Overview of the ImageCLEF 2013 Robot Vision Task

Jesus Martinez-Gomez[1], Ismael Garcia-Varea[1], Miguel Cazorla[2], and Barbara Caputo[3,4]**

[1] University of Castilla-La Mancha
Albacete, Spain
[2] University of Alicante
Alicante, Spain
[3] Idiap Research Institute, Martigny, Switzerland
[4] University of Rome La Sapienza, Italy
[1] {Jesus.Martinez, Ismael.Garcia} @uclm.es
[2]miguel.cazorla@ua.es [3,4]bcaputo@idiap.ch

Abstract. This article describes the RobotVision@ImageCLEF 2013 challenge, which addresses two problems: place classification and object recognition. Participants of the challenge were asked to classify rooms on the basis of image sequences captured by cameras 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 elevator, in the toilet, etc.) and "which objects can you see?" (I can see a table and a chair but not a computer) when presented with a test sequence. The number of times a specific set of object appears in a frame was not considered but if they appeared or not in it. The test sequence was acquired within the same environment but with different lighting conditions than the training sequences. The main novelty of the 2013 edition of the task is the object recognition problem. For both problems: place classification and object recognition, depth and visual images were provided. Moreover, participants were allowed to take advantage from the temporal continuity of the test sequence. The winner of the 2013 edition of the Robot Vision task was the MIAR ICT group, from China.

1 Introduction

This paper describes the ImageCLEF 2013 Robot Vision challenge [12], a competition that started in 2009 within the ImageCLEF [1] [2] as part of the Cross

** 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/

Language Evaluation Forum (CLEF) Initiative [2]. Since its origin, the Robot Vision task has been addressing the problem of place classification for mobile robot localization.

The 2009@ImageCLEF edition of the task [13], 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 [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, 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, 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.

For the ImageCLEF 2013 Robot Vision challenge we changed the visual data, providing the traditional RGB images and its 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.

Regarding the participation, in this edition, we received a total of 16 runs, from 6 different participant groups. The best result was obtained by the MIART ICT research group from Beijing, China.

The rest of the paper details the challenge and is organized as follows: Section 2 describes the 2013 ImageCLEF edition of the RobotVision task. Section 3

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

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 2013 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 fifth edition of the Robot Vision challenge was focused on the problem of multi-modal information retrieval from indoor scenes. Participants had to detect, for each test image, the presence or lack of a set of objects. They also had to determine the kind of room where the image was acquired from. All the images were captured by a perspective camera (visual images) and a Kinect device (depth images) 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 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 sequences were acquired

within the same building at two different floors but with some variations in the lighting conditions and in the acquisition procedure (clockwise and counter clockwise, ground floor first or ground floor last).

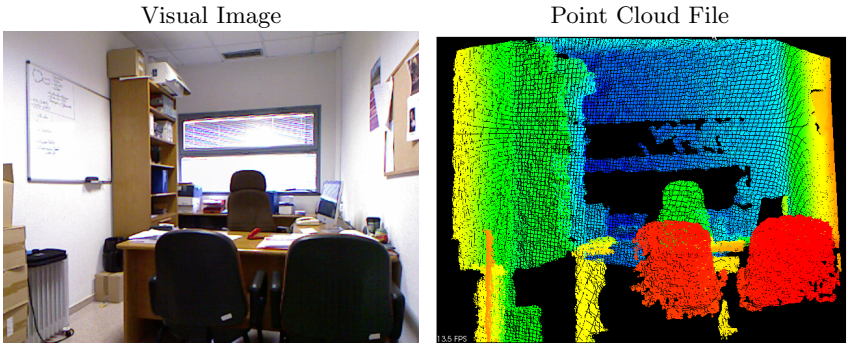


Fig. 2. Visual, depth and 3D point cloud files.

In one hand, and as opposite to previous editions of the challenge, a single task was considered this time, therefore no sub-tasks were defined. All the room and object categories included in the test sequence were previously seen during training. On the other hand, the use of the temporal continuity of the test sequence was allowed.

2.2 The Data

In the 2013 edition of the RobotVision challenge the O-VIDA Robot Vision dataset was used. This dataset consists of different training, validation and test sequences of depth and visual images acquired within an indoor environment: a department building at the University of Alicante. 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, elevator area and warehouse. The 8 different objects are: extinguisher, computer, chair, printer, urinal, screen, trash and fridge.
From this dataset two different labelled sequences were selected for training, one labelled sequence for validation, and one unlabelled sequence for testing. The frequency distribution for room categories in the training, validation and test sequences are depicted in Table 1.

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

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

	Number of frames			
Room Category	Training 1	Training 2	Validation	Test
Corridor	891	1262	764	1317
Hall	103	228	000	297
ProfessorOffice	124	192	200	222
StudentOffice	155	276	282	318
TechnicalRoom	136	281	214	240
Toilet	121	242	188	198
Secretary	098	195	181	201
VisioConference	149	300	000	306
Warehouse	070	166	000	127
ElevatorArea	100	174	040	289
All	1947	3316	1869	3515

situation makes it easier the classification of test frames as Corridor while other classes as Warehouse or Toilet are more challenging. The validation sequence was released for providing participants an additional sequence for testing their preliminary proposals. It was also released for preventing the extreme lighting conditions present in the test sequence. The validation sequence was acquired just in the first floor of the building and it does not contains any frame for three rooms: Hall, VisioConference and Warehouse. The frequency distribution for object categories in the training, validation and test sequences are depicted in Table 2, where can be observed that there are no presence of Screens in the validation sequence.

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

	Number of presences / lacks			
Room Category	Training 1	Training 2	Validation	Test
Extinguisher	259 / 1688	529 / 2787	286 / 1583	520 / 2995
Computer	289 / 1658	466 / 2850	416 / 1453	473 / 3042
Chair	470 / 1477	767 / 2549	567 / 1302	889 / 2626
Printer	210 / 1737	292 / 3024	255 / 1614	279 / 3236
Urinal	054 / 1893	110 / 3206	070 / 1799	090 / 3425
Screen	081 / 1866	190 / 3126	000 / 1869	151 / 3364
Trash	406 / 1541	451 / 2865	253 / 1616	662 / 2853
Fridge	057 / 1890	104 / 3212	099 / 1770	114 / 3401
All	1826 / 13750	2909 / 23610	1946 / 13006	3178 / 24942

The differences between all the room categories can be observed in Figure 3, where a single visual image for each of the 10 room categories is shown.

Fig. 3. Examples of visual images (one for of the 10 different categories) from the O-VIDA Robot Vision 2013 dataset

Fig. 4 shows four examples of visual images for each of the 8 different objects appearing in the dataset.

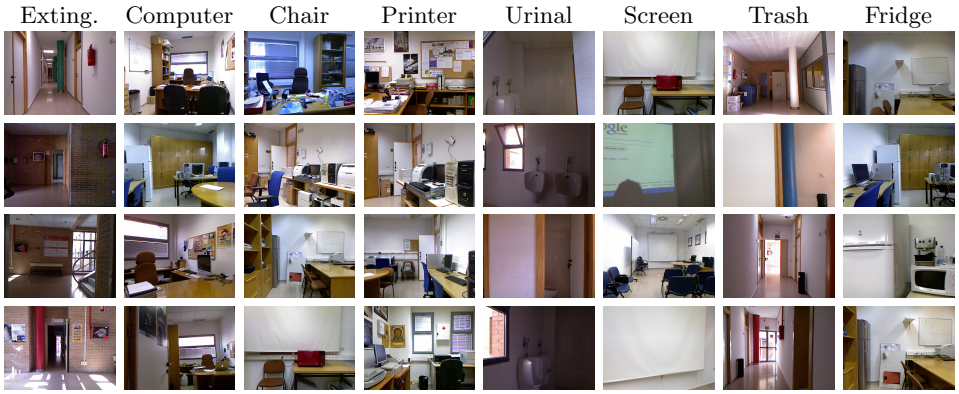


Fig. 4. Examples of visual images (four for of the 8 different objects) from the O-VIDA Robot Vision 2013 dataset

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. Due to the fact that wrong room classifications and/or wrong object detections 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 7030 points.

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

Class / Room 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 correctly classified object whitin the frame	+0.125 points
For each misclassified object whitin the frame	-0.125 points
For each not classified object whitin the frame	+0.000 points

2.4 Additional information provided by the organization

As in the previous edition [5], 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. Information about the point cloud library [14] and a basic technique for taking advantage of the temporal continuity[3] was also provided. In this regard, and in order to evaluate the performance of the baseline classification system (which was built using uniquely the provided techniques, briefly described below) we submitted a single runs. The results obtained with such proposal [4] can be considered as baseline results, which all the participants were expected to improve.

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

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

(PHOW) [1], Pyramid histogram of Local Binary Patterns (PLBP) [6], Self-Similarity-based PHOW (SS-PHOW) [15], and Compose Receptive Field Histogram (CRFH) [3].

Depth Features NARF features is a novel descriptor technique that has been included in the point cloud library [14]. 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) [9]. This SVM-based multi-class 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 [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 2013, 39 participants registered to the Robot Vision task but only 6 submitted, at least, one run accounting for a total of 16 different runs. These participants were:

- NUDT: National University of Defense Technology, Changsha, China.
- MIAR ICT: Beijing, China.
- MICA: Hanoi university of Science and Technology, Hanoi, Vietnam
- REGIM: University of Sfax National School of Engineers, Tunisia
- GRAM:University of Alcala de Henares, Spain
- SIMD: University of Castilla-La Mancha, Albacete, Spain.
  - Out of competition organizers contribution using proposed techniques

4 Results

This section presents the results of the Robot Vision task of ImageCLEF 2013.

4.1 Overall Results

The scores obtained by all the submitted runs are shown in Table 4. The maximum score that could be achieved was 7030 and the winner (MIAR ICT) obtained a score of 6033.5 points. NUDT and SIMD teams ranked second and third respectively and their score was higher than 71% of the maximum score (the one obtained with the baseline system, SIMD result in the table).

Table 4. Overall ranking of the runs submitted by the participant groups to the 2013 Robot Vision task

Rank	Group Name	Score	% Max. Score
1	MIAR ICT	6033.500	85.83
2	MIAR ICT	5924.250	84.27
3	MIAR ICT	5924.250	84.27
4	MIAR ICT	5867.500	83.46
5	MIAR ICT	5867.000	83.46
6	NUDT	5722.500	81.40
7	SIMD*	5004.750	71.19
8	REGIM	4368.250	65.98
9	MICA	4479.875	63.73
10	REGIM	3763.750	53.54
11	MICA	3316.125	47.17
12	MICA	2680.625	38.13
13	GRAM	-487.000	<0.00
14	GRAM	-497.000	<0.00
15	GRAM	-497.000	<0.00
16	NUDT	-866.250	<0.00

* SIMD organizers submission was out-of-competition, it was provided to be considered a baseline score. The organizers only used the techniques proposed in the webpage of the challenge [4]. Concretely, PHOW features were extracted from visual images and then, a Support Vector Machine was trained using DOGMA [7].

4.2 Detailed Results

Here we present a deeper analysis of the best submission for each participant group. We have computed separately the score for the class classification and the recognition sub-problems. All these results can be seen in Table 5 and Fig. 5. As it can be observed, in one hand, that MIAR ICT and NUDT groups obtained similar scores, with better results for room classification than for object recognition. On the other hand, REGIM and MICA proposals ranked better for object recognition than for room classification.

We have also analysed the specific performance for the different room categories and objects. For each room class and object considered, we have computed the percentage of right and wrong classifications. We also have computed the percentage of times of not providing information for room classes or objects. All these data can be observed in Fig. 6 and Fig. 7 for room classes and objects, respectively.

From the reported results, we can state that Hall and Elevator Area are the most challenging rooms, while Corridor is the easiest one. The number of training frames containing these classes can be one of the most important reasons for this

[4] http://www.imageclef.org/2013/robot

Table 5. Detailled ranking of the best runs submitted

Rank	Group Name	Score Class	Score Object	Score Total
1	MIAR ICT	3168.5	2865.000	6033.500
2	NUDT	3002.0	2720.500	5722.500
3	SIMD*	1988.0	3016.750	5004.750
4	REGIM	2223.5	2414.750	4368.250
5	MICA	2063.0	2416.875	4479.875
6	GRAM	-487.0	0.000	-487.000

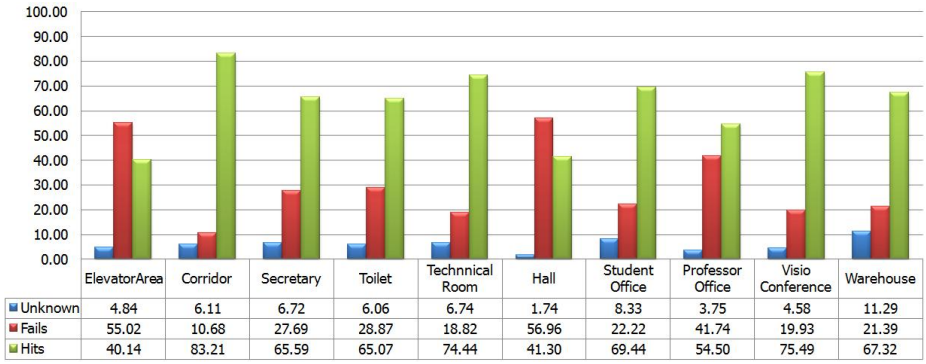
Fig. 5. Best results obtained for each group

fact. That is, the number of frames containing Corridor is about one order of magnitude higher than the ones containing Hall or Elevator Area.

It can be noticed that all the objects are managed properly by the participant proposals. Urinal was the object that obtained the highest percentage of right detections, while Trash obtained the lowest one. It should be pointed out that, for all the objects (see Table 2), the appearance ratio in less than 30%. Classifying all test frames as "there are no objects in the scene" would obtain a high positive score, especially for Urinal, Fridge or Screen. Chair and Trash could be considered the most challenging test objects because their appearance ratio is higher than for the rest of the objects. There are two possible reasons explaining that participants obtained better results for Chair than for Trash: (1) trashes are considerably smaller than chairs, and (2) trashes can appear in most of the room categories while chairs are only present in 6 rooms (TechnicalRoom, ProfessorOffice, StudentOffice, Secretary, VisioConference and Warehouse).

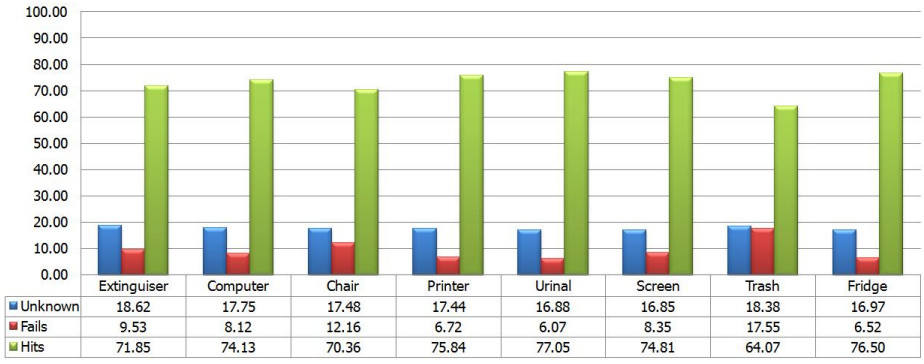
5 Conclusions and Future Work

In this paper the overview of the 2013 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 each run submitted by the participants.



	ElevatorArea	Corridor	Secretary	Toilet	Technnical Room	Hall	Student Office	Professor Office	Visio Conference	Warehouse
Unknown	4.84	6.11	6.72	6.06	6.74	1.74	8.33	3.75	4.58	11.29
Fails	55.02	10.68	27.69	28.87	18.82	56.96	22.22	41.74	19.93	21.39
Hits	40.14	83.21	65.59	65.07	74.44	41.30	69.44	54.50	75.49	67.32

Fig. 6. Percentage of hits, fails and unknowns for all room categories



	Extinguiser	Computer	Chair	Printer	Urinal	Screen	Trash	Fridge
Unknown	18.62	17.75	17.48	17.44	16.88	16.85	18.38	16.97
Fails	9.53	8.12	12.16	6.72	6.07	8.35	17.55	6.52
Hits	71.85	74.13	70.36	75.84	77.05	74.81	64.07	76.50

Fig. 7. Percentage of hits, fails and unknowns for all objects

As a novelty for this edition, we have introduced the additional problem to recognize specific objects that can appear within an image. That provides and additional component to the classical place classification problem, turning it into a multimodal classification problem.

According to the obtained results we can conclude that the introduction of the object recognition task was not as challenging as we expected: most of the participants were able to identify those objects properly. With respect to the scores obtained by the different runs, almost half of them improved the baseline results provided by the organizers, obtaining score higher than the 80% of the maximum score.

For future editions we plan to continue in the direction of including new challenging variations to the problem of scene classification problem. In particular, as the next step forward we will focus on providing the number of occurrences a specific object appears in an image.

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. B. Caputo, H. Müller, B. Thomee, R. Paredes, D. Zellhofer, H. Goeau, P. Bonnet, J. Martinez Gomez, I. Garcia Varea, and M. Cazorla. Imageclef 2013: the vision, the data and the open challenges. In Springer LNCS, editor, *CLEF 2013*, 2013.
3. O. Linde and T. Lindeberg. Object recognition using composed receptive field histograms of higher dimensionality. In *Proc. ICPR*. Citeseer, 2004.
4. Jesus Martinez-Gomez, Ismael Garcia-Varea, and Barbara Caputo. Baseline multimodal place classifier for the 2012 robot vision task. In *CLEF (Online Working Notes/Labs/Workshop)*. CLEF, 2012.
5. Jesus Martinez-Gomez, Ismael Garcia-Varea, and Barbara Caputo. Overview of the imageclef 2012 robot vision task. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.
6. 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.
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. F. Orabona, L. Jie, , and B. Caputo. Online-Batch Strongly Convex Multi Kernel Learning. In *Proc. of Computer Vision and Pattern Recognition, CVPR*, 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. 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.
13. Andrzej Pronobis, Li Xing, and Barbara 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.
14. Radu Bogdan Rusu and Steve Cousins. 3d is here: Point cloud library (pcl). In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1–4. IEEE, 2011.
15. 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.