CPPP/UFMS at ImageCLEF 2014: Robot Vision Task

Rodrigo de Carvalho Gomes, Lucas Correia Ribas, Amaury Antônio de Castro Junior, Wesley Nunes Gonçalves

Federal University of Mato Grosso do Sul - Ponta Porã Campus
Rua Itibiré Vieira, s/n, CEP 79907-414
Ponta Porã - MS, Brazil
rodrigo.firewall@hotmail.com, lucascorreiaribas@gmail.com,
{amaury.junior,wesley.goncalves}@ufms.br

Abstract. This paper describes the participation of the CPPP/UFMS group in the robot vision task. We have applied the spatial pyramid matching proposed by Lazebnik et al. This method extends bag-of-visual-words to spatial pyramids by concatenating histograms of local features found in increasingly fine sub-regions. To form the visual vocabulary, k-means clustering was applied in a random subset of images from training dataset. After that the images are classified using a pyramid match kernel and the k-nearest neighbors. The system has shown promising results, particularly for object recognition.

Key words: Scene recognition, object recognition, spatial pyramid matching

1 Introduction

In recent years, robotics has achieved important advances, such as the intelligent industrial devices that are increasingly accurate and efficient. Despite the recent advances, most robots still represent the surrounding environment by means of a map with information about obstacles and free spaces. To increase the complexity of autonomous tasks, robots should be able to get a better understanding of images. In particular, the ability to identify scenes such as office, kitchen, as well as objects, is an important step to perform complex tasks [1]. Thus, place localization and object recognition becomes a fundamental part of image understanding for robot localization [2].

This paper presents the participation of our group in the 6th edition of the Robot Vision challenge[1] [3, 4]. This challenge addresses the problem of semantic place classification and object recognition. For this task, the bag-of-visual-words approach (BOW) [5, 6] is one of the most promising approaches available. Although the approach have advantages, it also has one major drawback, the absence of spatial information. To overcome this drawback, a spatial pyramid

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

framework combined with local features extractors, such as such as Scale Invariant Feature Transform (SIFT) [7] and Speeded Up Robust Features (SURF) [8], was proposed by Lazebnik et al. [9]. This method showed significantly improved performance on challenging scene categorization tasks. The image recognition system used in our participation is based on the improved BOW and multi-class classifiers.

Experimental results have shown that the image recognition system provides promising results, in particular to the object recognition task. Among four systems, the proposed system ranked second using the number of cluster $k = 400$ and number of images $M = 150$ for training the vocabulary.

This paper is described as follows. Section 2 presents the image recognition system used by our group in the robot vision challenge. The experiments and results of the proposed system are described in Section 3. Finally, conclusions and future works are discussed in Section 4.

2 Image Recognition System

In this section, we describe the image recognition system used in the challenge. This system can be described into 3 steps: i) feature extraction; ii) spatial pyramid matching; iii) classification. The following sections describe each step in details.

2.1 Feature Extraction

In the feature extraction step, the system extracts SIFT descriptors from 16×16 patches computed over a grid with spacing of 8 pixels. For each patch i, 128 descriptors are calculated, i.e., it is calculated a vector $\varphi_i \in \Re^{128}$. To train the visual vocabulary, we perform k-means clustering of a random subset of descriptors $D = \{\varphi\}$ from the training set according to Equation 1. Throughout the paper, the number of clusters will be referred to as k and the size of the random subset of images will be referred to as M.

$$C = \text{k-means}(D) \tag{1}$$

where $C \in Re^{k \times 128}$ represents the clusters.

Then, each vector descriptor φ_i is associated to the closest cluster according to the Euclidean distance (Equation 2). The index associated is usually called visual word in the bag-of-visual-word approach.

$$\lambda_i = \arg \min_{j=1}^{k} |\varphi_i, C_i| \tag{2}$$

where $|.|$ is the Euclidean distance.

2.2 Spatial Pyramid Matching

The pyramid matching was proposed to find an approximate correspondence between two sets, such as histograms. It works placing a sequence of grids over the space and calculating a weighted sum of the number of matches. Consider a sequence of grids at resolution $l = 0, \ldots, L$. A grid at level l has 2^{dl} cells, where d is the space dimension which in our case is $d = 2$. In each cell i, it calculates the histogram $H^l(i)$ of visual words λ. The number of matches at level l for two image X e Y is given by:

$$I^l = \sum_{i=1}^{2^{dl}} \min\left(H_X^l(i), H_Y^l(i)\right) \tag{3}$$

In order to penalize matches at larger cells, the pyramid match kernel between images X e Y, considering all levels l, is given by:

$$\kappa^L(X,Y) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1} I^l} \tag{4}$$

The kernel above is calculated to each visual word, such that:

$$K^L(X,Y) = \sum_{j=1}^{k} \kappa^L(X_j, Y_j) \tag{5}$$

where X_j indicates that the kernel will consider only the visual word j.

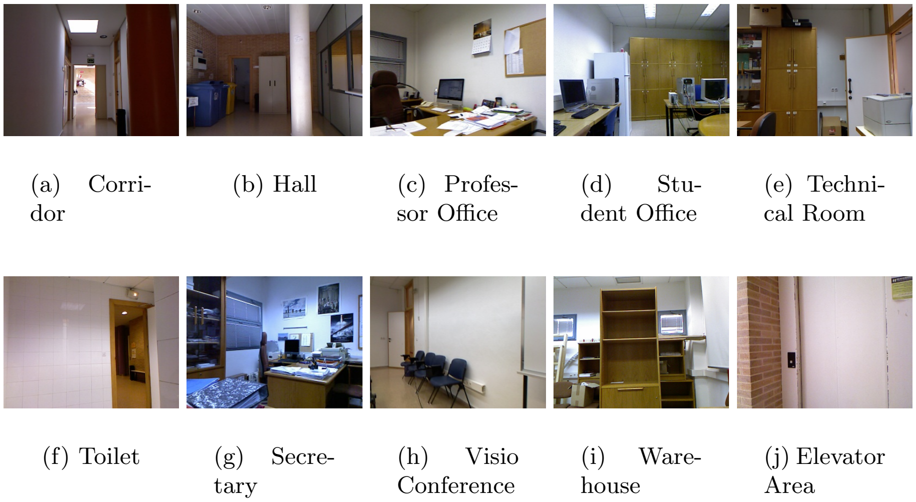
2.3 Classification

The multi-class classification is done with the k-nearest neighbor using the kernel K^L described above. Given a test image, it calculates the kernel value for all training images and assigns the room/category of the closest training image. The same procedure is done for object recognition, i.e., it is detected the presence of the objects of the closest training image.

3 Experiments and Results

In this section we describe the experiments and results of the proposed system. To train the system, we have used 5000 visual images divided into 10 rooms/categories: Corridor, Hall, ProfessorOffice, StudentOffice, TechnicalRoom, Toilet, Secretary, VisioConference, Warehouse, ElevatorArea. An example of each category can be seen in Figure 1. The train dataset also provides 8 objects: Extinguisher, Phone, Chair, Printer, Urinal, Bookself, Trash, Fridge. Examples of images containing each of the objects can be seen in Figure 2. The number of images for each room/category and object is summarized in Table 1.

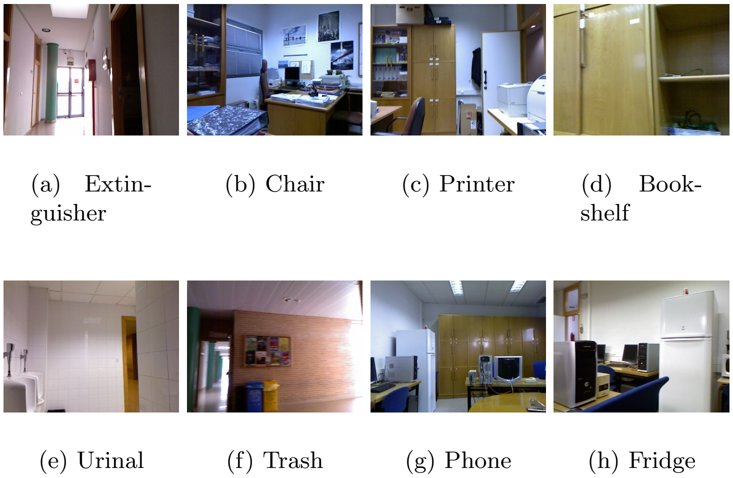
To test the proposed system, we have used the validation dataset composed by 1500 images. The results for different values of number of cluster k and images



(a) Corridor (b) Hall (c) Professor Office (d) Student Office (e) Technical Room

(f) Toilet (g) Secretary (h) Visio Conference (i) Warehouse (j) Elevator Area

Fig. 1. Example of each room/category from the training dataset. Images have 640 × 480 pixels.



(a) Extinguisher (b) Chair (c) Printer (d) Bookshelf

(e) Urinal (f) Trash (g) Phone (h) Fridge

Fig. 2. Example of each object from the training dataset. Images have 640 × 480 pixels.

Table 1. Number of images from the training dataset for the room and object recognition.

Category	N. Images	Perc. (%)
Corridor	1833	36.66
Hall	306	6.12
ProfessorOffice	355	7.10
StudentOffice	498	9.96
TechnicalRoom	437	8.74
Toilet	389	7.78
Secretary	336	6.72
VisioConferene	364	7.28
Warehouse	174	3.48
ElevatorArea	308	6.16

Object	N. Images	Perc. (%)
Extinguisher	770	15.40
Chair	1304	26.08
Printer	473	9.46
Bookself	802	16.04
Urinal	162	3.24
Trash	813	16.26
Phone	267	5.34
Fridge	190	3.80

used to obtain the vocabulary M can be seen in Table 2. The final size of the descriptor is given by $k \times \sum_{l=0}^{L} 2^{2l}$. For $L = 2$ and $k = 300$, the final size is 6300. Despite high number of descriptors, the system takes on average 0.9545 seconds to process an image. For each image, our system provided the room category and the presence of objects. The scores shown in the table are the sum of all the scores obtained for the images. The rules shown in Table 3 are used when calculating the final score for an image.

Table 2. Results in the validation dataset for different values of k and M.

Parameters	Score Rooms	Score Object	Score Total
$k = 300, n_t = 50$	952.5	1051	2003.5
$k = 300, n_t = 150$	960	1044.75	2004.75
$k = 400, n_t = 50$	952.5	1038.5	1991
$k = 400, n_t = 150$	963	1049.25	2012.25
$k = 500, n_t = 50$	948	1014.25	1962.25
$k = 500, n_t = 150$	966	1042	2008

Table 3. Rules for calculating the score for place localization and object recognition.

The room category has been correctly classified:	+1.0 points
The room category has been wrongly classified:	-0.5 points
The room category has not been classified:	0.0 points
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 (False Negative):	-0.25 points

Finally, the Table 4 shows the results for all participations in the robot vision challenge. Three groups have submitted solutions to this challenge and the baseline indicates the results obtained by the dataset provided script (Color & Depth histogram + SVM). For the competition we submitted four runs with different values for the parameters k and M (see Section 2.1). Our system ranked second, achieving 1738.75 points on this task for $k = 400$ and $M = 150$.

Table 4. Results for all groups in the robot vision challenge. The baseline indicates the results by using Color & Depth histogram + SVM.

#	Group	Score Rooms	Score Objects	Score Total
1	NUDT	1075.50	3357.75	4430.25
2	UFMS	219.00	1519.75	1738.75
3	Baseline Results	67.5	186.25	253.75
4	AEGEAN	-405	-995	-1400

4 Conclusions

This paper described the participation of our group in the Robot vision challenge. In this challenge, the proposed system ranked second among four others systems. Thus, the image recognition system has shown promising results, particularly for object recognition.

As future work, we intend to extend the approach for color images and use other classifiers, such as Support Vector Machine, which it is known to provide better results than k-nearest neighbors. In addition, the system will be applied in the depth images.

Acknowledgments.

RCG and LCR were supported by the CNPq and PET-Fronteira, respectively.

References

1. Ulrich, I., Nourbakhsh, I.: Appearance-based place recognition for topological localization. In: Robotics and Automation, 2000. Proceedings. ICRA '00. IEEE International Conference on. Volume 2. (2000) 1023–1029
2. Martinez-Gomez, J., Garcia-Varea, I., Cazorla, M., Caputo, B.: Overview of the imageclef 2013 robot vision task. In: Working Notes, CLEF 2013. (2013)
3. Caputo, B., Müller, H., Martinez-Gomez, J., Villegas, M., Acar, B., Patricia, N., Marvasti, N., Üsküdarlı, S., Paredes, R., Cazorla, M., Garcia-Varea, I., Morell, V.: ImageCLEF 2014: Overview and analysis of the results. In: CLEF proceedings. Lecture Notes in Computer Science. Springer Berlin Heidelberg (2014)

4. Martinez-Gomez, J., Cazorla, M., Garcia-Varea, I., Morell, V.: Overview of the ImageCLEF 2014 Robot Vision Task. In: CLEF 2014 Evaluation Labs and Workshop, Online Working Notes. (2014)
5. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: In Workshop on Statistical Learning in Computer Vision, ECCV. (2004) 1–22
6. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2. ICCV '03, Washington, DC, USA, IEEE Computer Society (2003) 1470–
7. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal Computer Vision **60**(2) (2004) 91–110
8. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. **110**(3) (2008) 346–359
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR '06, Washington, DC, USA, IEEE Computer Society (2006) 2169–2178