Baseline Multimodal Place Classifier for the 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. The objective of this article is reporting the participation of the SIMD-IDIAP group in the RobotVision@ImageCLEF 2012 challenge. This challenge addresses the problem of multimodal place classification, and the 2012 edition has been organized by the members of the SIMD-IDIAP team. One of the main novelties in the 2012 edition of the task has been the proposal of several techniques for features extraction and cue integration. This paper details how to use all these techniques for developing a multimodal place classifier and describes the results obtained with it. Our approach ranked 7th for task 1 and 4th for task 2. The complete results for all the participants of the 2012 RobotVision task are also reported in this article.

1 Introduction

This article describes the participation of the SIMD-IDIAP team at the fourth edition of the Robot Vision task [7]. This competition addresses the problem of multimodal place localization for mobile robots in indoor environments. Since its release in 2009, the information provided by the organizers consisted on visual images acquired with a mobile robot while moving within indoor environments. However, the 2012 edition has introduced the use of range images as well as proposes useful techniques for the development of the approaches.
The SIMD-IDIAP team has developed a multimodal place classifier based on the use of Support Vector Machines (SVMs) [13]. The classifier has been trained using a combination of visual and depth features extracted from sequences of images. The selected features, as well as the classifier, are those proposed by the organizers of the task, members of the SIMD-IDIAP team. Therefore, the results

** 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.)

achieved by the method presented in this article can be considered as baseline results that any participant is expected to improve.
The rest of the paper is organized as follows: Section 2 describes the 2012 edition of the RobotVision task. Section 3 gives an overview of the SIMD-IDIAP proposal, while the feature extraction and classification techniques are described in Section 4 and Section 5 respectively. We report the results obtained in Section 6, and finally, in Section 7, conclusions are drawn and future work are outlined.

2 The RobotVision Task

2.1 Description

The fourth edition of the RobotVision challenge is focused on the problem of multi-modal place classification. Participants are 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 have available visual images and range images that can be used to generate 3D cloud point files. The difference between visual images, range images and 3D point cloud files can be observed in Figure 2. Training and test sequences have been 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 are considered in the RobotVision 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

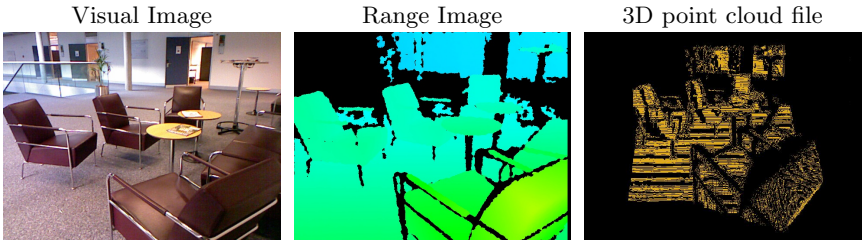

Fig. 2. Visual, depth and 3D point cloud files.

seen during training. The difference between both tasks is the presence (or lack) of kidnappings in the final test sequence, and the availability on the use of the temporal continuity of the sequence.

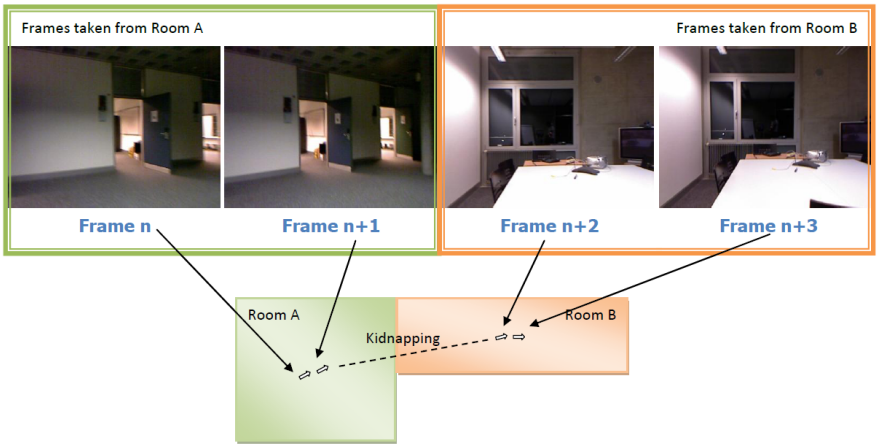

Fig. 3. Example of kidnapping.

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

The Data Three different sequences of frames are provided for training and two additional ones for the final experiment. All training frames are labelled with the name of the room they were acquired from. There are 9 different categories of rooms:

– Corridor

- Elevator Area
- Printer Room
- Lounge Area
- Professor Office
- Student Office
- Visio Conference
- Technical Room
- Toilet

Figure 4 shows an exemplar visual image for each one of the 9 room categories.

Fig. 4. Examples of images from the RobotVision 2012 database.

Task 1 This task is mandatory and the test sequence has 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 are not kidnappings in the final test sequence.

Task 2 This task is optional and participants can take advantage of the temporal continuity of the test sequence. There are kidnapping in the final test sequences that allow participants to obtain additional points when they are managed correctly

Performance Evaluation The proposals of the participants are compared using the score obtained by their submissions. These submissions are the classes or room categories assigned to the frames of the test sequences, and the score is calculated using the rules that are shown in the Table 1. Due to wrong classifications obtain negative points, participants are allowed to not classify test frames.

Table 1. 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)

2.2 Information provided by the organizers

The organizers of the 2012 RobotVision task propose the use of several techniques for features extraction and cue integration (classifier). Thanks to the use of these techniques, participants can focus on the development of new features while using the proposed method for cue integration or vice versa.
The organizers also provide information as the point cloud library [11] and a basic technique for taking advantage of the temporal continuity[1]. All these techniques have been used for generating the multimodal place classifier proposed in this article and are explained in Section 4 and Section 5.

3 Overall Description

The SIMD-IDIAP proposal for the 2012 RobotVision task can be split into two steps: training and classification. The first step is performed by generating a SVM classifier [13] with a combination of features extracted from visual and depth images. We opted for the visual and depth features proposed by the task organizers and a SVM classifier. Both features are then concatenated for generating a single feature. The complete training process is shown in Figure 5.

The second step corresponds with the classification of test frames. Before classifying a test frame, it is necessary to extract the same features extracted in the training step. After that, the features are classified using the SVM previously generated, which obtains a decision margin for each class. All these margins are then processed to decide whether to classify the frame or not, depending on the level of confidence of the decision. Low confidence frames will not be classified in order to avoid obtaining negative points. The complete process for classification and post-processing will be explained in Section 5.

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

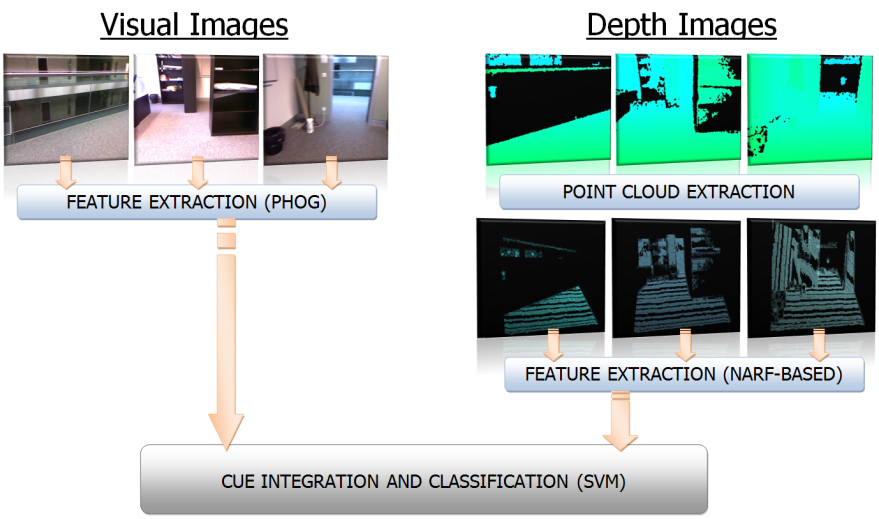

Fig. 5. Overall overview of the training step for generating the classifier.

4 Feature Extraction

As features to extract from the sequences of frames, we have chosen the Pyramid Histogram of Orientated Gradients (PHOG) [2] and the Normal Aligned Radial Feature (NARF). These descriptors can be extracted from visual images and 3D point cloud files respectively, and they are the features proposed by the task organizers.
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) [8], Self-Similarity-based PHOW (SS-PHOW) [12], and Compose Receptive Field Histogram (CRFH) [5].
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. In order to extract descriptors with the same length, we have computed a new feature from the NARF points extracted from a 3D point cloud file.

4.1 Visual Features - PHOG

The Pyramid Histogram of Orientated Gradients (PHOG) is inspired by the pyramid representation presented in [4] and the Histogram of Gradient Orientation (HOG) described in [3]. This descriptor consists then of a histogram of orientation gradients over each image sub-region at each level. It represents local image shape and its spatial layout, both together with a spatial pyramid kernel. Two parameters should be fixed when extracting a PHOG descriptor from a

visual image: the number of levels of the pyramid and the number of bins of the histogram. Each bin represents the number of edges with orientations within a certain angular range. In our approach, we opted for 20 bins, 3 levels (0, 1 and 2) and the range of orientations [0,360]. Using these parameters, for each visual image we obtain a 420 bytes descriptor.

4.2 Depth Features - NARF

The process to generate depth features from range images consists of three steps: a) convert the range image into a 3D point cloud file, b) extract NARF features from keypoints, and c) compute a descriptor pyramidly from the NARF features.
The first step has been done by using a pyhton script provided by the task organizers. This script has been used due to the specific format of the RobotVision@ImageCLEF 2012 range images, but the step can be skipped when using the PCL software to register point cloud files directly from the kinect device.
The second step extracts NARF features from the keypoints detected in a point cloud file. The keypoints are detected by using the neighbour cloud points and taking into account the borders. For each keypoint, a 36 bytes NARF feature is extracted, and we also store the ¡x,y,z¿ position of the 3D point.
For the third step, we compute pyramidly a descriptor from the data generated in the previous step. This is done by following the pyramid representation presented in [4]. In a similar way as for the PHOG descriptor, we have to fix the number of bins of the histogram and the number of levels. We selected 100 bins and 2 levels (0 and 1), obtaining a 500 bytes descriptor.
Both PHOG and NARF-based descriptors are directly concatenated to obtain a 920 bytes descriptor by frame.

5 Classification and post-processing

5.1 Classifier

The algorithm that was proposed by the organizers for cue integration was the Online-Batch Strongly Convex mUlti keRnel lEarning (OBSCURE) [10]. In this work, we have used a classifier similar to OBSCURE based on the use of Online Independent Support Vector Machines (OI-SVM) [9]. This classifier is named Memory Controlled OI-SVM [6] and allows to keep under control the memory growth while the algorithm learns incrementally. This is done by applying a forgetting strategy over the stored Training Samples (TSs), while preserving the stored Support Vectors and it approximates reasonably well the original optimal solution.
The Memory Controlled OI-SVM is trained using the combination of visual and depth features described in the previous section. When a test frame is classified, the classifier obtains a decision margin for each class, and this output has to be processed to obtain the most feasible class.

5.2 Low confidence detection

In order to detect low confidence classifications, we process the obtained outputs in the following way: (i) we normalize the margins by dividing all values by the maximum margin, and (ii) we test whether the normalized outputs pass two conditions. On the basis of the output margins $M_{n\,i=1}^{i\,C}$, with C= number of classes, for each frame n, the two conditions used to detect challenging frames are:

1. $M_n^i < M_{max}{}_{i=1}^{C}$: for each of the possible classes C none obtains a high level of confidence;
2. $|M_n^i - M_n^j| < \Delta_{i=1}^{C}$: there are at least two classes with high level of confidence, but the difference between them is too small to allow for a confident decision.

If a frame has been detected as challenging, the algorithm will not assign them any room category and it will not be classified. In all the experiments, we have used $M_{max} = 0.2$ and $\Delta = 0.2$.

5.3 Temporal Continuity

For the task 2, the temporal continuity of the sequence can be taken into account. In order to take advantage of this situation, we used the solution proposed by the task organizers: detect prior classes and use them to classify challenging frames. Once a frame has been identified as challenging (and not classified), we use the classification results obtained for the last n frames to solve the ambiguity: if all the last n frames have been assigned to the same class C_i, then we can conclude that all frames come from the same class C_i, we consider it a prior class, and the label will be assigned accordingly. We have used $n = 5$ for all the experiments.

6 Results

We only submitted two runs: one for the task 1 and one for the task 2. The process for generating both runs includes all the techniques that have been explained above.
We extracted a PHOG and NARF-based feature descriptor for each one of the training and test frames. Both descriptors were concatenated to generate a 920 bytes descriptor and then, we trained a Memory Controlled OI-SVM using the training1, training2, and training3 sequences. Finally, we classified the test sequences (test 1 for task 1 and test 2 for task 2) using the MC-OI-SVM.
For both tasks, we post-processed the output generated with the classifier to detect challenging frames. These challenging frames were not labelled. For task 2, we used the temporal continuity of the sequence for classifying the challenging fames when possible (a prior class using the last 5 frames has been detected).

6.1 Task 1

Eight different groups submitted runs for the task 1 and all the results can be observed in Table 2. The maximum score that could be achieved was 2445 and the best group (CII UTN FRC) obtained 2071 points. Our proposal ranked 7th with 462 points and most of the teams, as expected, achieved better than us. The submitted run classified only 1526 frames (37.5% of the test frames were detected as challenging), with 994 hits and 532 fails.

Table 2. Ranking of the runs submitted by the groups for the task 1.

Rank	Group Name	Best Score
1	CIII UTN FRC	2071
2	NUDT	1817
3	UAIC2012	1348
4	USUroom409	1225
5	SKB Kontur Labs	1028
6	CBIRITU	551
7	SIMD-IDIAP - baseline results	462
8	BuffaloVision	-70

6.2 Task 2

For the optional task, the maximum score was 4079 and only 4 groups submitted runs. Our submission ranked 4th with 1041 points and the winner for the task 2 was the CIII UTN FRC group with 3930 points. All the results can be seen in Table 3. In this task, our algorithm discarded 1205 challenging frames but 97 of them were classified using a prior class detected with the temporal continuity. Concretely, the final submission consisted of 2915 frames classified and 1108 (27.54%) frames that were not labelled.

Table 3. Ranking of the runs submitted by the groups for the task 2.

Rank	Group Name	Best Score
1	CIII UTN FRC	3930
2	NUDT	3925
3	CBIRITU	3169
4	SIMD-IDIAP - baseline results	1041

7 Conclusions and Future Work

This article describes the participation of the SIMD-IDIAP group to the RobotVision task at ImageCLEF 2012. We developed an approach using the techniques

for feature extraction, cue integration, and temporal continuity proposed by the task organizers.
We submitted runs for task 1 (mandatory) and task 2 (optional) using the information extracted from visual and depth images. Due to such runs were generated using the techniques proposed by the organizers, these results can be considered as baseline results.
Our best runs in the two tracks ranked respectively seventh (task 1) and fourth (task 2), showing that most of the teams ranked better than the baseline results. For future work, we have in mind to improve the NARF-based descriptor introduced in this article. We also have plans to evaluate the use of larger descriptors, which can be done by using a higher number of levels for the pyramid and bins for the histogram.

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. 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*, page 408. ACM, 2007.
3. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. Ieee, 2005.
4. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, 2006.
5. O. Linde and T. Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proc. ICPR*. Citeseer, 2004.
6. J. Martinez and B. Caputo. Towards semi-supervised learning of semantic spatial concepts for mobile robots. *Journal of Physical Agents*, 4(3):19–31, 2010.
7. Jesus Martinez-Gomez, Ismael Garcia-Varea, and Barbara Caputo. Overview of the imageclef 2012 robot vision task. In *CLEF 2012 working notes*, 2012.
8. 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.
9. 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.
10. F. Orabona, L. Jie, , and B. Caputo. Online-Batch Strongly Convex Multi Kernel Learning. In *Proc. of Computer Vision and Pattern Recognition, CVPR*, 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.
13. V.N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York Inc, 2000.