Combination of Classifiers for Indoor Room Recognition

CGS participation at ImageCLEF2010 Robot Vision Task

Walter Lucetti Emanuel Luchetti

Gustavo Stefanini Advanced Robotics Research Center
Scuola Superiore di Studi e Perfezionamento Sant'Anna
w.lucetti@sssup.it e.luchetti@sssup.it

Abstract. This paper represents a description of our approach to the problem of topological localization of a mobile robot using visual information. Our method has been developed for ImageCLEF 2010 Robot Vision Task challenge. The challenge was focused on the problem of visual place classification, with a special focus on generalization. The goal was to recognize rooms by the images captured with a stereo camera mounted on a mobile robot within an office environment. Algorithms should be able to reply to question "Where are you?", saying "I don't know" if the room analyzed was not presented during training phase. For the challenge three sequences were given: Training Set, Validation Set and Test Set acquired on three different floors of the same building. We chose to approach the challenge realizing a multi-Level machine learning architecture, made of a first "weak" classifiers Level based on visual features extracted from images and of a second Level performing fusion of first Level outputs. We developed four configurations to determine the best approach to problem solving: **C**ommittees **o**f **E**xperts, **S**tacked **R**egression with **S**upport **V**ector **M**achines stage, **S**tacked **R**egression with **A**rtificial **N**eural **N**etwork stage, Weighted Linear Combination of all the three previous methods. Finally the result of twenty RUNs, five RUNs for each of the four different system configurations, were submitted at ImageCLEF challenge.

Keywords: ImageCLEF, Visual Place Classification, Features Extraction, Support Vector Machines, Bayes Classifier, Artificial Neural Network, Committees of Experts, Stacked Regression, Stacked Generalization, Stereo Vision, Hough Transform, Discrete Fourier Transform.

1 Introduction

ImageCLEF[1] hosted in 2010 the third edition of Robot Vision Challenge. The task addressed the problem of Visual Place Classification, this edition with a special focus on generalization [1].

We chose to approach the challenge using Classifiers Combination method [2], after the analysis of the provided training set images. Together to Training Set a Validation Set was released, Validation Set consisted of 2069 image couples acquired on a different floor of the same building used to create the training set, but only seven known rooms was visited and three "*Unknown*" rooms were added to be able to test "Unknown" recognition techniques.

Finally after about one month from Training Set release, a final Test Set was released. It consisted of 2741 unlabeled image couples acquired to a third floor of the same building. Analogue rooms were visited and other "Unknown" rooms were presented. We performed twenty different RUNs on test set: five RUNs for each of the four different Classifiers Combination methods we developed.

For each submitted RUN was given a score as described in [1].

[1] ImageCLEF: http://www.imageclef.org

2 System Architecture

The approach we chose to follow is illustrated in **Fig. 1**: each stereo image couple is processed to extract different kind of visual features that are analyzed by a first stage of classifier (*Level-0*) to produce a set of label. The set of label produced by Level-0 becomes the input for the second stage (*Level-1*) that produces a set of values indicating the *Confidence Level* of every possible output class. A final stage analyzes Level-1 outputs and gives in output the final reply.

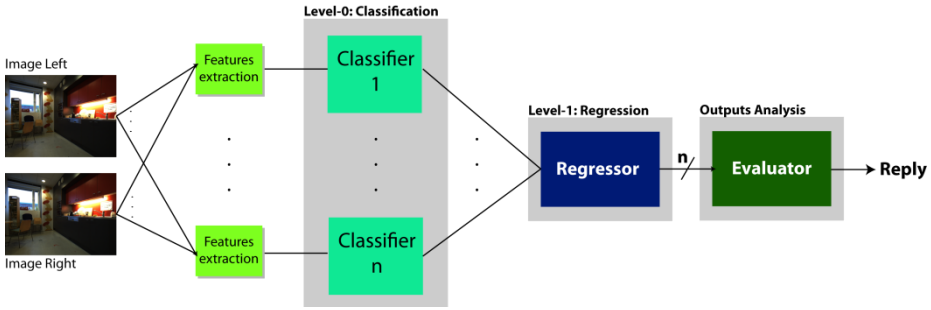

Fig. 1. Combination of Classifiers scheme

Motivation

Many method of Image Classification using machine learning techniques were implemented using a single type of features extracted from single image or Stereo Images couple. The classification made on one type of features works well only on a well stated kind of environment (i.e. texture features extraction is right for highly repetitive images, but is not informative for homogeneous images). Combining different kind of *Level-0* classifiers allows to choose the classifier that gives the correct reply according to the image presented as input at the system. *Level-1* is capable of taking several classifier outputs as input and to learn from training data how well they perform and how their outputs should be combined. Final results confirmed the strength of the method (Par. 7).

3 Level-0: a Pool of Experts

The *Level-0* stage that we realized is composed of a total of ten classifiers, five set of two kind of classifiers working in couple: Support Vector Machines [10] and Normal Bayes Classifier [11].
Five different sets of visual features are extracted from every pair of stereo image (**Fig. 2**) of the dataset:

- Color Features
- Texture Features
- Segment Features
- Depth Features
- 3D Space Features

Every couple of classifiers is trained on the same set of visual features such to obtain two different replies of the same type, such to increase the probability of obtaining a correct classification.

Fig. 2. Stereo Couple Images

3.1 Color Features

Left image of every frame stereo couple is converted from RGB to *Luv color mode* (**Fig. 3**) and divided in nine sub-image. From every sub-image are extracted *mean* and *standard deviation* of each of the three image channels for a total of 54 features.

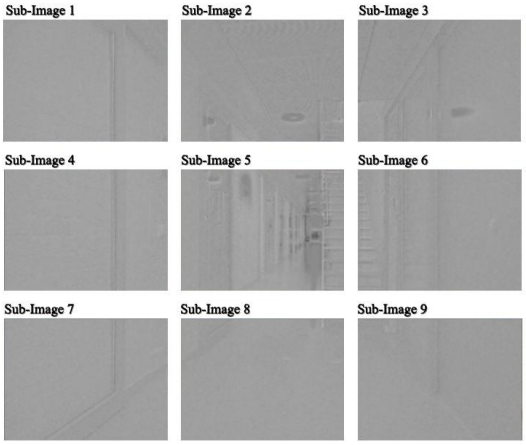

Fig. 3. Color Features Extraction

3.2 Texture Features

Left image of every frame stereo couple is divided in nine sub-image that are processed with *Discrete Fourier Transform* to extract frequency components (**Fig. 4**). Magnitude spectrum of each sub-image is calculated and are extracted *frequency* and *phase* of the ten higher power components, for a total of 180 features.

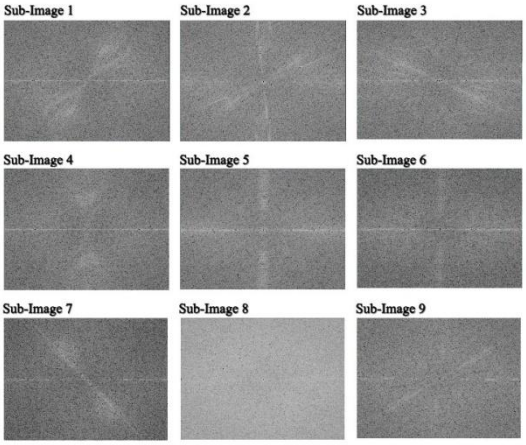

Fig. 4. Texture Features Extraction

3.3 Segment Features

Left image from every frame is filtered using *Canny Edge Detector* algorithm [12] and from result image are extracted 30 line segments (**Fig. 5**) using *Probabilistic Hough Transform* [8]. For every segment are extracted *length* and *angle* for a total of 60 features.

Fig. 5. Segment Features Extraction

3.4 Depth Features

From every frame stereo couple disparity map is calculated (**Fig. 6**) using *Semi-Global Block Matching Stereo Correspondence* algorithm [9]. Disparity map is divided in nine grayscale sub-images, and from each sub-image are extracted *mean* and *standard deviation* for a total of 18 features.

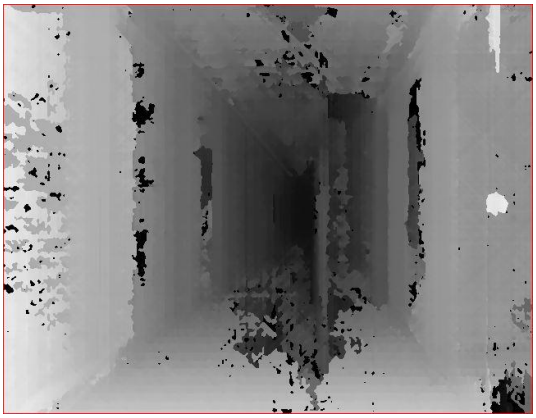

Fig. 6. Depth Features Extraction

3.5 3D Space Features

Using stereo system *Intrinsic Parameters* every pixel of Disparity Map is projected in 3D space coordinates system (**Fig. 7**). From 3D Space Information are extracted seven features:

- *mean* and *standard deviation* of *distance* of each 3D point from robot reference system origin;
- *mean* and *standard deviation* of *height* of every 3D point;
- *mean*, *standard deviation* and *maximum* of *depth* (Z coordinate) values.

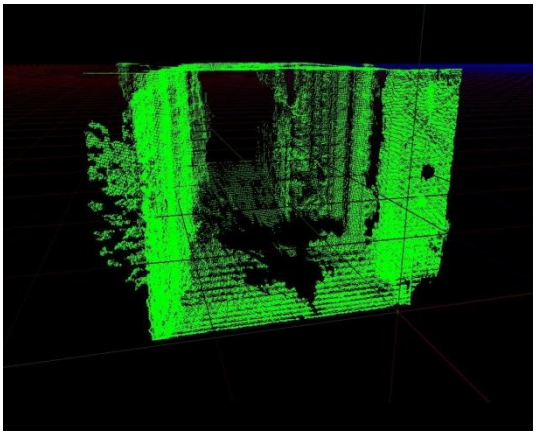

Fig. 7. 3D Features Extraction

4 Level-1: Regression Stage

Classifiers labels obtained at Level-0 stage are combined to produce a set of *Confidence Values*. To choose the best approach we analyzed four different algorithms for Level-1 stage:

- Committee of Experts
- Stacked Regression using Support Vector Machines
- Stacked Regression using Artificial Neural Network
- Linear Weighted Combination of the previous three

4.1 Committee of Experts

Committee **o**f **E**xperts (**CoE**) was initially described as a method to improve regression estimates in [4] and [5], but can be used for both regression and classification. We use a modified version of **CoE** that has multiple output values instead of single output label. The algorithm is represented in **Fig. 8**: at Level-0 a pool of *N* experts estimates a target function $h(x)$ giving N replies $1_j(i)$, with $j = 1..N$ *and* $i = 1..M$, where M is the number of the classes and

$$1_j(i) = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

is the Indicator Function.
The reply of each *j-th* Level-0 expert ($1_j(i)$) indicates one (*i-th*) of the M available classes. The N outputs coming from Level-0 are linearly combined using weights:

$$Out_i = \sum_{j=1}^{N} w_j * 1_j(i)$$

$$\sum_{j=1}^{N} w_j = 1.0$$

The weights w_j are manually assigned to each expert, according to the accuracy evaluated in classifying Validation Set images.

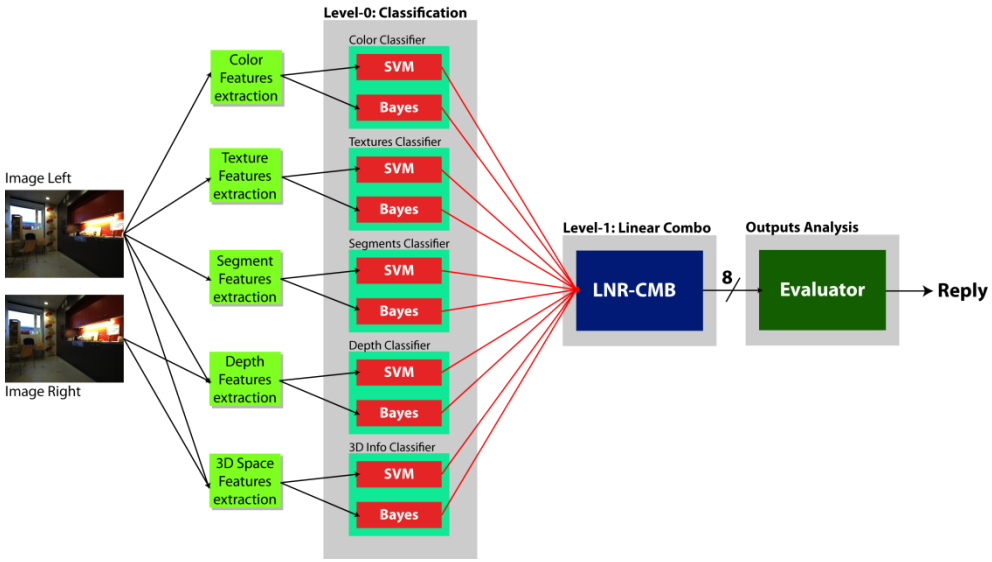

Fig. 8. Committee of Experts

4.2 Stacked Regression

Stacked **R**egression (**SR**) [6] is based on **S**tacked **G**eneralization (**SG**) method, introduced by Wolpert in 1990 [3] and was initially presented as a method to combine multiple models for classification. Next **SG** was also used for regression and even unsupervised learning [7].

Like in **CoE** method we have a pool of experts estimating a target function $h(x)$: the pool composes the so called "*Level-0 generalizer"* and it is trained in the first stage of **SR.** The second stage consists in training a Regression Level that takes as inputs the outputs of Level-0 generalizers and try to estimate the Confidence Values of every possible output

class. This stage is called "*Level-1 Regression*" and its purpose is to learn the biases of Level-0 generalizers.

It is very important that Level-1 and Level-0 machine learning networks are trained using different dataset: in this way generalization capabilities are granted, and over fitting probabilities is decreased. The training approach chosen will be detailed in Par. 6. We chose to evaluate two kinds of Level-1 approach:

- Array of **S**upport **V**ector **M**achines **R**egressors (**SVM-R**), one for every possible output class (**Fig. 9**).

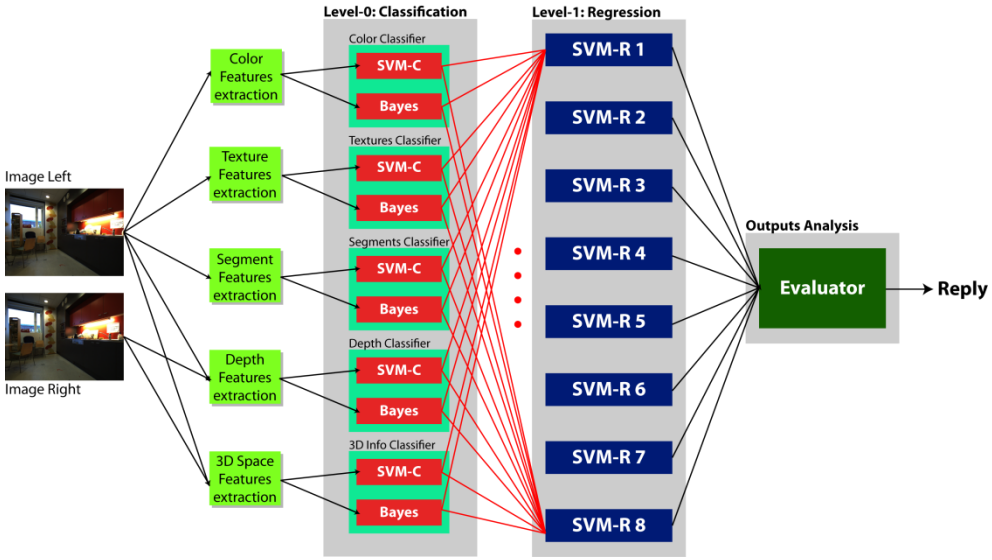

Fig. 9. Stacked Regression with SVM-R array

- **A**rtificial **N**eural **N**etwork - **M**ulti **L**ayer **P**erceptron (**ANN-MLP**) with an output for every possible class (**Fig. 10**).

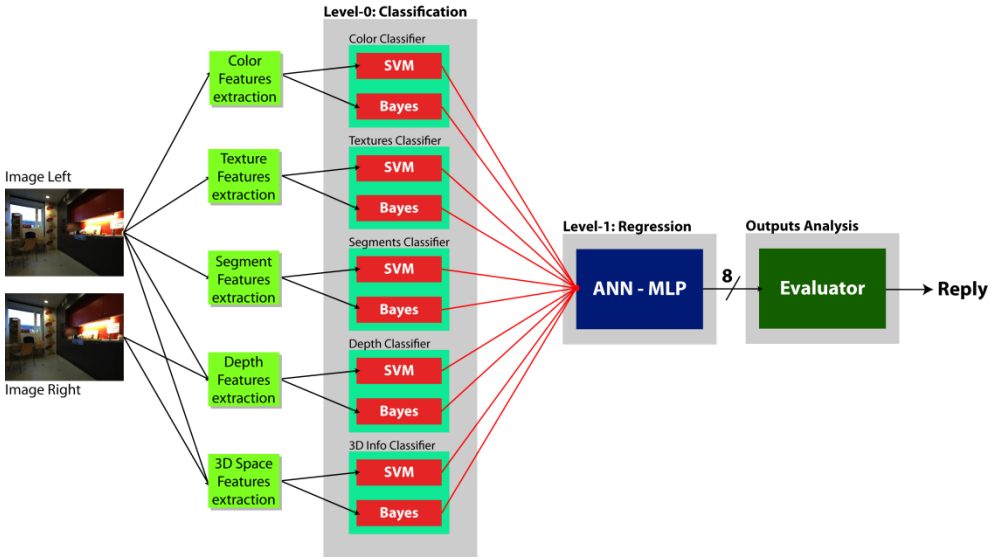

Fig. 10. Stacked Regression with ANN-MLP

SVM-R approach is different from **ANN-MLP** since every Level-1 output is independent from each other. Every output is obtained using a different SVM-R network trained to recognized its own class and to regret all the other.

ANN-MLP outputs are instead obtained by a single network with multiple outputs, where each output is influenced by the full connection between units of hidden-outputs layer.

4.3 Linear Combination of methods

The three methods previously analyzed can be linearly combined using weights such to obtain a unique more precise reply; the combination scheme is illustrated in **Fig. 11**. The linear combination is weighted and weights w_i have been chosen according to generalization capabilities evaluated during Validation Phase on Validation Set images, where

$$\sum_{i=1}^{3} w_i = 1.0$$

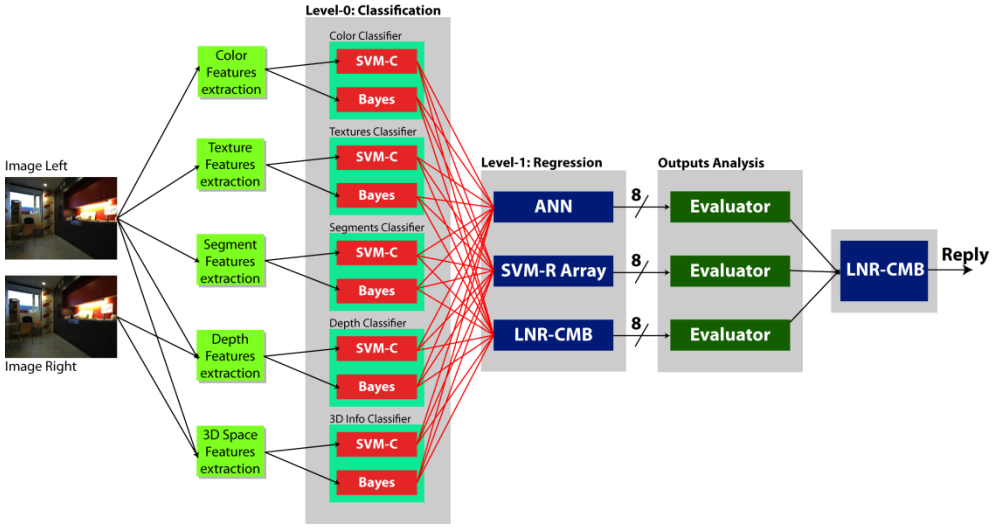

Fig. 11. Linear Weighted Combination of Regressors

5 Final stage: Level-1 outputs analysis

For every room class available Level-1 stage gives an output in the range [0.0, 1.0], where 0.0 implies *total rejection* and 1.0 *total agreement*. At every available class is assigned a *Level of Confidence* based on the activation threshold scheme as illustrated in **Fig. 12**. *Th_h* and *Th_l values* are respectively high and low activation threshold. If only one output was in *Activation Zone* the class related to that output has been chosen as final label reply; if more than one output or none was in *Activation Zone*, we chose to reply "*UNKNOWN*" or to not reply according to the number of class present in every zone.

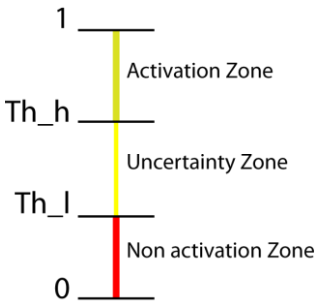

Fig. 12. Level-1 output activation scheme

6 Training phase

Training phase is a critical step for Machine Learning approach. A good configuration of training set will bring to a good performance of machine learning system. Training Set available for Robot Vision task was highly unbalanced. There were rooms with 1146 examples images and rooms with only 192 images on a total of 4782 examples. Using original training set we went to a system very performing on rooms with a large number of example (i.e. Corridor) and very weak for the others (i.e. Printer Area).

To avoid this issue, Training Set has been reorganized according to this scheme:

- examples was randomly distributed to avoid "sequenced training";
- for every class was chosen 600 examples, replicating frames for those classes with less than 600 image couples;
- for **SG** configurations, examples chosen to be presented to Level-0 stage were not inserted in training set for Level-1;

With this configuration scheme we obtained a remarkable improvement in performances two different training sets for Level-0 and Level-1, each composed of 4800 examples (just compare RUN #2 and RUN #10 scores in **Table 1**).

7 Final Results and Future Works

For ImageCLEF 2010 Robot Vision Task we submitted twenty RUNs (**Table 1**) on a total of 42 RUNs submitted by all the team participating. Our best score has been 253 (the winner totalized 677), obtained in RUN #14 and we placed 4th.

The twenty submitted RUNs can be subdivided in five groups of four RUNs. Each group was composed of four tests on all the four methods previously illustrated and was different from the other for at least one feature. The first group was an initial test made training Level-0 Classifiers and Level-1 Regressors on the original unbalanced Training Set, for the other groups the same Balanced Training Set was used (see Par. 6). The second group was the first test carried on recognition of "Unknown Rooms"; "Unknown Rooms" recognition has been disabled in the third group where in case of not dominant Level-1 reply we chose to not give a reply label for the analyzed frame. In the last two groups Level-1 Regressors was trained again on the same Level-0 outputs of second and third groups, with Stereo Vision disabled such to determinate its real contribute to final results, since we noticed that Disparity Map and 3D Space Classifiers performances were very poor.

Analyzing result table is evident the strength of Combining of Classifiers method.

Looking at RUNs from #9 to #12 we can notice that the best classifier in Level-0 (columns from 7 to 16) obtained a score of -725, while the worst result (column 2) in Level-1 configurations is -342. This is evident also looking at RUNs from #13 to #16.

Our best RUN, the #14, was obtained using **SR** with **SVM** approach, choosing to not classify unknown rooms (giving indeed no reply) and deactivating Stereo Vision Features that was not very performing (as highlighted in columns of runs from #5 to #12).

The weakness of our method lies in unknown rooms recognizing phase. To improve this phase we developed a method that introduces a new Level of classification that, taking "Level-1" outputs as input, would make distinction between "known" and "unknown" rooms in the case that there is not a dominant Level-1 output (Par. 5). The short time available did not allow us to fully complete this enhancement, that is scheduled to realized for our future works on Image Classification.

						Color Features		Texture Features		Hough Features		Disparity Map Features		3D Space Features	
RUN #	**Score**	**Type**	**Unknown Available**	**Training Set**	**Stereo Used**	**SVM-C**	**Bayes**	**SVM-C**	**Bayes**	**SVM-C**	**Bayes**	**SVM-C**	**Bayes**	**SVM-C**	**Bayes**
1	185	CoE	NO	Unbalanced	YES	-869	-823	-1409	-1226	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
2	52	SR-SVM	NO	Unbalanced	YES	-869	-823	-1409	-1226	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
3	-971	SR-ANN	NO	Unbalanced	YES	-869	-823	-1409	-1226	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
4	90	Lin. Combo	NO	Unbalanced	YES	-869	-823	-1409	-1226	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
5	-618	CoE	YES	Balanced	YES	-881	-725	-1467	-1179	-1295	-1465	-2093	-1965	-2243	-1455
6	-624	SR-SVM	YES	Balanced	YES	-881	-725	-1467	-1179	-1295	-1465	-2093	-1965	-2243	-1455
7	-1101	SR-ANN	YES	Balanced	YES	-881	-725	-1467	-1179	-1295	-1465	-2093	-1965	-2243	-1455
8	-1092	Lin. Combo	YES	Balanced	YES	-881	-725	-1467	-1179	-1295	-1465	-2093	-1965	-2243	-1455
9	48	CoE	NO	Balanced	YES	-881	-725	-1467	-1179	-1295	-1465	-2093	-1965	-2243	-1455
10	228	SR-SVM	NO	Balanced	YES	-881	-725	-1467	-1179	-1295	-1465	-2093	-1965	-2243	-1455
11	-342	SR-ANN	NO	Balanced	YES	-881	-725	-1467	-1179	-1295	-1465	-2093	-1965	-2243	-1455
12	-131	Lin. Combo	NO	Balanced	YES	-881	-725	-1467	-1179	-1295	-1465	-2093	-1965	-2243	-1455
13	9	CoE	NO	Balanced	NO	-881	-725	-1467	-1179	-1295	-1465	n.u.	n.u.	n.u.	n.u.
14	**253**	**SR-SVM**	**NO**	**Balanced**	**NO**	**-881**	**-725**	**-1467**	**-1179**	**-1295**	**-1465**	n.u.	n.u.	n.u.	n.u.
15	5	SR-ANN	NO	Balanced	NO	-881	-725	-1467	-1179	-1295	-1465	n.u.	n.u.	n.u.	n.u.
16	-172	Lin. Combo	NO	Balanced	NO	-881	-725	-1467	-1179	-1295	-1465	n.u.	n.u.	n.u.	n.u.
17	-391	CoE	YES	Balanced	NO	-881	-725	-1467	-1179	-1295	-1465	n.u.	n.u.	n.u.	n.u.
18	-560	SR-SVM	YES	Balanced	NO	-881	-725	-1467	-1179	-1295	-1465	n.u.	n.u.	n.u.	n.u.
19	-1206	SR-ANN	YES	Balanced	NO	-881	-725	-1467	-1179	-1295	-1465	n.u.	n.u.	n.u.	n.u.
20	-926	Lin. Combo	YES	Balanced	NO	-881	-725	-1467	-1179	-1295	-1465	n.u.	n.u.	n.u.	n.u.

Table 1. Submitted RUNs Scores and System Configurations (n.a. = *used, but scores not available* / n.u. = *not used*)

References

[1] Andrzej Pronobis, Marco Fornoni, Henrik I. Christensen, and Barbara Caputo, "The Robot Vision Task at ImageCLEF 2010" In the Working Notes of CLEF 2010, Padova, Italy, 2010.

[2] Dima, C. S., Nicolas, V., & Martial, H., "Classifier Fusion for Outdoor Obstacle Detection", In *International Conference on Robotics and Automation* - Vol. 1, pp. 665-671, 2004.

[3] Wolpert, D. H., "Stacked Generalization" Los Alamos, NM, Tech. Rep. LA-UR-90-3460, 1990.

[4] M. Perrone, "Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization", Ph.D. dissertation, Brown Universitiy, 1993.

[5] M.P. Perrone and L.N. Cooper, "When networks disagree: Ensemble methods for Hybrid Neural Networks" in *Neural Network for speech and Image Processing*, R. J. Mammone, Ed. Chapmann-Hallpp.126-142, 1993.

[6] L. Breiman, "Stacked Regression", *Machine Learning*, vol. 24, no. 1, 1996.

[7] P. Smyth and D. Wolpert, "An evaluation of linearly combining density estimators via stacking", Information and Computer Science Department, University of California, Irvine, Tech. Rep., 1998.

[8] Matas, J. and Galambos, C. and Kittler, J.V., "Progressive Probabilistic Hough Transform", *BMVC98*, 1998.

[9] H. Hirschmuller, "Stereo Vision in Structured Environments by Consistent Semi-Global Matching", *CVPR (2)'06*, pp. 2386-2393, 2006.

[10] J. Shawe-Taylor, N. Cristianini, "Support Vector Machines and other kernel-based learning methods"- Cambridge University Press, 2000.

[11] K. Fukunaga. "Introduction to Statistical Pattern Recognition. second ed.", New York: Academic Press, 1990.

[12] "Canny Edge Detector" - http://en.wikipedia.org/wiki/Canny_edge_detector.