Overview of the ImageCLEF 2014 Robot Vision Task

Jesus Martinez-Gomez[1,3], Ismael García-Varea[1], Miguel Cazorla[2], and Vicente Morell[2] **

[1] University of Castilla-La Mancha, Albacete, Spain
[2] University of Alicante, Alicante, Spain
[3] University of Malaga, Malaga, Spain
{Jesus.Martinez, Ismael.Garcia}@uclm.es
{Miguel.Cazorla, Vicente.Morell}@ua.es

Abstract. This article describes the RobotVision@ImageCLEF 2014 challenge, which addresses two problems: place classification and object recognition. Participants of the challenge were asked to classify rooms on the basis of visual and depth images captured by a Kinect sensor mounted on a mobile robot. They were also asked to detect the appearance or lack of several objects. The proposals of the participants had to answer two questions: "where are you?" (I am in the corridor, in the kitchen, etc.), and "list the objects that you can see?", from a predefined list (I can see a table and a chair but not a computer) when presented with a test frame (a visual and a depth image). The number of times a specific object appears in a frame is not relevant. Two different sequences of frames were provided for training and validation purposes, respectively. The final test sequence included images acquired in a similar but different indoor office environment, which is considered the main novelty of the 2014 edition of the task. In contrast to previous editions of the task, sequences do not represent the temporal continuity in the acquisition procedure and therefore, test frames have to be processed sparsely. The winner of the 2014 edition of the Robot Vision task was the NUDT group, from China.

1 Introduction

This paper describes the ImageCLEF 2014 Robot Vision challenge [9], a competition that started in 2009 within the ImageCLEF[4] [1] as part of the Cross Language Evaluation Forum (CLEF) Initiative[5]. Since its origin, the Robot Vision task has been addressing the problem of place classification for mobile robot localization.

** This work has been partially funded by FEDER funds and the Spanish Government (MICINN) through projects TIN2010-20900-C04-03, DPI2013-40534-R and by the Innterconecta Programme 2011 project ITC-20111030 ADAPTA.

[4] http://imageclef.org/
[5] http://www.clef-initiative.eu//

The 2009@ImageCLEF edition of the task [12], 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 in 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 [10] was held in conjunction with ICPR 2010 conference. In that edition, where 9 groups participated, the use of stereo images and two types of different training sequences (easy and hard), that had to be used separately, were introduced. The 2010@ImageCLEF edition [11], with 7 participating groups, was focused on generalization: several areas could belong to the same semantic category.

In 2012 [5], stereo images were 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 consisted of two types of images and the challenge become a problem of multimodal (place) classification. In addition to the use of depth images (using a visual representation), the optional task contained kidnappings and unknown rooms (not previously seen in training sequences) not appeared in the test sequences. Moreover, several techniques for features extraction and cue integration were proposed to the participants.

In 2013 [6], the visual data was changed, providing the traditional RGB images and their corresponding point cloud information. The main difference from 2012 edition was that no depth image was provided but the point cloud itself. The purpose of that was to encourage the participants to make use of 3D image processing techniques, in addition to visual ones, with the aim to obtain better classification results. Furthermore, for some specific rooms, we provided completely dark images for which the use of the 3D information had to be used in order to classify such a room. In addition to the use of the point cloud representation, the 2013 edition of the task included object recognition.

For the ImageCLEF 2014 Robot Vision challenge, we have introduced two main changes. Firstly, the temporal continuity from the image acquisition has been completely removed in the training, validation and test sequences. That is, consecutive frames in the provided sequences do not represent consecutive frames during the acquisition procedure. The second change is the inclusion of validation and test frames acquired in a different environment. Namely, we acquired new frames in a different building that contains the same type of rooms

and objects imaged in the training and part of the validation sequence. Regarding the participation, in this edition, we received a total of 17 runs, from 4 different participant groups. The best result was obtained by the NUDT research group from the National University of Defense Technology, Changsha, China.

The rest of the paper details the challenge and is organized as follows: Section 2 describes the 2014 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, the main conclusions are drawn and some ideas for future editions are outlined.

2 The RobotVision Task

This section describes the details concerning the setup of the ImageCLEF 2014 Robot Vision task. In Section 2.1 a description of training, validation and test sequences is provided. In Section 2.2 the performance evaluation criteria is detailed. Finally, in Section 2.3 a brief description of the baseline visual place classification system provided by the organizers, as well as other relevant details concerning the task are presented.

2.1 Description

The sixth edition of the Robot Vision task is focused on the use of multimodal information (visual and depth images) with application to semantic localization and object recognition. The main objective of this edition is to address the problem of robot localization in parallel to object recognition from a semantic point of view, with a special focus on generalization. Both problems are inherently related: the objects present in a scene can help to determine the room category and vice versa. Solutions presented should be as general as possible while specific proposals are not desired. In addition to the use of visual data, a 3D point cloud representation of the scene acquired from a Microsoft Kinect device was used, which has shown as a de facto standard in the use of depth images. In this new edition of the task, we introduced strong variations between training and test scenarios with the aim to solve the object recognition and localization problems in parallel and for a great variety of different scenarios.

Participants were given visual images and depth images in Point Cloud Data (PCD) format. Fig. 2 shows the same scene represented in a visual image and a point cloud data file. Training, validation and test sequence were acquired within two different buildings presenting a similar structure but with some variations in the objects distribution. All the room and object categories included in the test sequence were previously seen during training and validation.

As for the 2013 edition of the challenge, no sub-tasks were defined and all participants have to prepare their submissions using the same test sequence.

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

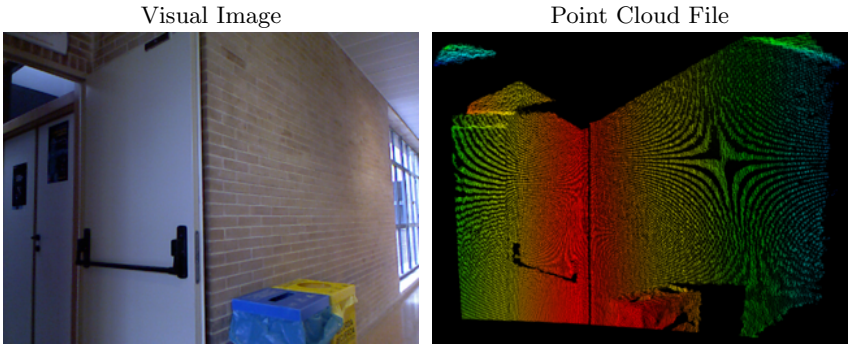

Fig. 2. Visual and 3D point cloud representation for a scene.

2.2 The Data

In the 2014 edition of the RobotVision challenge, a new version of the unreleased Robot Vision dataset was created. This specific dataset consists of three sequences (training, validation and test) of depth and visual images acquired within the following indoor environment: two department buildings at the University of Alicante, in Spain. Visual images were stored in PNG format and depth ones in PCD. Every image in the dataset was manually labelled with its corresponding room category/class and with a list of eight different objects to appear or not within it. The 10 different room categories are: Corridor, Hall, ProfessorOffice, StudentOffice, TechnicalRoom, Toilet, Secretary, VisioConference, Warehouse and ElevatorArea. The 8 different objects are: Extinguisher, Phone, Chair, Printer, Urinal, Bookself, Trash and Fridge. The dataset used in the task includes two labelled sequences used for training and validation with 5000 and 1500 images respectively. The unlabeled sequence used for test consists of 3000 different images. The frequency distribution for room categories in the training, validation and test sequences are depicted in Table 1. Regarding the building used in the acquisition, all the 5000 training images were acquired in

the building A, the same used for the 2013 edition dataset. The validation sequence included 1000 images from building A but 500 new images from building B. Finally, all 3000 test images were acquired in building B.

Table 1. Distribution of room categories for the training, validation and test sequences.

	Number of frames		
Room Category	Training (Building A)	Validation (Building A+B)	Test (Building B)
Corridor	1833	479	772
Hall	306	103	202
ProfessorOffice	355	149	372
StudentOffice	498	174	419
TechnicalRoom	437	110	242
Toilet	389	094	141
Secretary	336	102	245
VisioConference	364	113	159
Warehouse	174	081	201
ElevatorArea	308	095	247
All	5000	1500	3000

It can be observed that in all sequences, Corridor is the class with higher number of frames. This is because most of the space of the University of Alicante building, suitable for robot navigation, belongs to several corridors. This situation makes it easier the classification of test frames as Corridor while other classes as VisioConference or ElevatorArea are more challenging. The frequency distribution for rooms in the different sequences is graphically presented in Fig. 3.

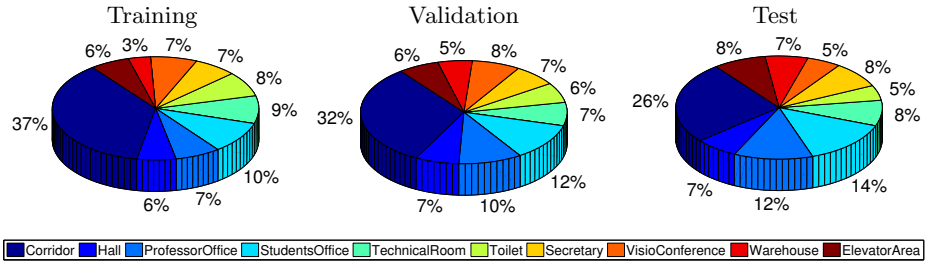

Fig. 3. Class distribution in training, validation and test sequences.

The distribution for object categories in the training, validation and test sequences is depicted in Table 2, while frequencies are presented in Fig. 4. Despite of small variations, it can be observed how classes and objects frequencies are maintained along training, validation and testing sequences.

The differences between all the room categories can be observed in Fig. 5, where a single visual image for each of the 10 room categories is shown. Moreover,

Table 2. Distribution of object presences or lacks for the training, validation and test sequences.

	Number of presences / lacks		
Object Category	Training (Building A)	Validation (Building A+B)	Test (Building B)
Extinguisher	770 / 4230	238 / 1262	566 / 2434
Chair	1304 / 3696	471 / 1029	1070 / 1930
Printer	473 / 4527	139 / 1361	265 / 2735
Bookshelf	802 / 4198	317 / 1183	896 / 2104
Urinal	162 / 4838	040 / 1460	060 / 2940
Trash	813 / 4187	323 / 1177	797 / 2203
Phone	267 / 4733	113 / 1387	303 / 2697
Fridge	190 / 4810	034 / 1466	047 / 2953
All	4781 / 35219	1675 / 10325	4004 / 19996

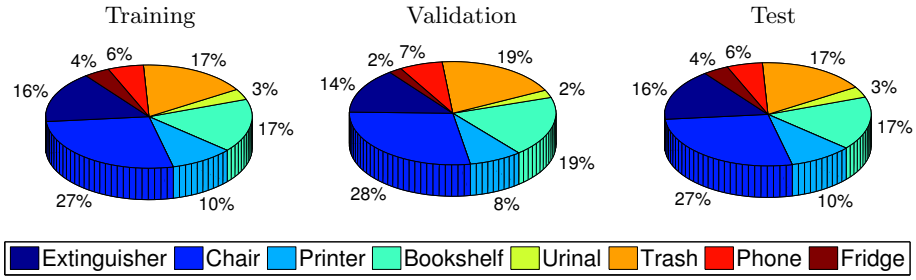

Fig. 4. Object distribution in training, validation and test sequences.

Fig. 6 shows four examples of visual images for each of the 8 different objects appearing in the dataset.

2.3 Performance Evaluation

The runs submitted for each participant were compared and sorted according the score assigned to each submission. Every submission consisted of the room category assigned to each test image and the corresponding list of the 8 detected/non-detected objects within that image. As we already mentioned above, the number of times a specific object appears in an image was not relevant to compute the score. The score was computed using the rules shown in Table 3. For a better understanding of the score computation, an example of three different hypothetical user decisions for a specific test image is shown in Table 4. Due to the fact that wrong room classifications account negatively to the score, participants were allowed to not providing such information, in which case the score is not affected. The final score was computed as the sum of the score obtained for each individual test frame. According to the test set released the maximum score to be obtained was 7004 points.

Fig. 5. Examples of visual images (one for each of the 10 different categories) from the Robot Vision 2014 dataset

Fig. 6. Examples of visual images (four for each of the 8 different objects) from the Robot Vision 2014 dataset

2.4 Additional information provided by the organization

In addition to all the image sequences, we created a Matlab script to be used as template for participants proposals. This script performs all the steps for generating solutions for the Robot Vision challenge: features generation, training, classification and performance evaluation. Basic features are generated for both visual and depth images (histograms) and training and classification is performed

Table 3. Rules used to calculate the final score for a test frame

Room class/Category	
Room class/category correctly classified	+1.0 points
Room class/category wrongly classified	-0.5 points
Room class/category not classified	+0.0 points
Object Recognition	
For each object correctly detected (True Positive)	+1.0 points
For each object incorrectly detected (False Positive)	-0.25 points
For each object correctly detected as not present (True Negative)	+0.0 points
For each object incorrectly detected as not present (FalseNegative)	-0.25 points

Table 4. Examples of performance evaluation of three different user decisions for a single test frame of a TechnicalRoom with two type of objects appearing in the scene: Phone and Printer, where the maximum score to be obtained with this test frame is 3.0

Correct values for the test frame

Room class/Category	Extinguisher	Phone	Chair	Printer	Urinal	Bookself	Trash	Fridge
TechnicalRoom	NO	YES	NO	YES	NO	NO	NO	NO

User decision a) It is TechnicalRoom with two type of objects appearing in the scene: Phone and Trash. Total Score: 1.5

Room class/Category	Extinguisher	Phone	Chair	Printer	Urinal	Bookself	Trash	Fridge
TechnicalRoom	NO	YES	NO	NO	NO	NO	YES	NO
1.0	0.0	1.0	0.0	-0.25	0.0	0.0	-0.25	0.0

User decision b) An Unknown room with two type of objects appearing in the scene: Phone and Printer. Total Score: 2.0

Room class/Category	Extinguisher	Phone	Chair	Printer	Urinal	Bookself	Trash	Fridge
Unknown	NO	YES	NO	YES	NO	NO	NO	NO
0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	0.0

User decision c) A Corridor with two type of objects appearing in the scene: Extinguisher and Printer. Total Score: 0.0

Room class/Category	Extinguisher	Phone	Chair	Printer	Urinal	Bookself	Trash	Fridge
Unknown	YES	NO	NO	YES	NO	NO	NO	NO
-0.5	-0.25	-0.25	0.0	1.0	0.0	0.0	0.0	0.0

using an On-line Independent Support Vector Machines [8] that, in comparison with SVM, dramatically reduces learning time and space requirements at the price of a negligible loss in accuracy.

3 Participation

In 2014, 28 participants registered to the Robot Vision task but only 4 submitted, at least, one run accounting for a total of 17 different runs. These participants were:

- NUDT: National University of Defense Technology, Changsha, China.
- UFMS CPPP: Federal University of Mato Grosso do Sul, Ponta Pora, Brazil
- AEGEAN: University of the Aegean Karlovassi, Greece
- SIMD: University of Castilla-La Mancha, Albacete, Spain.
  - Out of competition organizers contribution using the techniques included in the MATLAB proposed script. It can be considered as a baseline result.

4 Results

This section presents the results of the Robot Vision task of ImageCLEF 2014.

4.1 Overall Results

The scores obtained by all the submitted runs are shown in Table 5. The maximum score that could be achieved was 7004 and the winner (NUDT) obtained a score of 4430,25 points. This maximum score is the addition of the maximum score computed from rooms classification (3000) and object recognition (4004).

Table 5. Overall ranking of the runs submitted by the participant groups to the 2014 Robot Vision task

Rank	Group Name	Score Room (% Max)	Score Objects (% Max)	Score Total (% Max.)
1	NUDT	1075,5 (35,85)	3354,75 (83,78)	4430,25 (63,25)
2	NUDT	1060,5 (35,35)	3354,75 (83,78)	4415,25 (63,04)
3	NUDT	1057,5 (35,25)	3354,75 (83,78)	4412,25 (63,00)
4	NUDT	1107,0 (36,90)	3276,00 (81,82)	4383,00 (62,58)
5	NUDT	1107,0 (36,90)	3245,50 (81,06)	4352,50 (62,14)
6	NUDT	1113,0 (37,10)	3233,75 (80,76)	4346,75 (62,06)
7	NUDT	1060,5 (35,35)	3096,75 (77,34)	4157,25 (59,36)
8	NUDT	1057,5 (35,25)	3096,75 (77,34)	4154,25 (59,31)
9	NUDT	1030,5 (34,35)	2965,25 (74,06)	3995,75 (57,05)
10	NUDT	1008,0 (33,60)	2870,00 (71,68)	3878,00 (55,37)
11	UFMS	0219,0 (07,30)	1519,75 (37,96)	1738,75 (24,83)
12	UFMS	0213,0 (07,10)	1453,00 (36,29)	1666,00 (23,79)
13	UFMS	0192,0 (06,40)	1460,50 (36,48)	1652,50 (23,59)
14	UFMS	0150,0 (05,00)	1483,00 (37,04)	1633,00 (23,32)
15	SIMD	0067,5 (02,25)	0186,25 (04,65)	0253,75 (03,62)
16	AEGEAN	-405,0 (-13,50)	-995,00 (-24,85)	-1400,00 (-19,99)
17	AEGEAN	-405,0 (-13,50)	-1001,00 (-25,00)	-1406,00 (-20,07)

SIMD organizers submission was out-of-competition submission, and it was provided to be considered a baseline score. For this submission, just the techniques proposed in the webpage of the challenge[6] were used. Concretely, it was generated an image descriptor by concatenating both depth and visual histograms. These descriptors were then used as input to train an Online Support Vector Machine [3] using DOGMA [7].

4.2 Details and participants approaches

A detailed view of the obtained results for the best submissions of the 4 participants is show in Fig. 7. This figure graphically presents how participants submission performed notoriously better for object recognition than for room classification. For example, the winner of the task achieved 83,78% of the maximum score (3354.75 out of 4004 points) for the object recognition problem, while they just obtained 1075,5 out of 3000 points (35,85%) for the scene classification problem.

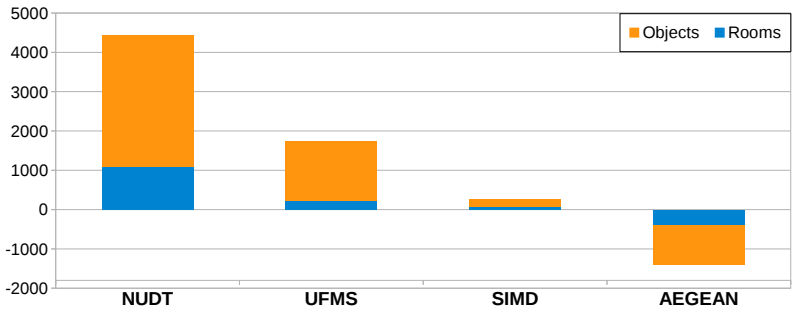

Fig. 7. Detailed results for best groups submissions.

In relation to the participant approaches, NUDT and UFMS groups submitted a working note with internal details of their submissions. The NUDT proposal [13] that ranked first followed a spatial pyramid matching approach [4] based on appearance and shape features. Concretely, they used a Bag of Words (BoW) representation to create the appearance descriptor from dense SIFT features. The shape was represented using Pyramid Histograms of Gradients (PHOG) approach. Shape and appearance descriptors were then concatenated to create a single image descriptor used as input for the classification step. The classification was performed using a multi-class SVM using an one versus all

[6] http://www.imageclef.org/2014/robot

strategy. The CPPP/UFMS proposal [2] also uses dense SIFT descriptors and the spatial pyramid approach. However, this approach is based on a k-nearest neighbor classifier and no PHOG descriptors are considered. None of the groups used the depth information encoded in the point cloud files that were released in conjunction to the visual images.

In view of the obtained results, we can conclude that room classification remains as an open problem when generalization is requested. That is, current approaches (as shown in previous task editions) perform well when the test environment has been previously imaged during training, but they present problems to classify frames acquired in new environments. On the other hand, we should point out the high performance of the submissions when facing the object recognition problem. This can be explained because object recognition does not rely on the scene generalization as for the room classifications. Namely, phones or chairs will always be recognized as their type (a phone or a chair, respectively) independently from the scene where they are placed.

5 Conclusions and Future Work

In this paper, the overview of the 2014 edition of the Robot Vision task at ImageCLEF has been presented. We have described the task, which had slightly variations from previous editions, and a detailed analysis of the results obtained for the participants proposals.

As a novelty for this edition, we have introduced physical changes in the environment where the test sequence has been acquired. That provides an additional component to the classical place classification problem, empathizing in the generalization. According to the obtained results, this novelty has resulted in a notorious decrease on the room classification performance: none of the submission achieved more than 40% of the maximum score. The inclusion of depth images in the participants proposals could have increased the performance of the room classifiers. With respect to the object recognition, it was properly managed by the NUDT group that ranked first.

As future work, we plan to manage both room classification and object recognition problems jointly. All the participants solutions are based on using the same technique to classify the room and to recognize objects. Both problems are solved without any type of correlation, a different way as humans do. Therefore, future work will focus on making participants classify rooms using as input the list of objects recognized in the scene.

References

1. B. Caputo, H. Müller, J. Martinez-Gomez, M. Villegas, B. Acar, N. Patricia, N. Marvasti, S. Üsküdarlı, R. Paredes, M. Cazorla, I. Garcia-Varea, and V. Morell. ImageCLEF 2014: Overview and analysis of the results. In *CLEF proceedings*, Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2014.

2. R. de Carvalho Gomes, L. Correia Ribas, A. Antônio de Castro Junior, and W. Nunes Gonçalves. CPPP/UFMS at ImageCLEF 2014: Robot Vision Task. In *CLEF 2014 Evaluation Labs and Workshop, Online Working Notes*, 2014.
3. Orabona F., Castellini C., Caputo B., Luo J., and Sandini G. On-line independent support vector machines. volume 43, pages 1402–1412, 2010.
4. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.
5. J. Martinez-Gomez, I. Garcia-Varea, and B. Caputo. Overview of the imageclef 2012 robot vision task. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.
6. J. Martinez-Gomez, I. Garcia-Varea, M. Cazorla, and B. Caputo. Overview of the imageclef 2013 robot vision task. In *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*, 2013.
7. F Orabona. Dogma: a matlab toolbox for online learning. *Software available at http://dogma. sourceforge. net*, 2009.
8. 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.
9. A. Pronobis and B. 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.
10. 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.
11. A. Pronobis, M. Fornoni, HI Christensesn, and B. Caputo. The robot vision track at imageclef 2010. *Working Notes of ImageCLEF*, 2010, 2010.
12. A. Pronobis, L. Xing, and B. Caputo. Overview of the clef 2009 robot vision track. In Carol Peters, Barbara Caputo, Julio Gonzalo, Gareth Jones, Jayashree Kalpathy-Cramer, Henning Müller, and Theodora Tsikrika, editors, *Multilingual Information Access Evaluation II. Multimedia Experiments*, volume 6242 of *Lecture Notes in Computer Science*, pages 110–119. Springer Berlin / Heidelberg, 2010.
13. Y. Zhang, J. Qin, F. Chen, and D. Hu. NUDT's Participation in ImageCLEF Robot Vision Challenge 2014. In *CLEF 2014 Evaluation Labs and Workshop, Online Working Notes*, 2014.