequences were acquired using two cameras: a perspective visual camera and a 3D range laser sensor Kinect camera. Therefore, there are

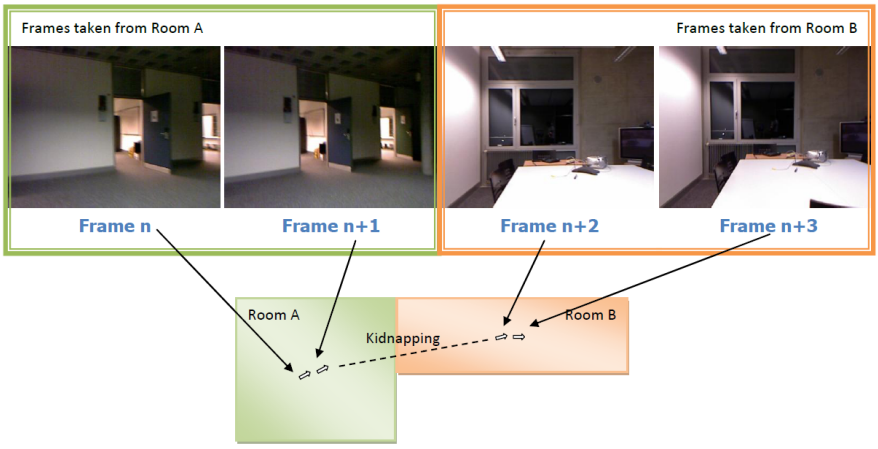

Fig. 3. Example of kidnapping.

Table 1. Distribution of room categories for the training sequences.

	Number of frames		
Room Category	Training 1	Training 2	Training 3
Corridor	438	498	444
Elevator Area	140	152	84
Printer Room	119	80	65
Lounge Area	421	452	376
Professor Office	408	336	247
Student Office	664	599	388
Visio Conference	126	79	60
Technical Room	153	96	118
Toilet	198	240	131
All	2667	2532	1913

two different types of images: RGB images and depth images.

Three different sequences of frames were provided for training and two additional ones for the final experiment. All training frames were labelled with the name of the room they were acquired from. There were 9 different categories of rooms, and the distribution for the training sequences can be observed in Table 1

The difference between all the room categories can be observed in Figure 4, where an exemplar visual image for each one of the 9 room categories is shown.

Fig. 4. Examples of images from the Robot Vision 2012 database.

2.3 Subtasks

All the participants of the ImageCLEF 2012 Robot Vision task were allowed to submit their runs to two different subtasks: task1 and task2.

Task 1 This task was mandatory and the test sequence had to be classified without using the temporal continuity of the sequence. Therefore, the order of the test frames cannot be taken into account. Moreover, there were not kidnappings in the final test sequence.

Task 2 This task was optional and participants could take advantage of the temporal continuity of the test sequence. There were kidnapping in the final test sequences that allowed participants to obtain additional points when they were managed correctly

2.4 Performance Evaluation

The proposals of the participants were compared using the score obtained by their submissions. These submissions were the classes or room categories assigned to the frames of the test sequences, and the score was computed using the rules that are shown in Table 2. Due to wrong classifications obtaining negative points, participants were allowed to not classify test frames.

Table 2. Rules used to calculate the final score for a run.

Each correctly classified frame	+1 points
Each misclassified frame	-1 points
Each frame that was not classified	+0 points
(Task 2) All the 4 frames correctly classified after a kidnapping	+1 points (additional)

2.5 Additional information provided by the organization

We proposed the use of several techniques for features extraction (PHOG and NARF) and cue integration (OBSCURE). Thanks to the use of these techniques, participants could focus on the development of new features while using the proposed method for cue integration or vice versa. We also provided information as the point cloud library [11] and a basic technique for taking advantage of the temporal continuity[3]. In order to evaluate this information, we submitted two runs (task 1 and task 2) that were obtained using only the provided techniques. The results obtained with such proposal [3] can be considered as baseline results that all the groups were expected to improve.

[3] http://imageclef.org/2012/robot

Visual Features PHOG features are histogram-based global features that combine structural and statistical approaches. Other descriptors similar to PHOG that could also be used are: Sift-based Pyramid Histogram Of visual Words (PHOW) [1], Pyramid histogram of Local Binary Patterns (PLBP) [4], Self-Similarity-based PHOW (SS-PHOW) [12], and Compose Receptive Field Histogram (CRFH) [2].

Depth Features NARF features is a novel descriptor technique that has been included in the point cloud library [11]. The number of descriptors that can be extracted from a range image is not fixed, in the same manner as SIFT points.

Cue Integration The algorithm proposed for cue integration was the Online-Batch Strongly Convex mUlti keRnel lEarning (OBSCURE) [6]. This SVM-based multiclass learning algorithm obtains state-of-the-art performance in a considerably lower training time. Other algorithm that could be used was the Online Independent Support Vector Machines [5] that, in comparison with SVM, dramatically reduces learning time and space requirements at the price of a negligible loss in accuracy.

3 Participation

In 2012, 43 groups registered to the Robot Vision task but only 8 submitted, at least, one run, namely:

- CIII UTN FRC: Universidad Tecnológica Nacional, Córdoba, Argentina.
- NUDT: National University of Defense Technology, Changsha, China.
- UAIC2012: Alexandru Ioan Cuza University, Iasi, Romania.
- USUroom409: Ural Federal University, Yekaterinburg, Russian Federation.
- SKB Kontur Labs: Kontur Labs, Yekaterinburg, Russian Federation.
- CBIRITU: Istanbul Technical University, Istanbul, Turkey.
- SIMD: University of Castilla-La Mancha, Albacete, Spain.
- BuffaloVision: University at Buffalo, New York, United States.

A total of 23 runs were submitted, with 18 runs submitted to the task 1 (mandatory) and 5 runs submitted to the task 2 (optional). The limit to the number of runs that could be submitted was 3.

4 Results

This section presents the results of the Robot Vision task of ImageCLEF 2012 for the two subtasks: task1 and task 2.

Table 3. Ranking of the runs submitted by the groups for the Task 1.

Rank	Group Name	Score	% Max. Score
1	CIII UTN FRC	2071	84.70
2	NUDT	1817	74.31
3	NUDT	1729	70.72
4	UAIC2012	1348	55.13
5	USUroom409	1225	50.10
6	USUroom409	1225	50.10
7	USUroom409	1193	48.79
8	UAIC2012	1049	42.90
9	UAIC2012	1049	42.90
10	SKB Kontur Labs	1028	42.04
11	SKB Kontur Labs	1006	41.15
12	SKB Kontur Labs	997	40.78
13	CBIRITU	551	22.54
14	CBIRITU	542	22.17
15	SIMD/IDIAP (baseline results)	462	18.86
16	BuffaloVision	-70	<0.00
17	BuffaloVision	-110	<0.00
18	BuffaloVision	-234	<0.00

4.1 Task 1

Eight different groups submitted runs for the task 1 as can be observed in Table 3. The maximum score that could be achieved was 2445 and the winner (CII UTN FRC) obtained 2071 points. CII UTN FRC and NUDT teams ranked first and second respectively and their score was higher than 70% of the maximum score.

The score obtained by the SIMD/IDIAP team could be considered as a baseline result that all the groups were expected to improve. Such score was obtained using the techniques provided by the organizers without new contributions. As it was expected, 6 out of 7 teams obtained higher scores. The results are summarized in Figure 5, where only the best run for each team has been considered.

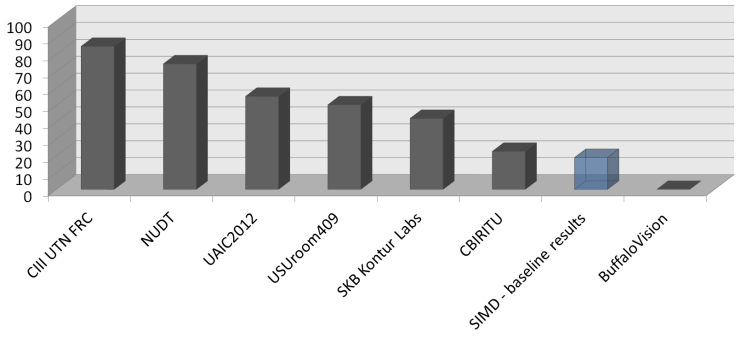

Fig. 5. Results obtained for the Task 1 as % of the maximum score

Table 4. Ranking of the runs submitted by the groups for the Task 2.

Rank	Group Name	Score	% Max. Score
1	CIII UTN FRC	3930	96.35
1	CIII UTN FRC	3925	96.22
2	NUDT	3859	94.61
3	CBIRITU	3169	77.69
4	SIMD/IDIAP (baseline results)	1041	25.52

4.2 Task 2

For the optional task, the maximum score was 4079 and only 4 groups submitted runs. The winner for the task 2 was the CIII UTN FRC group with 3930 points, only 71 more than the NUDT group, which ranked second. All the results can be seen in Table 4.

In view of these results, it should be remarked the high quality of the participant proposals, due to the score obtained by CIII UTN FRC, NUDT and CBIRITU groups was higher than 75% of the maximum score. All the results for task 2 are summarized in Figure 6, where only the best run for each team has been considered.

5 Conclusions and Future Work

We have described in this article the fourth edition of the Robot Vision task at ImageCLEF 2012, which attracted a considerable attention with 8 groups submitting runs. There are 2 main conclusions that can be drawn from the proposals: (i) despite of two types of images were provided (visual and range), depth features were not commonly used, and (ii) most of the proposals were based on the use of SVMs. We plan to continue the task in the next years with new challenges related to place categorization. Concretely, we have plans to introduce object categorization in the following editions.

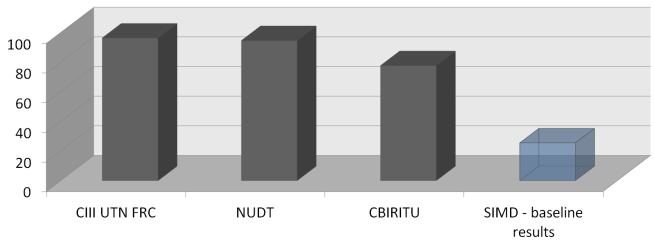

Fig. 6. Results obtained for the Task 2 as % of the maximum score

References

1. A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *International Conference on Computer Vision*, pages 1–8. Citeseer, 2007.
2. O. Linde and T. Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proc. ICPR*. Citeseer, 2004.
3. Jesus Martinez-Gomez, Ismael Garcia-Varea, and Barbara Caputo. Baseline multimodal place classifier for the 2012 robot vision task. In *CLEF 2012 working notes*, 2012.
4. T. Ojala, M. Pietikäinen, and T. Mäenpää. Gray scale and rotation invariant texture classification with local binary patterns. *Computer Vision-ECCV 2000*, pages 404–420, 2000.
5. F. Orabona, C. Castellini, B. Caputo, J. Luo, and G. Sandini. Indoor place recognition using online independent support vector machines. In *Proc. BMVC*, volume 7, 2007.
6. F. Orabona, L. Jie, , and B. Caputo. Online-Batch Strongly Convex Multi Kernel Learning. In *Proc. of Computer Vision and Pattern Recognition, CVPR*, 2010.
7. A. Pronobis, H. Christensen, and B. Caputo. Overview of the imageclef@ icpr 2010 robot vision track. *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 171–179, 2010.
8. A. Pronobis, M. Fornoni, HI Christensesn, and B. Caputo. The robot vision track at imageclef 2010. *Working Notes of ImageCLEF*, 2010, 2010.
9. A. Pronobis, L. Xing, and B. Caputo. Overview of the clef 2009 robot vision track. pages 110–119. Springer, 2010.
10. Andrzej Pronobis and Barbara Caputo. The robot vision task. In Henning Muller, Paul Clough, Thomas Deselaers, and Barbara Caputo, editors, *ImageCLEF*, volume 32 of *The Information Retrieval Series*, pages 185–198. Springer Berlin Heidelberg, 2010.
11. R.B. Rusu and S. Cousins. 3d is here: Point cloud library (pcl). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1–4. IEEE, 2011.
12. E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07*, pages 1–8, 2007.