This is the accepted version of:

F. Rubio, J. Martínez, M. J. Flores and J.M. Puerta
Integration of Contextual Information into the Scene
Classification Problem

Robotics and Autonomous Systems 97 (2017)

<https://doi.org/10.1016/j.robot.2017.08.010>

Please, visit the provided url to obtain the published version.

Integration of contextual information into the scene classification problem

Fernando Rubio^{a,*}, Jesus Martínez-Gómez^a, M. Julia Flores^a, José M. Puerta^a

^a*Computing Systems Department, University of Castilla-La Mancha, Spain*

Abstract

The task of identifying the semantic localization of a robot has commonly been treated as a classification problem, where images are taken as input and a set of predefined labels is the output. While traditional approaches have focused on the performance of the image features extracted from computer vision techniques, the contextual information that can come with the images has not been taken into account. In this work, we present an approach for integrating this information in a scene classification pipeline where we opt for Bayesian network classifiers in addition to standard support vector machine ones. The approach is evaluated in two scenarios, one in which the contextual information is directly provided with the images, and the other where it must be inferred in an additional stage. The evaluation was performed using two families of classifiers over two datasets, and the results obtained show how the scene classification problem can benefit from the integration of contextual information.

Keywords: scene classification, descriptor generation, machine learning, robotics

*Corresponding author. Tel.: +34 967 599200 ext. 2677; fax: +34 967 599224.

Email addresses: `Fernando.Rubio@alu.uclm.es` (Fernando Rubio),
`Jesus.Martinez@uclm.es` (Jesus Martínez-Gómez), `julia.flores@uclm.es` (M. Julia Flores), `jose.puerta@uclm.es` (José M. Puerta)

1. Introduction

Localization is one of the main focuses in the field of robotics, as any robot must understand its current localization in order to make appropriate decisions. The robot localization problem has evolved enormously by integrating novel sensors and computer vision techniques, but also by using different representations. In addition to the metric representation of the environment, where a robot’s location is defined by its (x, y, z, θ) pose, we can adopt a semantic representation of the environment, which is commonly known as semantic localization [1].

Semantic localization defines the location of a robot by means of semantic categories that describe the surrounding scene [2]. The semantic categories or labels describe the environment as a human does, and “office”, “corridor” or “kitchen” are common options when dealing with indoor locations. The use of this semantic representation improves the human-robot interaction process, as humans can tell the robot to move to the “dining room” instead of using metric locations as navigation goals.

Given that cameras are the main robot sensor [3], semantic localization is usually addressed as a classification problem in which input images must be categorized using a prior set of semantic labels. In this process, we can identify two main tasks: extracting the appropriate features from the input images, and selecting the optimal classification model to be fed with these features. Regarding the image features, state-of-the-art techniques such as HOG [4], SIFT [5] or SURF [6] extract continuous descriptors that have proven to work well, especially when combined, with support vector machine (SVM) classifiers [7].

In addition to the information that is self-encoded in the images, these can be annotated with contextual information related to miscellaneous aspects such as the overall quality of the image, a list of keywords describing it, or the type of image from a set of pre-defined options such as social, animals, sports or landscape. These annotations should be inferred when dealing with images acquired via the robot, but they may be available when working with some of the existing image databases¹. As opposed to the descriptors extracted with computer vision techniques, the contextual annotations are encoded with discrete values.

In this article, we propose the use of Bayesian network classifiers (BNCs)

¹<http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm>

to properly integrate the contextual information with the descriptors extracted from the images. In a recently published work [8], BNCs have produced promising results when used to predict scene classification labels. Furthermore, the capabilities of this paradigm seem to provide a favourable environment for interpretability and for the natural integration of distinct sources of “input information”.

We evaluate this approach by selecting two datasets in which images acquired with a robot are annotated both with the semantic category of the scene and some additional contextual information. Over these datasets, we evaluate three scenarios using state-of-the-art image descriptors, and incorporating SVMs in the comparison: 1) a complete lack of contextual information; 2) access to this contextual information assuming that it is fully available; and 3) the contextual information is inferred by using estimators trained from data. Scenario 1 corresponds to the classical scene classification problem in which no annotations are available. Scenario 2 represents a non-realistic (ideal) situation where input images have contextual annotations, and it is used to show the maximum improvements that can be achieved by the integration of this information. Finally, Scenario 3 proposes a realistic benchmark where the effective integration of the contextual information is evaluated. A complete scheme for our proposal can be seen in Fig. 1, where green represents our contribution and red corresponds to traditional approaches. The last step represents the validation of the experimentation, which shows the effectiveness of our proposal.

The rest of the paper is organized as follows. In Section 2 we introduce the semantic localization problem, and describe both the descriptor extraction procedure and the classification models of two families: SVMs and BNCs. The methodology to infer the contextual information and its subsequent integration is explained in Section 3. Section 4 presents the experimental procedure and the datasets proposed for the experimentation. Finally, in Section 5 and Section 6, respectively, the experimental results and conclusions are presented.

2. Related Work

2.1. Semantic Localization

The generic problem of localization in robotics consists of answering the question “where am I?” [9]. In this task, robots acquire data through their

Figure 1: Flow chart showing the proposal presented in this work. Our contributions are marked in green, while traditional approaches are in red.

sensors (e.g. cameras) and estimate their location by interpreting such information. Semantic localization is a specific type of localization in which the robot provides semantic information about the surrounding environment by means of natural language. There are different types of semantic information but recent efforts have focused on the acquisition of semantic maps. Recent literature [10] considers contextual information and scene classification as fundamental aspects of semantic localization. As there are some deficiencies in the overall management of these tasks, this work represents an initial approach in the use of different contextual information to increase the performance of scene classifiers.

The most common interpretation of scene classification is by means of semantic annotations describing the robot’s surroundings [11]. In this case, the representation of the environment is provided by a set of predefined labels. Several ways of defining the environment by means of labels exist but the use of scene categories is the most widely accepted one [12]. Based on this approach, the location of the robot can be given by using terms such as “office”, “toilet” or “basement”.

Scene classification is a supervised classification problem formulated as follows. Let I be an input image, $g(I) = \mathbf{x}$ a function that generates an image descriptor $\mathbf{x}$ from I , and M a classification model that provides the class posterior probability $P_M(c|\mathbf{x})$, where c is a class label from a predefined set of semantic locations $\mathcal{C}$. Thus, without loss of generality, the problem is to find the optimal label $\hat{c}$ according to the following equation:

$$\hat{c} = \arg \max_{c \in \mathcal{C}} P_M(c|g(I)) \quad (1)$$

Using state-of-the-art functions for generating descriptors from images, $g(I)$, the problem consists of selecting the most appropriate classification model, M . In addition to this selection, the image descriptor ($\mathbf{x}$) may require a pre-processing stage to meet the requirements of the selected model. For instance, we may face this scenario when using continuous image descriptors in conjunction with discrete classification models.

2.1.1. Descriptors

The descriptor generation process is commonly carried out in two stages: extraction of local features in the overall image, and then the obtaining of a fixed-dimensionality global descriptor from them. Local features encode information such as color, shape or gradients. While each local feature presents a fixed-dimensionality (e.g. SIFT is encoded with 128 elements), the number of extracted features depends on both the keypoint detection algorithm – e.g. canny edge detector [13] or uniform sampling – and the size of the input image. This leads to descriptors of which the dimensionality is not constant. Most common classification models, such as support vector machines or Bayesian networks, do not support this kind of features, a fact which encourages the generation of fixed-dimensionality descriptors using the extracted local features as input.

Some techniques extract global descriptors that can be used directly in the classification task, such as GIST [14], Centrist [15] and HOG [4], which have proven their value in a wide range of applications. On the other hand, local features such as SIFT [5], SURF [6] and ORB [16] need a posterior step to obtain a fixed-dimensionality descriptor. In the literature, the most frequently used technique for obtaining a global descriptor from local features is the Bag-of-Words process [17]. This procedure is based on a prior clustering stage using the extracted local features as input. Each cluster represents a “word”, and the complete set of n words defines the codebook or

dictionary. This dictionary makes it possible to represent any input image as a bag of words by mapping the extracted local features to words. Finally, a histogram based on the word frequencies is computed and used as a global image descriptor.

2.1.2. Classification

Traditionally, classification in semantic localization tasks is performed by means of support vector machines (SVMs). These are a set of supervised learning methods that have been widely applied in many problems. Here, we provide a very brief introduction to them. For a comprehensive introduction to the subject, please refer to [18] or [19]. Assume $\{\mathbf{x}_i, y_i\}_{i=1}^l$, with $\mathbf{x}_i \in R^m$ and $y_i \in \{-1, 1\}$, is a set of samples and labels drawn from an unknown probability distribution; we wish to find a function $f(\mathbf{x})$ such that $\text{sign}(f(\mathbf{x}))$ best determines the category of any future sample $\mathbf{x}$. In the most general setting,

$$f(\mathbf{x}) = \sum_{i=1}^l \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \quad (2)$$

where $b \in R$ and α_i are Lagrangian coefficients. In practice, most of the α_i are found to be zero after training; the vectors with an associated α_i different from zero are called support vectors. The function $K(\mathbf{x}_1, \mathbf{x}_2)$ represents the type of kernel and it should be selected according to the input data and the problem. With this brief introduction, the selection of the most suitable classifier has been dealt with as the search for the best kernel for a specific problem.

Most classification models can handle multi-label classes but the models (SVMs) we have described above can only deal with binary ones. To address this limitation, multi-label SVM classifiers can be created from binary ones by using one-versus-one or one-versus-all strategies [20].

2.2. Contextual Information

Descriptions of the semantic localization of a robot usually consist of annotations about the environment. These annotations are labels that determine aspects such as the lighting conditions, presence/absence of objects, applied aesthetic filters and so on. This set of tags, known as contextual information, permits a better human comprehension of the localization. As can be seen, scene classification can be understood as a kind of contextual

information, and these labels for localization are very closely related to other types of contextual information.

Research into labelling contextual information has evolved enormously recently, and several methods exist for inferring semantic information from images. Traditionally, contextual information was manually labelled, and there are several databases encoding such information [21]. As the variety of these tags can be limited, we can opt to use APIs capable of providing lexical labels describing any input image. For instance, we can find the Google Cloud Vision² (GCV) API that relies on large scale datasets to automatically generate these labels, or Amazon Mechanical Turk³ (AMT), in which labelling is performed by human agents known as “Turkers”. However, the labels provided by GCV are too generalist or heterogeneous to solve specific problems. Moreover, different problems have appeared in the AMT platform, some of them related to cultural factors and others related to the misunderstanding of the specified requirements [22].

2.3. Bayesian network classifiers

Bayesian network classifiers (BNCs) have never had a significant impact on scene classification due to the poor interpretability of the descriptors. A competitive solution was proposed in [23], which uses probabilities and the naive Bayes theorem to solve the classification problem. Nevertheless, our approach implies the use of contextual information in the models, and now the interpretability of this information provides a substantial benefit for probabilistic models. For this reason we provide a brief introduction to BNCs below.

A Bayesian network (BN) [24] consists of two components: a) a directed acyclic graph (DAG) of which the nodes represent the random variables $\mathbf{X}$ in the problem, and where the set of directed edges connect pairs of vertices representing the direct dependencies between variables; b) the numerical part of the BN corresponds to the relationship between variables quantified by conditional probabilities, usually in the form of tables (CPTs).

Bayesian network classifiers (BNCs) are Bayesian networks in which one of the variables represents the class C , and each possible label of the class is a state of this variable. In all BNCs the main task is to compute the posterior

²<https://cloud.google.com/vision/>

³<https://www.mturk.com/mturk/welcome>

maximum probability of the class variable conditioned to the input values of a new object $\vec{x}$, i.e., under the maximum a posteriori hypothesis (MAP) we assign the label with maximum probability:

$$c_{MAP} = \operatorname{argmax}_{c_l} P(c_l|\vec{x})$$

where $P(c_l|\vec{x})$ is computed based on the BN yielding the different BNC models.

2.3.1. Naive Bayes

The naive Bayes (NB) classifier is the simplest BNC, and has a fixed structure (see Fig. 2). This simple structure codifies that any pair of predictive variables X_i and X_j are conditionally independent given the class variable C . This assumption is hard to hold with real data but it makes this model extremely efficient.

Figure 2: BNC structures for n features + class: NB (left), TAN (middle), and SPODE (right).

Despite its simplicity, NB has shown very good results in real applications such as computing, marketing and medicine [25].

2.3.2. Semi-naive models

Several techniques have been proposed in order to improve the accuracy of NB by relaxing the attribute interdependence assumption. We can refer to them as semi-naive classifiers, a term introduced by Kononenko in [26]. All this family of techniques shares the fact that the variable C is the parent of all variables X_i . In this paper, we focus on two well-known models, namely TAN (tree augmented naive Bayes) and AODE (averaged one-dependence estimator). We selected these models due to the good performance shown in previous works, and the fact that they still maintain a very good tradeoff between complexity and effectiveness.

The **TAN** model [27] relaxes the conditional independence restriction, allowing only one arc between variables X_i (see Fig. 2), without a large increase in the complexity of computation.

Learning TAN models is divided into two phases: a) the structural learning, a connected tree with variables X_i , and b) the parameter estimation of the conditional distribution $P(X_i|X_j, C)$ and $P(C)$.

AODE [28] avoids the structural learning phase, as does the NB model, by using an ensemble of 1-dependence classifiers but with a structure that is previously fixed for each model of the ensemble. AODE selects a limited ensemble of 1-dependence classifiers and aggregates the predictions in order to obtain the class. Each of the models in the ensemble is named SPODE (super parent one dependence estimator). The structure for each SPODE can be seen in Fig. 2. Similarly to the NB model, all the variables X_j have the class variable C as parent but in this case another variable X_i is also parent of the remaining variables X_j , i.e. we have a super-parent with the Cartesian product of the variable X_i and the class variable C in an NB model. Therefore, if we have n variables and the class, we have n SPODE classifiers.

3. Integration of Contextual Information

The main objective of the current article is to integrate contextual information into the semantic localization problem, which is seen as a scene classification task. We try to solve Eq. 1, where the contextual information is included. However, the contextual information can be present in many ways, e.g. the lighting conditions when the image was taken, the aesthetic techniques applied to the image, the presence/absence of objects, and so on. This contextual information is usually discrete, and for instance a common label for lighting is “sunny” instead of the brightness level, which is a continuous value.

Labels can be transformed into discrete variables in different ways, and it depends on whether the information is exclusive, i.e. if the lighting conditions are “sunny” it cannot be “night”, or not, i.e. more than two objects can be in the same image. If the contextual information is exclusive, we can take two approaches: 1) create a new variable with as many values as the contextual information can be labelled with, and 2) create a binary variable for each value the contextual information can take. On the other hand, if the information is not exclusive, the most common option is to create as many binary variables as the contextual information we have. Fig. 3 shows two kinds of contextual information in different images and the possible variables that we can create.

Figure 3: Variable generation from contextual information. Extraction of binary variables from non-exclusive information (left). Two possible options for exclusive information (right): binary or multivalued variables.

Suppose that L represents a set of n variables obtained from contextual information where $L = \{l_1, \dots, l_n\}$. If we modify Eq. 1 to take into account the contextual information, the result we obtain is Eq. 3:

$$\hat{c} = \arg \max_{c \in \mathcal{C}} P_M(c|g(I), L) \quad (3)$$

In this equation we have continuous variables corresponding to $g(I)$ and discrete ones from L . Now, our problem has an input of continuous and discrete values, which implies that the use of different models (P_M) is not trivial. Another alternative consists in preprocessing continuous data to discrete values. However, some descriptors from the image cannot be directly transformed into discrete values.

In addition to the representation of the contextual information, we evaluate two scenarios in this proposal. The first one assumes that the contextual information is directly provided with the input images, while the second scenario involves an additional stage in which such information must be inferred. The first scenario represents an ideal situation, and allows us to evaluate the integration capabilities of the proposal. On the other hand, the second scenario represents the realistic end application of the semantic localization system, in which we may face errors in the inference of the contextual information. The inference of the contextual information is managed as a preliminary stage, which consists of a set of classifiers using the image descriptor as input. Each classification model from this stage separately

handles a variable from the contextual information. A scheme for these two propositions can be seen in Fig. 4.

Figure 4: Experimentation pipeline where contextual information is assumed as available (ideal situation) or inferred from the images (realistic situation).

4. Experimental Procedure

4.1. Datasets

We selected two datasets for the evaluation of the proposal: KTH-IDOL2 [29], and the ViDRILO dataset [30]. These datasets were obtained within indoor environments using a mobile robot. Each image in the dataset is annotated with the semantic category of the scene in which it was acquired. More details are provided below.

4.1.1. KTH-IDOL2

The KTH Image Database for rObot Localization (IDOL2) [29] consists of 24 image sequences acquired with two mobile robot platforms: a PeopleBot (Minnie) and a PowerBot (Dumbo). These sequences were generated within an indoor environment, which consists of five rooms of different functionality (one-person office, two-person office, corridor, kitchen, and printer area), and under three lighting conditions (cloudy weather, sunny weather, and night).

Each image is annotated with the semantic category of the scene in which it was acquired. We select the subset of images acquired with the PeopleBot since the position of its camera is more similar to most current robots. Fig. 5 shows exemplar images for each semantic category in the dataset. In addition, Fig. 6 presents the same scene acquired under three lighting conditions. The class distribution and the lighting conditions can be seen in Table 1.

Figure 5: Semantic categories used in the KTH-IDOL2 dataset.

Figure 6: Lighting conditions of the images from the KTH-IDOL2 dataset.

Table 1: Semantic categories and lighting conditions distribution in the KTH-IDOL2 dataset.

Scene	Printer Area	Corridor	1-Person Office	2-Person Office	Kitchen	All
Cloudy	488	1632	458	518	656	3752
Sunny	482	1582	468	518	556	3606
Night	534	1704	522	633	611	4004
All	1504	4918	1669	1448	1823	11362

4.1.2. ViDRILO

ViDRILO [30] is a dataset containing 5 sequences of annotated RGB-D images acquired with a mobile robot, in two office buildings under challenging lighting conditions. Sequences in the dataset are annotated with 10 semantic categories (e.g. corridor, kitchen, or office) within an indoor office environment. An example of these images can be seen in Fig. 7. In addition, the presence or lack of 15 objects (e.g. fridge, chair or computer) is also included in the ground truth of the dataset. Fig. 8 shows an example of each object, and the distribution of the categories and objects is presented in Table 2 and 3, respectively. In order to extract the same descriptors from the two selected datasets, the depth information encoded in the RGB-D images from ViDRILO is not considered. This allows us to focus the study on the classification models.

Figure 7: Semantic categories used in the ViDRILO dataset.

Table 2: Distributions of the annotations in the ViDRILO dataset for semantic categories

Scene	CR	EA	HA	PO	SE	SO	TO	TR	VC	WH
#Images	10273	1269	1181	1722	1284	2166	993	1341	1320	905

Table 3: Distributions of the annotations in the ViDRILO dataset for objects

Object	Exting.	Urinal	Computer	Screen	Chair	Trash	Printer	Fridge
#Images	4723	421	3308	568	5937	5796	1684	395
Object	Table	Board	Bookshelf	Sink	HD	Phone	Bench	-
#Images	6214	6312	4299	190	1461	395	2914	-

Figure 8: Examples of each annotated object in the ViDRILO dataset.

4.2. Descriptors and Classifiers

Using the images included in the KTH-IDOL2 and ViDRILO datasets as input, we first extract the two descriptors presented in Section 2. The first one is the Histogram of Oriented Gradients (HOG), which extracts a global histogram from the image. In this work, the HOG implementation corresponds to [31], where just level 0 is used (whole image) and we are able to select the size n of the resulting histogram. The second descriptor is the combination of SIFT local features with the Bag-of-Words (BoW) process. The whole process is known as the Histogram of Visual Words (HoVW), and we are also able to select the size n of the BoW clustering, which corresponds to the size of the descriptor.

Taking into account that both descriptors are encoded as a sequence of continuous values, we need to perform a discretization step to use them as input in any discrete classification model. For this purpose we use Fayyad & Irani’s MDL method [32], which selects the best intervals for the continuous data by employing mutual information. Furthermore, we evaluate different descriptor dimensionalities by: a) merging adjacent angles in the HOG computation; and b) selecting different numbers of visual words, n , during the HoVW computation. Specifically, we opt for four dimensionalities that are shared by the two descriptors: 50, 100, 200 and 300. The combination of the discretization and dimensionality reduction stages results in different sets of

discrete predicting attributes.

With respect to classification, four models are evaluated: SVMs, NB, TAN and AODE. The first one corresponds to the implementation of the LIBSVM toolbox [33] with all its default parameters, i.e. the strategy to deal with the multi-nominal problems is one-versus-one. With respect to the kernels, two of them are assessed:

- linear kernel (SVM_{lin}): it is one of the most common kernel functions, and is given by the inner product plus an additional coefficient: $\mathbf{x}_1 \cdot \mathbf{x}_2 + coef_0$.
- χ^2 kernel (SVM_{χ^2}): this kernel comes from the χ^2 distribution and is the best of those studied for feature vectors based on histograms:

$$1 - \sum_{i=1}^n \frac{(\mathbf{x}_1 - \mathbf{x}_2)^2}{\frac{1}{2}(\mathbf{x}_1 + \mathbf{x}_2)}$$

With regards to NB, just the discrete version of the classifier is considered. Traditional implementations of TAN and AODE can only use discrete variables. BN implementations correspond to Weka⁴ and its default parameters. This results in two classifiers (SVM_{lin} and SVM_{χ^2}) that can cope with continuous data, and three classifiers (NB, TAN and AODE) that need to carry out the discretization step to use the descriptor information as input.

Each classification model is evaluated by means of a 5-fold cross validation (CV) using the accuracy as measure, as is widely accepted in the evaluation of the performance of classifiers. In order to maintain the class distribution in the evaluation, we perform a stratified fold selection. Furthermore, the same fold selection is maintained during the evaluation of different classifiers and descriptors to avoid the effect of randomness.

4.3. Contextual Information

In this work, the contextual information refers to the annotations that are integrated in the two selected datasets, namely KTH-IDOL2 and ViDRILO. Specifically, this information corresponds to the lighting conditions (cloudy, night or sunny) during the image acquisition in KTH-IDOL2, as well as the presence or lack of 15 objects in the scene in ViDRILO. The integration is performed by adding binary variables to the image descriptor extracted. In particular, 3 binary variables are added to either HOG or HoVW when using

⁴<http://www.cs.waikato.ac.nz/ml/weka/>

the KTH-IDOL2 dataset, and 15 variables when ViDRILO is considered. With respect to the objects, the contextual information is not exclusive, as a binary variable is made for each object. Despite lighting conditions being exclusive, we select the binary representation in order to have more variables for experimentation.

In Section 3 we discuss two scenarios with respect to the contextual information: ideal and realistic situations. The first one assumes that the contextual data is known, and this assumption is made to understand the behaviour of classification models when these variables are included. In this way, we can appreciate whether their contribution improves the model’s behaviour, which is what one would expect since the input information is richer. The second way to obtain contextual information is by means of learning a classification model. To do so, we create a classifier for each binary variable, where the input is the information extracted by the descriptors. Therefore, some of the attributes we will use in our scene classifier (lighting conditions and presence of objects) will be predicted by an ”inner” classifier. In this case, we expect to observe how the error produced by these classification models that predict the contextual information is propagated to those models in charge of the localization. It may also be that the predictions are quite good. The point is that this would be a more realistic test because we cannot obtain these values from an image automatically unless it is externally annotated.

5. Results

Three experiments were carried out to evaluate the proposal. The first experiment aims to perform an initial selection of the classification model. It is based on the baseline experimentation, in which scene classification solely relies on the descriptor extracted from the images. The second experiment focuses on the inference of the contextual information, which is integrated in the semantic localization system in the third experiment. The last experiment also evaluates the integration of the contextual information when it is directly provided and not inferred. All the experimentation was performed using two datasets, namely KTH-IDOL2 and ViDRILO, and the results obtained are analysed and discussed.

5.1. Scene classification from image descriptors

Five classification models were evaluated in the first experiment. The first two classifiers are support vector machines that differ in the type of kernels: linear (SVM_{lin}) and χ^2 (SVM_{χ^2}). These two classifiers can cope with continuous data, while the other classification models (NB_d , TAN and AODE) needed a Fayyad-Irani pre-processing stage to discretize the input data. In addition to the five classification models and two datasets, we evaluated two image descriptors as input, namely HOG and HoVW, using four dimensionalities: 50, 100, 200, and 300. For all these parameters, we computed the accuracy for scene classification using 5-fold cross validation. Fig. 9 shows the results obtained, and it can be observed how AODE and SVM_{χ^2} outperformed the other classifiers. HoVW obtained higher accuracy values than HOG, and both descriptors benefited from larger dimension values for most of the classifiers. This tendency, however, was not maintained for TAN and SVM_{lin} , which displayed unexpected behaviour.

5.2. Contextual information inference from image descriptors

The second experiment focuses on the inference of contextual information using the image descriptor as input. We tested several classification models in the same way as in the previous section but here for predicting contextual information in both datasets from the descriptors. In this experiment, we considered 3 binary variables representing the lighting conditions in the KTH-IDOL2 dataset, and 15 representing the presence or lack of each object annotated in the ViDRILO dataset. An independent classifier was trained from each one of these variables using as input the different histograms from the HOG and HoVW descriptors.

We performed a more detailed study of the five classifiers to obtain a thorough prediction of the contextual information. This study revealed AODE and SVM_{χ^2} , in conjunction with HoVW_{300} , to be the two combinations of classification model and descriptor with the best performance. Taking into account that our study focuses on the subsequent integration of such contextual information, we finally selected AODE. We made this choice because AODE classifications are provided with the probability encoding the confidence of the decision, which can be exploited for different purposes.

As we wish to evaluate binary classifiers in contrast to multilabel ones, we selected the AUC (Area Under the Curve) metric instead of accuracy. The results obtained with the combination of AODE and HoVW_{300} are shown in Tables 4 and 5 for the KTH-IDOL2 and ViDRILO datasets, respectively.

Figure 9: Accuracy for the five classifiers with different n values in each descriptor, HOG (top) and HoVW (bottom), and for each dataset, KTH-IDOL2 (left) and ViDRiLO (right).

Table 4: AODE classifier’s AUC for inferring the lighting conditions in KTH-IDOL2.

	Cloudy	Night	Sunny
AUC	0.742	0.843	0.856

Table 5: AODE classifier’s AUC for inferring object presence in ViDRiLO

Object	Exting.	Urinal	Computer	Screen	Chair	Trash	Printer	Fridge
AUC	0.881	0.983	0.935	0.964	0.924	0.8388	0.914	0.970
Object	Table	Board	Bookshelf	Sink	HD	Phone	Bench	-
AUC	0.942	0.876	0.927	0.990	0.985	0.936	0.954	-

We can first conclude that both lighting conditions and object presence/absence can be well inferred using the proposed classifiers. However,

AODE is more suitable for identifying the objects than for inferring the lighting conditions.

5.3. Scene classification from image descriptors and contextual information

The third experiment evaluates the scene classification when contextual information is integrated in conjunction with the image descriptors. Therefore, we combined the image descriptor (either HOG or HoVW) with the contextual information annotated in each dataset (lighting conditions or object presence) to generate a new input for the classification pipeline. Two new scenarios were evaluated in this experiment but they only differ in the classification stage, as mentioned above. The ideal scenario assumed the availability of contextual information, and we directly used the annotations included in the dataset. On the other hand, the realistic scenario inferred the contextual information using the systems developed in the previous experiment. Both scenarios were compared with the baseline one, in which the scene is classified using only the descriptor extracted from the image.

The experimentation included the three most promising classification models (SVM_{χ^2} , TAN and AODE), two datasets (KTH-IDOL2 and ViDRILO) and two descriptors, namely HoVW and HOG. Regarding the descriptor dimension, we considered four options (50, 100, 200, and 300) in order to discover the relationship between the descriptor dimension and the improvements obtained. Using 5-fold cross validation, we computed the accuracy of scene classification and obtained the results that are shown in Fig. 10.

The integration of the contextual information in the ideal scenario always improved the accuracy of the baseline classifier, as we see in the differences between the red and green bars. In Fig. 10, the bottom graphs from the ViDRILO dataset show substantial gains for the ideal scenario. This increase is considerably greater in comparison with KTH-IDOL2 because contextual annotations (presence of objects in the scene) provide more discriminative capabilities. With respect to the image descriptor, the room for improvement is greater for those with the lowest baseline accuracy. In other words, the highest accuracy increases were obtained when using low-dimensionality descriptors, especially for HOG. This could be expected as the weight of the contextual variables in the classification directly depends on the initial size of the descriptor. In the above graphs, low-size descriptors (such as those with 50 bins) have a big impact on classification where contextual information is known. Regarding the classification model, SVM_{χ^2} performed a more

Figure 10: Scene classification accuracy when the contextual information is not integrated (red), assumed as available (green), and inferred in an additional stage (blue).

suitable integration than AODE and TAN in the ideal scenario. This can be seen in more detail when analyzing Fig. 11.

The results obtained were substantially different for the realistic scenario. The SVM χ^2 classifier was negatively affected when integrating contextual information that had been inferred previously. In fact, the accuracy of the scene classifier was lower than the baseline result in most of the combinations of dataset and descriptor. In Fig 10, we see only two cases in which the SVM χ^2 baseline results (red) are surpassed by the realistic ones (blue). On the other hand, AODE and TAN properly integrated the inferred contextual information and obtained accuracies that are closer to the ideal scenario than the baseline. The blue bars in AODE and TAN are closer to the green ones than to the red. Again, the highest accuracy increases centred on the low-dimensionality HOG descriptor.

The improvement or degradation in accuracy is graphically shown in more detail in Fig. 11. From this figure, we can draw the following conclusions. First, KTH-IDOL2 is not suitable for improving in the current proposal as the contextual information included in it is not valuable. Indeed, lighting conditions do not provide information about the scene in which the robot is situated. On the other hand, the presence or lack of certain objects can contribute to determining the location of the robot. Second, there is a rela-

Figure 11: Accuracy variation of ideal (solid) and realistic (dashed) situation for SVM (red), AODE (blue) and TAN (green) classifiers.

tionship between the descriptor’s dimension and the potential accuracy increase. Low-dimensional descriptors are more suitable for a further increase in scene classification accuracy when integrating contextual information. In this specific problem, a stable solution is reached with 300 variables, while larger-sized descriptors obtain similar or worse results. We select 300 bins to maximize the contribution of the contextual information. Finally, even when the ideal accuracy increase for TAN and AODE is not as high as for SVM_{χ^2} , the realistic scenario obtained results that are closer to this increase. This means that SVM_{χ^2} is more sensitive to incorrect classifications of contextual instances, while TAN and AODE are more robust against the propagated errors of contextual information classifiers.

6. General conclusions and further work

We have presented and evaluated the integration of contextual information in combination with image descriptors for the problem of scene classification. The integration has been carried out using two scenarios in which the information was assumed as available, or inferred in a preliminary stage. The proposal was evaluated by using two datasets (KTH-IDOL2, ViDRILLO) and two image descriptors (HOG, HoVW), as well as classification models belonging to the SVM or BNC families. On the basis of the results obtained,

we can conclude that contextual annotations are a valuable source of information for robot scene classification. However, the proper integration of such information depends on the classification model and the combination of descriptor and dataset used for the experimentation. SVMs were shown to be the optimal classification model when the contextual information is directly provided, but their performance was severely impaired when such information is inferred, which can result in misinferences. On the other hand, the TAN and AODE classifiers carried out a promising integration of inferred contextual information.

Our interest in using BNCs in the integration of contextual information is due to the fact that BNCs can handle this process in a natural way through their related (in)dependences, and they can deal with probabilities through the inference. In the experimentation, the output probability from the contextual information inference was not exploited. We plan to investigate how these probabilities could be used for generating more robust models. We also have the extension of this proposal with richer contextual information in mind. This information can be provided by other tools such as the Google Cloud Vision API, or even automatically generated by services such as Amazon Mechanical Turk.

Acknowledgements

This work has been partially funded by FEDER funds and the Spanish Government (MICINN) through project TIN2013-46638-C3-3-P, and TIN2015-65686-C5-3-R, and by Consejería de Educación, Cultura y Deportes of the JCCM regional government through projects PEII-2014-049-P and PPII-2014-015-P.

References

- [1] Andreas Nüchter and Joachim Hertzberg. Towards semantic maps for mobile robots. *Robotics and Autonomous Systems*, 56(11):915–926, 2008.
- [2] I. Kostavelis and A. Gasteratos. Semantic mapping for mobile robotics tasks: A survey. *Robotics and Autonomous Systems*, 66:86–103, 2015.
- [3] Jesus Martínez-Gómez, Antonio Fernández-Caballero, Ismael García-Varea, Luis Rodríguez, and Cristina Romero-González. A taxonomy of

- vision systems for ground mobile robots. *Int J Adv Robot Syst*, 11:111, 2014.
- [4] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893 vol. 1, 2005.
 - [5] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
 - [6] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *Computer vision and image understanding*, 110(3):346–359, 2008.
 - [7] Andrzej Pronobis, O Martinez Mozos, Barbara Caputo, and Patric Jensfelt. Multi-modal semantic place classification. *The International Journal of Robotics Research*, 29(2-3):298–320, 2010.
 - [8] Fernando Rubio, Jesus Martínez-Gómez, M Julia Flores, and José M Puerta. Comparison between bayesian network classifiers and svms for semantic localization. *Expert Systems with Applications*, 64:434–443, 2016.
 - [9] John J Leonard and Hugh F Durrant-Whyte. Mobile robot localization by tracking geometric beacons. *IEEE transactions on robotics and automation*, 7(3):376–382, 1991.
 - [10] Qiang Liu, Ruihao Li, Huosheng Hu, and Dongbing Gu. Extracting semantic information from visual data: A survey. *Robotics*, 5(1):8, 2016.
 - [11] Stephanie Lowry, Niko Sünderhauf, Paul Newman, John J Leonard, David Cox, Peter Corke, and Michael J Milford. Visual place recognition: A survey. *IEEE Transactions on Robotics*, 32(1):1–19, 2016.
 - [12] J. Wu, H. I. Christensen, and J. M. Rehg. Visual place categorization: Problem, dataset, and algorithm. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, pages 4763–4770. IEEE, 2009.

- [13] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.
- [14] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.
- [15] Jianxin Wu and Jim M Rehg. Centrist: A visual descriptor for scene categorization. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1489–1501, 2011.
- [16] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. IEEE, 2011.
- [17] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval*, pages 197–206, 2007.
- [18] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines: And Other Kernel-based Learning Methods*. Cambridge University Press, New York, NY, USA, 2000.
- [19] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.
- [20] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *The Journal of Machine Learning Research*, 5:101–141, 2004.
- [21] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.
- [22] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. Running experiments on amazon mechanical turk. *Judgment and Decision making*, 5(5):411–419, 2010.
- [23] Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Learning deep nbnn representations for robust place categorization. *arXiv preprint arXiv:1702.07898*, 2017.

- [24] Kevin B. Korb and Ann E. Nicholson. *Bayesian Artificial Intelligence, Second Edition*. CRC Press, Inc., 2nd edition, 2010.
- [25] M. Julia Flores, José A. Gámez, and Ana M. Martínez. Supervised classification with Bayesian networks. In Rafael Magdalena et al., editor, *Intelligent Data Analysis for Real-Life Applications: Theory and Practice*, pages 72–102. IGI Global, 2012.
- [26] Igor Kononenko. Semi-naive bayesian classifier. In *Machine Learning EWSL-91*, pages 206–219. Springer, 1991.
- [27] N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.
- [28] Geoffrey I Webb, Janice R Boughton, and Zhihai Wang. Not so naive bayes: aggregating one-dependence estimators. *Machine learning*, 58(1):5–24, 2005.
- [29] Jie Luo, Andrzej Pronobis, Barbara Caputo, and Patric Jensfelt. The KTH-IDOL2 Database. Technical Report CVAP304, KTH Royal Institute of Technology, CVAP/CAS, Stockholm, Sweden, 2006.
- [30] Jesus Martínez-Gómez, Ismael García-Varea, Miguel Cazorla, and Vicente Morell. Vidrilo: The visual and depth robot indoor localization with objects information dataset. *The International Journal of Robotics Research*, 34(14):1681–1687, 2015.
- [31] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.
- [32] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *Thirteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1022–1027. Morgan Kaufmann Publishers, 1993.
- [33] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.