Expert Systems With Applications 79 (2017) 181–193

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

On robot indoor scene classification based on descriptor quality and efficiency

Cristina Romero-González*, Jesus Martínez-Gómez, Ismael García-Varea, Luis Rodríguez-Ruiz

Departamento de Sistemas Informáticos, Universidad de Castilla-La Mancha, Campus Universitario s/n. Albacete, Spain

ARTICLE INFO

Article history:
Received 29 August 2016
Revised 24 February 2017
Accepted 25 February 2017
Available online 28 February 2017

Keywords:
Indoor scenes
Semantic classification
Multi-source classification
Descriptor combination
Spatial pyramid technique

ABSTRACT

Indoor scene classification is usually approached from a computer vision perspective. However, in some fields like robotics, additional constraints must be taken into account. Specifically, in systems with low resources, state-of-the-art techniques (CNNs) cannot be successfully deployed. In this paper, we try to close this gap between theoretical approaches and real world solutions by performing an in-depth study of the factors that influence classifiers performance, that is, size and descriptor quality. To this end, we perform a thorough evaluation of the visual and depth data obtained with an RGB-D sensor to propose techniques to build robust descriptors that can enable real-time indoor scene classification. Those descriptors are obtained by properly selecting and combining visual and depth information sources.

1. Introduction

Nowadays, general purpose robots are not common at home environments, since the current state of the art in robotics is still not capable of fully solving some crucial problems. One of the challenges is to make a robot understand where it is located to properly behave and interact with its environment. This is known as indoor scene classification (Wu, Christensen, & Rehg, 2009), and it can be defined as the problem of describing the place where the robot currently is by means of a set of pre-defined labels (bathroom, corridor or kitchen can be examples of the labels typically used).

In this work, we rely on two different type of images to solve the indoor scene classification problem, both acquired with a RGB-D sensor. The first type corresponds to typical visual images representing color information, and the second type is depth images that encode depth information in the color values. An example of a visual and its corresponding depth image, acquired from the same viewpoint, are shown in Fig. 1.

Convolutional Neural Networks (CNNs Lee, Grosse, Ranganath, & Ng, 2009) are the state-of-the-art method for computer vision applications. However, this approach involves some non trivial requirements. Specifically, CNNs require large training sets to converge, which are hard to obtain. The classification problem that we face in this paper has to deal with a specific dataset oriented to indoor environments suitable for robotic applications. This results in specific classes that are not available in general pre-trained CNN models, and with a relatively small number of samples to generate discriminant CNN classifiers. Additionally, CNNs are usually computationally demanding.

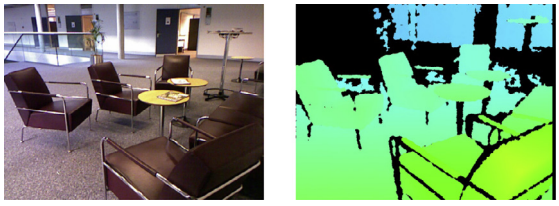
Consequently, we must rely on a more classical approach, that can run in real-world scenarios, even with a system with limited resources (such us most robotic platforms). In general, the indoor scene classification problem is addressed using machine learning techniques, where a model is trained from a set of samples. So, we can identify the following steps for designing and building an indoor scene classification system:

1. Extract a set of relevant and distinctive features from every image in a training dataset.
2. Group or map those features into a set of image descriptors.
3. Design a classifier capable of discriminating among the different types of scenes, and train it using the descriptors built in the previous step.

In this paper, we are going to focus on the second step. This way, state-of-the-art techniques and models will be adopted for the remaining two tasks (feature extraction and classification). Specifically, we will employ SIFT-based features (Lowe, 2004) and Online Independent Support Vector Machines (OISVMs) (Orabona, Castellini, Caputo, Jie, & Sandini, 2010) as classification models. An overview of this process is shown in Fig. 2.

* Corresponding author.
E-mail addresses: Cristina.RGonzalez@uclm.es (C. Romero-González), Jesus.Martinez@uclm.es (J. Martínez-Gómez), Ismael.Garcia@uclm.es (I. García-Varea), Luis.RRuiz@uclm.es (L. Rodríguez-Ruiz).

http://dx.doi.org/10.1016/j.eswa.2017.02.040
0957-4174/



(a) RGB image (b) Depth image

Fig. 1. Example of visual and depth images from the RobotVision@ImageCLEF 2012 dataset.

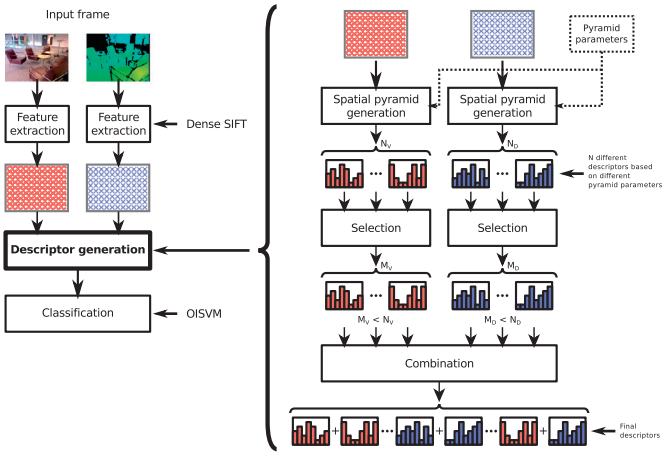


Fig. 2. Overview of the classification process, and the contribution of this paper in descriptors generation, selection and combination.

Therefore, the main goal of this work is to address the problem of defining, selecting and combining a set of initial descriptors to work in real-world scenarios. Here, we rely on the spatial pyramid approach, as it is described in Section 3.1, to obtain the initial set of descriptors. This method generates a global descriptor for each image based on the spatial distribution of the same set of features on it, which results in a high correlation between the representation at different levels. This might suggest that specific configurations (specific levels or type of divisions) are key for a distinctive representation of the image. At the same time, in Section 5 we establish a direct relation between descriptor size and computation time. So, an appropriate trade-off between descriptor quality and size has to be achieved to both obtain representative descriptors and do it efficiently. We conducted an extensive evaluation of the visual and depth data obtained, which leads to a data driven procedure to select the best parameters for a classifier.

Finally, we would like to remark the inclusion in the process of real-time constraints in the form of descriptor size. This results in a procedure that can be used in practical scenarios of semantic classification, and ensures its applicability in platforms with different processing capabilities.

The remainder of the paper is organized as follows. Section 2 provides an overview of related work in indoor scene classification. In Section 3, we present the formal description and formulation of the indoor scene classification problem and how to define the descriptors used (Section 3.1). Section 4 sets the experimental framework for the rest of the sections. In Section 5 we study some requirements to perform this task in real-time. Section 6 explains how the first subset of these descriptors are generated and selected. In Section 7, we describe the descriptor combination procedure and present the results. These results are discussed in Section 8. In Section 9 we compare our results with the submissions of the Robot Vision challenge at ImageCLEF 2012. Finally, Section 10 presents the main conclusions that can be drawn from this work.

2. Related work

Indoor scene classification is still considered an open problem where approaches for outdoor environments perform poorly. This is mainly caused by the high variability associated to indoor scenes (Quattoni & Torralba, 2009).

The entire process consists in acquiring an image, generating a suitable representation (image descriptor), and classifying (labeling) the scene. This classification can be performed according to (a) high-level features of the environment, like detected objects, (b) global image representations, or (c) local features. Moreover, several descriptors can be combined to increase the robustness of the classifier.

In spite of the fact that some works from late nineties addressed the indoor scene classification problem (Yamauchi & Langley, 1997), it was between 2002 and 2006 when it was formulated as it is known today and, since then, several proposals have been made. In Torralba, Murphy, Freeman, and Rubin (2003) a method for scene classification based on global image features was presented. In that work, the temporal continuity between consecutive images was exploited using a Hidden Markov Model (HMM). In Martinez Mozos, Stachniss, and Burgard (2005), on the other hand, a scene classifier with range data as input information and AdaBoost as the classification model is proposed. In 2006, Pronobis, Caputo, Jensfelt, and Christensen (2006) developed a visual scene classifier using composed receptive field histograms (Linde & Lindeberg, 2004) and Support Vector Machines.

The use of the Bag of Words (BoW) technique (Csurka, Dance, Fan, Willamowski, & Bray, 2004; Fei-Fei & Perona, 2005) can also be considered a remarkable milestone for visual scene classification. The BoW process starts by creating a visual dictionary of representative features. Next, each extracted feature is assigned to the closest word in the dictionary. Then, a histogram representing the number of occurrences of each visual word is computed. This histogram is finally used as the image descriptor.

In Lazebnik, Schmid, and Ponce (2006) is presented a spatial pyramid as an extension of the BoW technique. Their proposal allows to merge local and global information into a single image descriptor. The spatial pyramid approach has been successfully applied to several semantic localization problems (Everingham, Van Gool, Williams, Winn, & Zisserman, 2010; Griffin, Holub, & Perona, 2007; Xiao, Hays, Ehinger, Oliva, & Torralba, 2010) and it can be considered a standard solution for generating descriptors.

In 2010, Pronobis et al. introduced a novel cue integration scheme (Pronobis, Martinez Mozos, Caputo, & Jensfelt, 2010). Such scheme allows to properly combine information extracted from different robot sensors (camera and laser), as well as to combine information from the same source (e.g. two image descriptors generated from different visual local features). Feature combination has also been studied in similar problems, like multiclass object detection (Gehler & Nowozin, 2009). Proposals for integrating information from different sources became more popular with the arrival of the Microsoft Kinect sensor (Martinez Mozos, Mizutani, Kurazume, & Hasegawa, 2012; Silberman & Fergus, 2011). All these works show how crucial the information encoded in the descriptors is for a data driven learning. For that reason, obtaining descriptors that are both efficient and accurate is a fundamental problem to be addressed in this field.

3. Problem formulation

The indoor classification problem can be formulated as a classical statistical pattern recognition problem as follows. Let I be an