FULL PAPER

Scene Classification from Semantic Labeling

José Carlos Rangel[a,*], Miguel Cazorla[a], Ismael García-Varea[b], Jesus Martínez-Gómez[b], Élisa Fromont [c]
and Marc Sebban [c]

[a]*Computer Science Research Institute. University of Alicante*
P.O. Box 99. E-03080. Alicante. Spain;
[b]*University of Castilla-La Mancha*
Albacete. Spain;
[c]*Jean Monnet University*
Saint Etienne.France

(*v1.0 released June 2015*)

Finding an appropriate image representation is a crucial problem in robotics. This problem has been classically addressed by means of computer vision techniques, where local and global features are used. The selection or/and combination of different features is carried out taking into account repeatability and distinctiveness, but also the specific problem to solve. **In this article, we propose the generation of image descriptors from general purpose semantic annotations.** This approach has been evaluated as source of information for a scene classifier**, and specifically using Clarifai as semantic annotation tool**. The experimentation has been carried out using the ViDRILO toolbox as benchmark, which includes a companion of state-of-the-art global features and tools to make comparisons among them. According to the experimental results, the proposed descriptor performs similarly as well-known image descriptors based on global features in a scene classification task. Moreover, the proposed descriptor is based on generalist annotations without any type of problem-oriented parameter tuning.

Keywords: Scene classification, Semantic Labeling, Machine Learning, Data Engineering

1. Introduction

The scene classification or indoor place categorization problem could be defined as the problem of classifying an image as belonging to a scene category from a set of predefined labels [1]. This problem is closely related to the semantic localization one, and it helps to identify the surrounding of an agent, like a mobile robot, by means of scene categories like corridor or kitchen. Scene classifiers are also helpful for specific robotic tasks [2] like autonomous navigation, high-level planning, simultaneous location and mapping (SLAM), or human-robot interaction.

Scene classification is commonly addressed as a supervised classification process [3], where input data correspond to perceptions, and classes to semantic scene categories. Current approaches are based in a two-stage building process: a) select the appropriate descriptors to be extracted from perceptions, and b) choose a classification model to be able to deal with the extracted descriptors.

Relying on the use of images as the main perception mechanism, the descriptor generation problem is tackled with computer vision techniques. In this process, the organization of the data extracted from the images plays an important role. This is clearly exposed in two of the most

---

*Corresponding author. Email: jcrangel@dccia.ua.es

widely-used approaches: the Bag-of-Words (BoW) [4, 5] and the spatial pyramid [6]. This two approaches allow for the generation of fixed-dimensionality descriptors, required for most of the state-of-the-art classification models, built from any type of local features.

There exist, however, novel approaches proposing the use of categorical information instead of numeric image descriptors. For instance, Fei et al. propose in [7] the use of an Object Filter Bank for scene recognition, where the bank is built upon image responses to object detectors previously trained. Lampert et al. present in [8] an object recognizer based on attribute classification. This proposal relies on a high-level description that is phrased in terms of semantic attributes, such as the shape or the color of the object. The novel approach proposed in [9] represents images by using the objects appearing in them. This high-level representation encodes both object appearances and spatial information.

The use of Deep Learning (DL) is considered a remarkable milestone in the research areas of computer vision and robotics [10]. DL provides classifiers capable not only to classify data but also to automatically extract intermediate features. This technique has been applied to image tagging with surprising results. For instance, the Clarifai team won the 2013 Imagenet competition [11] by using Convolutional Neural Networks [12]. In addition to very large amounts of annotated data for training, DL requires high processing capabilities for classification. While these two requirements are not always met, we can take advantage of some existing solutions that provide the DL capabilities through application programming interfaces (APIs). Clarifai [1] is one of the well-known systems offering remote image tagging. Specifically, any input image is labeled with the semantic categories better describing the image content.

This article proposes a general framework to generate image descriptor from semantic labels. Namely, we rely on the use of the annotation scheme provided by Clarifai, and then use this labels to build image descriptors. The obtained descriptors are evaluated as input for the scene classification problem. We have performed an exhaustive comparison with state-of-the-art global descriptors in the ViDRILO dataset [13]. **The first goal of this paper is then to determinate whether Clarifai descriptors are competitive with other state-of-the-art image descriptors suitable for scene classification. It should be pointed out that Clarifai descriptors are generated from a general purpose labeling system. On the other hand, the rest of image descriptors included in the experimentation have been specifically selected by their scene representation capabilities. This work also aims at discovering (and discussing) the novel capabilities offered with the use of general purpose annotations.**

The rest of the paper is organized as follow: in Section 2, we introduce and formulate the scene classification problem. Section 3 describes the different descriptors used in this study and presents a detailed explanation of the Clarifai system and its performance. In Section 4, the experimental results are presented, and a discussion is carried out in Section 5. Finally, in Section 6 the main conclusions and future works are outlined.

2. Scene classification from semantic labels

The scene classification problem can be formulated as a classical statistical pattern recognition problem as follows. Let I be a perception (commonly an image), $d(I)$ a function that generates a specific descriptor given I, and M a classification model that provides the class posterior probability $P_M(c|d(I))$, where c is a class label from a set of predefined scene categories $\mathcal{C}$. Then, this problem can be established as the problem of finding the optimal label $\hat{c}$ according to:

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

---

[1]http://www.clarifai.com

In our case, I corresponds to a RGB image. The problem then involves two main stages: a) designing the descriptor generation process to obtain an appropriate representation of I ($d(I)$), and b) selecting a classification model capable of discriminating among the set of predefined scene categories.

This work is focused on the first stage: descriptor generation, and we propose representing every image I_i as a sequence of semantic annotations obtained from an external labeling system, namely Clarifai. That is, $d(\cdot)$ is designed as a black box procedure where every image is translated into a set of N labels $\mathcal{L} = \{l_1, \ldots, l_N\}$ corresponding to the semantic annotations obtained from the Clarifai system. While some labels can partially represent an input image I_i, a set of probabilities $\mathcal{W}_i = \{w_{i1} \ldots w_{iN}\}$ is obtained in conjunction to $\mathcal{L}$. Each entry $w_{i,j}$ represents the likelihood of describing the image I_i using the label l_j. Thanks to the use of a predefined set of semantic labels, we can obtain fixed-dimensionality image descriptors as explained in Section 3.

There exist some similarities between this approach and the classical Bag-of-Words [4] one. In both representations, image descriptors consist of a set of terms describing the input perception. However, the main difference comes for the semantic component of our approach. That is, the dictionary of words (or codebook) in the BoW approach does not fully represents semantic concepts, as it is computed from a set of numeric local features in an unsupervised way, usually through a k-means clustering algorithm.

3. Descriptors generation and description

This work proposes the use of Clarifai labels as input for a scene classifier. To evaluate this proposal, we carry out a comparison with classical approaches where descriptors are directly computed from images. This two alternatives are shown in Fig. 1. The classical approach relies on computer vision techniques to extract image descriptors. This stage should be carefully designed to select the appropriate features. In this design, we should also take into account aspects as efficiency, the programming language or the requirements of external libraries dependencies.
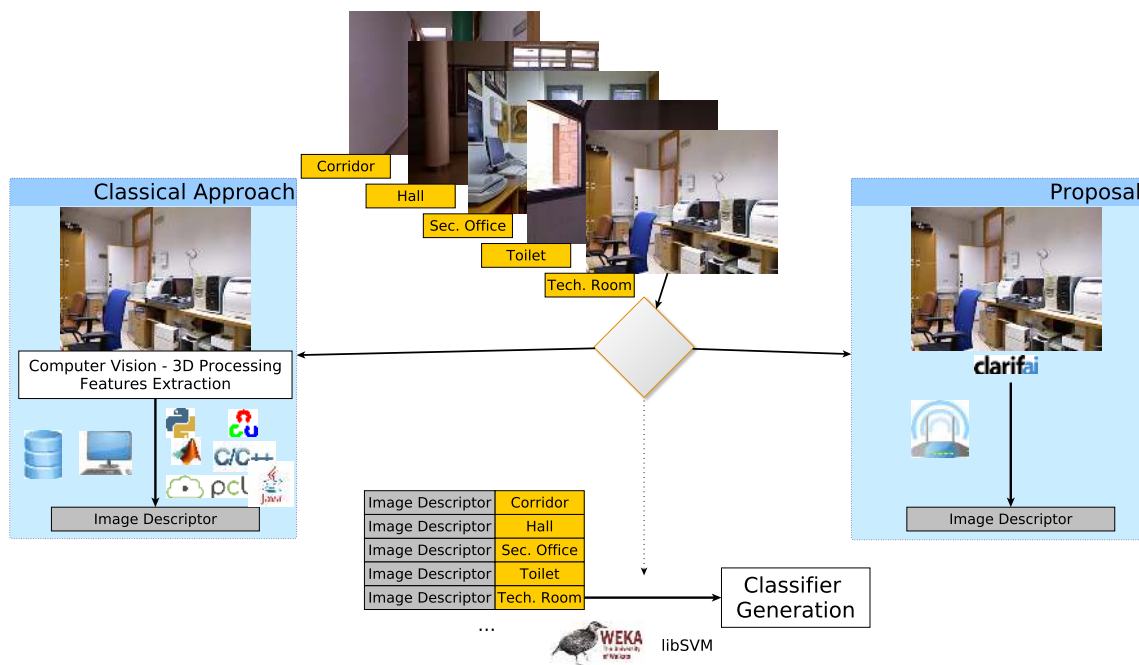

Figure 1. Methodology overall pipeline

On the other hand, the generation of descriptors from Clarifai labels is performed delegating

this step to an external system whose internal details do not need to be known. As it is shown in Fig. 1, the descriptors obtained are used as input for further classification tasks independently from the way they are generated.

In order to validate this approach in the ViDRILO (Visual and Depth Robot Indoor Localization with Objects information) dataset, we compare it against the baseline results obtained with state-of-the-art descriptors presented in [13]. The feature extractions techniques, as well as the use of the Clarifai system, are detailed in the following.

3.1   *Descriptor generation from visual and depth features*

Three baseline descriptors are proposed and released in conjunction with ViDRILO. These baseline descriptors are: Pyramid Histogram of Oriented Gradients [14] (PHOG), GIST [15], and Ensemble of Shape Functions [16] (ESF). Both PHOG and ESF are computed from perspective images, while ESF relies on the use of depth information.

3.1.1   *Pyramid Histogram of Oriented Gradients*

PHOG descriptors are histogram-based global features that combine structural and statistical approaches. This descriptor takes advantage of the spatial pyramid approach [6] and the Histogram of Gradients Orientation [17].

The generation of PHOG descriptors is shown in Fig. 2, and it depends on two parameters: the number of bins of each histogram B, and the number of pyramid levels L. At each pyramid level l_i in $[0 \ldots L]$, this process produces 4^{l_i} histograms with B bins. According to the baseline ViDRILO experimentation, we opt for using $B = 30$ and $L = 2$, which results in a descriptor size of 630 $((4^0 + 4^1 + 4^2) \times 30)$.

Figure 2. Generation of the PHOG Descriptor.

3.1.2   *GIST*

The GIST descriptor is intended to model the shape of a scene by using a holistic representation of the spatial envelope. In the GIST generation process, $S \times O$ transformations are performed, **with scales(S) and orientations(O),** over $N \times N$ patches in the image, as it is presented in Fig. 3. This transformations allow to represent each path by a low-dimensional $(S \times O)$ vector, which encodes the distribution of O orientations and S scales in the image along with a coarse description of the spatial layout. Using the standard implementation, the dimensionality of the GIST descriptors is 512 $(N = 4, O = 8, S = 4)$.

3.1.3   *Ensemble of Shape Functions*

The Ensemble of Shape Functions (ESF) **is a 3D point cloud descriptor that** consists of a combination of 3 different shape functions that result in 10 different histograms. It encodes relationships between the points in the cloud. **These functions encode the distance, the area and the angle between a set of random points in the cloud using 3 histograms for each function. These histograms are:** *in* **(inside the space),** *out* **(outside the**
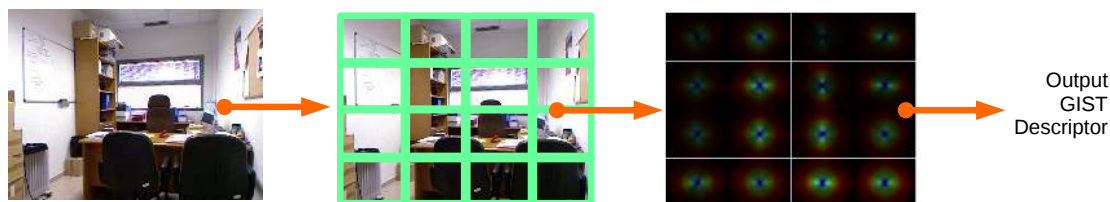
Figure 3. Generation of the GIST Descriptor.

space) or *mixed* (a combination of them). **An additional histogram encoding the lying distance of the cloud is also integrated in the ESF descriptor.** On the contrary to other 3D descriptors, ESF does not require normal information, which makes it robust to noise and partial occlusions. Each histogram contains 64 bins, and the final dimension of the descriptor is 640. **Fig. 4 shows the three histograms (in, out and mixed) obtained for the specific distance function.**
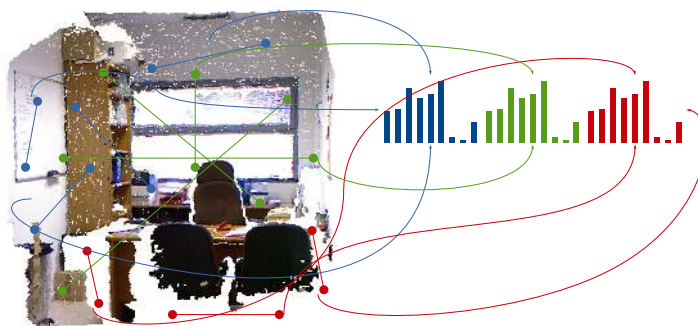

Figure 4. Generation of the ESF Descriptor.

3.2 *Descriptor generation from the Clarifai system*

Clarifai [18] is a research project aiming at developing a high-level image (and video) processing system by means of Convolutional Neural Networks (CNNs [12]). CNNs are hierarchical machine learning models that learn a complex representation of images using vast amounts of data. They are inspired by the human visual system and learn multiple layers of transformations, which extract a progressively more sophisticated representation of the input. Clarifai's working scheme is shown in Fig. 5. **Clarifai started as a research group that presented a solution to the image classification problem using the ImageNet dataset [11]. However, the Clarifai service is now a closed system whose details about the datasets used for training (which determine the dimension of the decision layer) and internal architecture are not provided. Therefore, we state the maximum number of the dimension for the extracted descriptors in a preliminary stage where we discover all the annotations that are extracted from the dataset. This is similar to the codebook identification when applying a Bag-of-Words approach.**

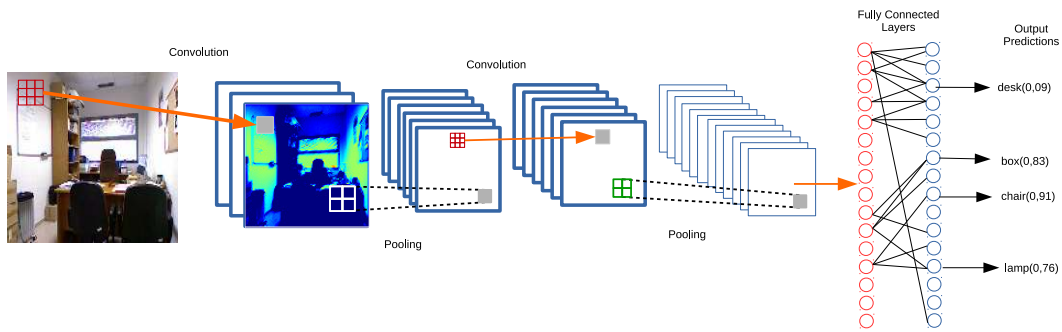We propose the use of the Clarifai labeling system as a black box procedure through its

Figure 5. Processing Scheme of the Clarifai API.

application programming interface (API[1]). From this API, we can automatically tag perspective images based on its content. It also provides with the presence probability of each label in the image.

There exists a commercial version of the Clarifai API with non-restricted access to advanced features of the system. However, we use a free trial version that presents some limitations, like the number of requests by hour. This version tags each input image with the 20 most feasible semantic labels, as well as their associated probabilities. This information is used to generate the image descriptors in this proposal.

The dimensionality-fixed descriptor generation process is carried out as follows. We firstly discover the total number of labels representing our dataset by: a) submitting all the images from our dataset to Clarifai, and b) identifying the unique values in the output label list. Then, we encode each image I_i using a sparse representation whose dimension corresponds to the length of the entire list $N = |\mathcal{L}|$. The descriptor consists of a set of entries $\mathcal{W}_i = \{w_{i,1}, ..., w_{i,N}\}$. Each entry $w_{i,j}$ contains the probability of representing I_i with label l_j only when the Clarifai response to this image request includes information for l_j. Otherwise, the entry $w_{i,j}$ encodes the lack of information about l_j using a zero value.

Using ViDRILO as dataset we obtain $N = 793$ different labels. Therefore, each Clarifai descriptor has a dimensionality of 793, even when it contains information only about 20 different labels. An example of this process (using 4 instead of 20 labels by service request, for clarity) is shown in Table 1. The semantic labels used in the Clarifai system represent meaningful concepts and the majority of them ($\approx 80\%$) are nouns. Some exemplar labels are: animal, flower, plant, sport, structure, vehicle or person.

Table 1.   Clarifai descriptor generation from sparse semantic labeling

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	...	l_{793}
	0.97	0.95	0.93	0.91	0	0	0	...	0
	0	0.94	0	0.92	0.93	0.94	0	...	0
	0.91	0	0	0	0.94	0.97	0.93	...	0
...	...	...	...	...	...	...	...	...	...
	0	0	0	0	0	0	0	...	0.91

6

4.   Experimental framework

All the experiments included in this work have been carried out using the ViDRILO dataset [13][1]. The main characteristics of this dataset are shown in Table 2, which provides five different sequences of RGB-D images captured by a mobile robot within an indoor office environment. **The dataset was acquired during a span of twelve months in order to provide inter sequence variability.**

Table 2.   Overall ViDRILO sequences distribution.

Sequence	Number of Frames	Floors imaged	Dark Rooms	Time Span	Building
Sequence 1	2389	1st,2nd	0/18	0 months	A
Sequence 2	4579	1st,2nd	0/18	0 months	A
Sequence 3	2248	2nd	4/13	3 months	A
Sequence 4	4826	1st,2nd	6/18	6 months	A
Sequence 5	8412	1st,2nd	0/20	12 months	B

Each RGB-D image is annotated with the semantic category of the scene it was acquired, from a set of ten different room categories. In Fig. 6, representative images for all the ten room categories are shown. **Different sequences from ViDRILO dataset were used as benchmark in the RobotVision challenge at ImageCLEF competition [19] during its last editions.**

| Corridor | Elevator Area | Hall | Professor Office | Secretary Office |
| Student Office | Toilet | Technical Room | Video Conf. Room | Warehouse |

Figure 6. Exemplar visual images for the 10 room categories in ViDRILO .

In the experimentation, several combinations of descriptors and classification models are evaluated in different scenarios. The main goal of the experiments is to determine the discrimination capabilities of Clarifai descriptors. Each scenario includes a training and a test sequence used to generate and evaluate the scene classifier, respectively. This evaluation computes the accuracy as the percentage of test images correctly classified with their ground truth category. Based on the five ViDRILO sequences, we are faced with 25 different scenarios (from Sequence 1 vs Sequence 1 to Sequence 5 vs Sequence 5). In addition to Clarifai, the 3 baseline descriptors previously described are evaluated: PHOG, GIST, and ESF. Regarding the classification model, we opted for three well-known alternatives: Random Forests (RFs) [20], k-Nearest Neighbor (kNN) [21], and chi-squared kernel SVM [22]; as proposed in the ViDRILO baseline experimentation.

Regarding the efficiency of the experimentation, any combination of baseline descriptor generation (PHOG, GIST, and ESF) and classification model (RFs, kNN and SVMs) allows for real-time processing. They were successfully evaluated under

---

[1]The ViDRILO dataset can be freely download from http://www.rovit.ua.es/dataset/vidrilo/

a 30fps frame rate on a Microsoft Kinect device in a preliminary stage. However, the running time for the Clarifai labeling highly depends on the internet connection and the overload of the system. This resulted in large time variations (in the range $[0.15 - 1.25]$ seconds per image), and it comes from the fact that there is no offline alternative to the online Clarifai labeling system.

4.1   *Evaluation of Clarifai as visual descriptor*

In the first experiment, we generated a Clarifai descriptor from all the images in the ViDRILO dataset. This process was carried out by following the methodology proposed in Section 3. Then, we integrated the Clarifai descriptors in the experimentation stage proposed by the dataset authors. **This experimentation proposed the evaluation of the five sequences available in the dataset, using them to firstly train a classifier and then, using another sequence to test the classifier. We followed this process using the combination of baseline classifiers and descriptors. The integration of Clarifai descriptors in the experimentation was performed by using them as input for the classifiers, as it was done for the PHOG, GIST, and ESF.**
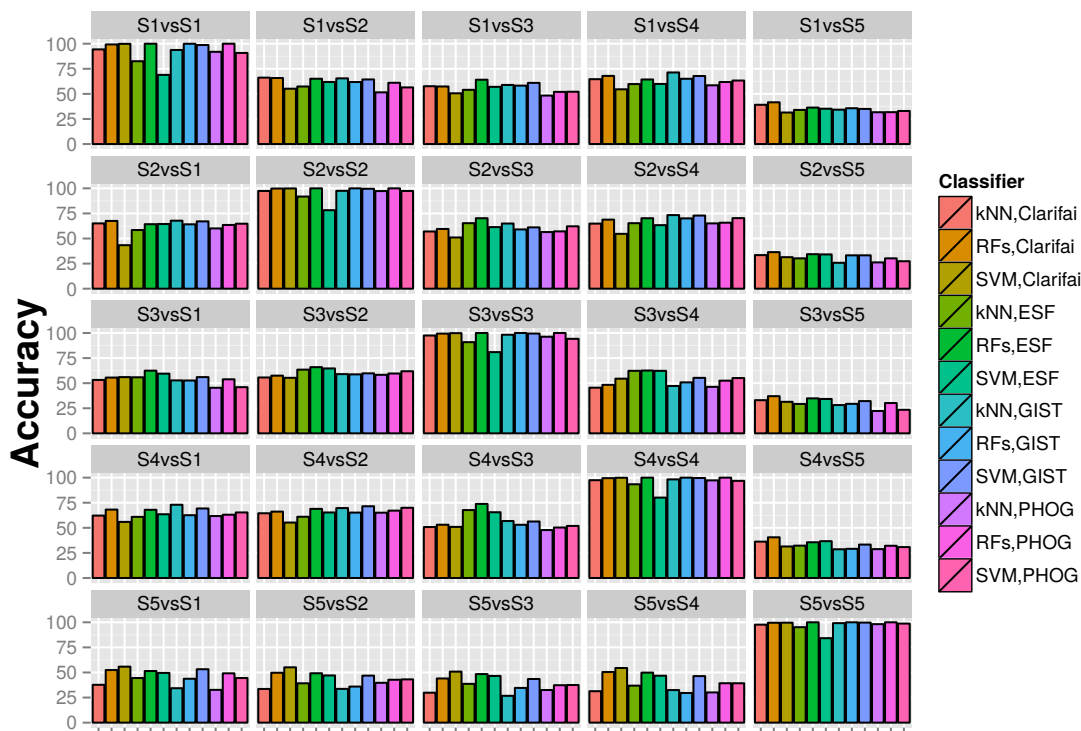

Figure 7. Accuracy obtained for all the classifiers, descriptors and training/test combinations.

The obtained results are shown in Fig. 7, **where each chart title denotes the sequences used for the experiment. That is, SAvsSB, means that the sequence A has been used to train the model, whereas the sequence B has been used to test such model.** The large amount of data avoids from extracting conclusions without a posterior analysis. Therefore, we post-processed the data to obtain the mean accuracy over the training/test combinations. **Fig. 8 graphically presents these results grouping them accordingly to the sequences combination. Fig. 8 left shows the mean accuracy obtained when using the same sequence for training and test. Fig. 8 right shows the mean accuracy for every combination of different training and test sequences.**
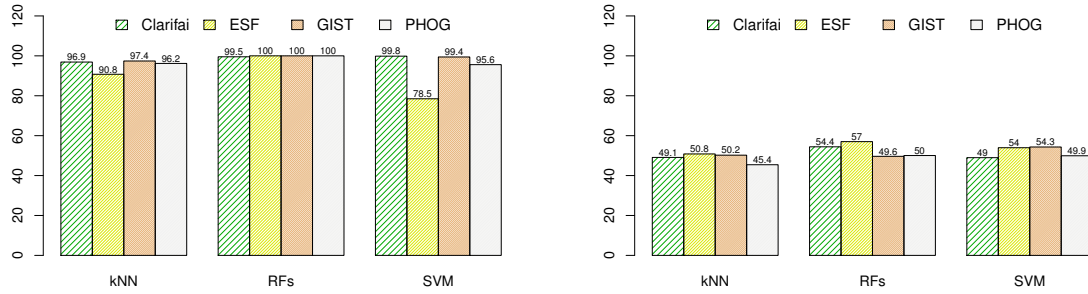
Figure 8. Accuracy averaged over the training/test combinations, using the same sequence(left) and using different sequences(right).

In order to carry out a fair comparison, we performed a Friedman test [23] with a 0.05 confidence level. The **null** hypothesis established all the descriptor/classifier combinations as equivalent when facing the scene classification problem. **This hypothesis was rejected, which encouraged us to follow a post-hoc statistical analysis as described in [24] to find out which of these combinations outperformed the rest. The comparison was done against the best combination, namely ESF with RFs, and obtained the values that are shown in Table 3. In view of these results, we can state that using Clarifai as descriptor, we can generate scene classifiers (using RFs as classifier) as competitive as those generated with ESF and GIST. It has been also exposed how Clarifai allows for results significant better than those obtained with PHOG**.

Table 3. Post-hoc analysis comparison for all descriptors/classifiers combination against the best combination(ESF/RFs)

Descriptor,Classifier	p-value	Rejected	Rank	Win	Tie	Loss
RFs,ESF	-	No	2.70	-	-	-
SVM,GIST	2.2433e-01	No	4.22	19	0	6
RFs,Clarifai	2.2433e-01	No	4.32	14	0	11
SVM,ESF	1.5541e-03	Yes	6.24	23	0	2
RFs,PHOG	7.6841e-04	Yes	6.54	20	5	0
RFs,GIST	7.6841e-04	Yes	6.56	20	4	1
kNN,GIST	6.7068e-04	Yes	6.64	19	0	6
SVM,Clarifai	4.5236e-05	Yes	7.30	21	0	4
SVM,PHOG	2.4383e-05	Yes	7.46	22	0	3
kNN,Clarifai	9.3896e-06	Yes	7.68	20	0	5
kNN,ESF	2.7732e-06	Yes	7.94	25	0	0
kNN,PHOG	4.7705e-13	Yes	10.40	25	0	0

4.2 *Coping with domain adaptation*

If we review the specifications of the sequences included in ViDRILO, we find out the Sequence 5 as the only one that has been acquired in a different building. Evaluating a scene classifier in an environment not seen previously makes the problem even more challenging. This has been exposed in Fig. 7, where the poorest results are obtained using Sequence 5 for training or test, **while when using Sequence 5 for both training and test the results are good.**

These scenarios (where Sequence 5 appears) evaluate the generalization capabilities of a scene classification system. Namely, the knowledge acquired during training should be generalist

enough to be applied to different environments. Therefore, we decided to perform an additional comparison between all the descriptors and classifiers in the 8 scenarios where Sequence 5 is used for training or test. The results obtained are presented in Fig. 9, where we can extract some interesting remarks. Firstly, all the methods ranking first by scenario use Clarifai as descriptor.

Moreover, these methods use a SVM classifier when using Sequence 5 for **training** (Fig. 9 **left**), and a Random Forest classifier when this sequence is used for **test** (Fig. 9 **right**). **This figure share the same notation that one used in the Fig 7.** The difference between the classification models can be explained by the number of images included in the ViDRILO sequences (see Table 2 for details). That is, SVMs require more training instances than Random Forests to achieve proper discrimination capabilities. Consequently, SVM perform better when trained from Sequence 5, which has 8412 images in contrast to the rest of sequences (3510 images on average).
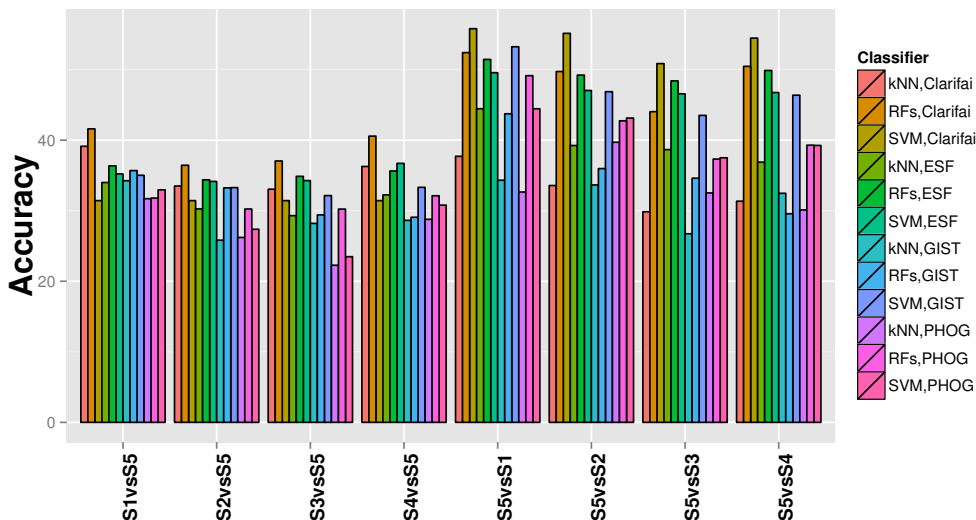

Figure 9.   Accuracy obtained for all the classifiers and descriptors combinations in those scenarios involving Sequence 5 as test (left) or training (right).

From these scenarios, we also carried out a Friedman test (0.05 confidence level) and a post-hoc statistical analysis. The null hypothesis that all descriptor/classifier combinations are equivalent was rejected. The post-hoc analysis was carried out against the best combination, Clarifai/RFs, and obtained the raking distribution shown in Fig. 10. This comparison stated that just ESF (in conjunction with RFs and SVM) and the combinations of GIST and Clarifai with SVMs are not significantly different from Clarifai/RFs. Fig. 10 illustrates the average rank position achieved with each combination of descriptor and classifier in the eight scenarios involving Sequence 5. It can be observed how the best combination always ranked between the first and fourth positions. This figure also helped us discover the low discriminating capabilities of the kNN classifier.

5.   Discussion

Clarifai descriptor have been shown as an appropriate representation for the scene classification problem, with no significant differences with respect to global features like GIST or ESF. However, two remarks can be obtained if we review the semantic labels generated from ViDRILO images using the Clarifai system. These labels are ranked by their distribution and graphically presented in Fig. 11. It is exposed how a small set of labels concentrates most of the presence
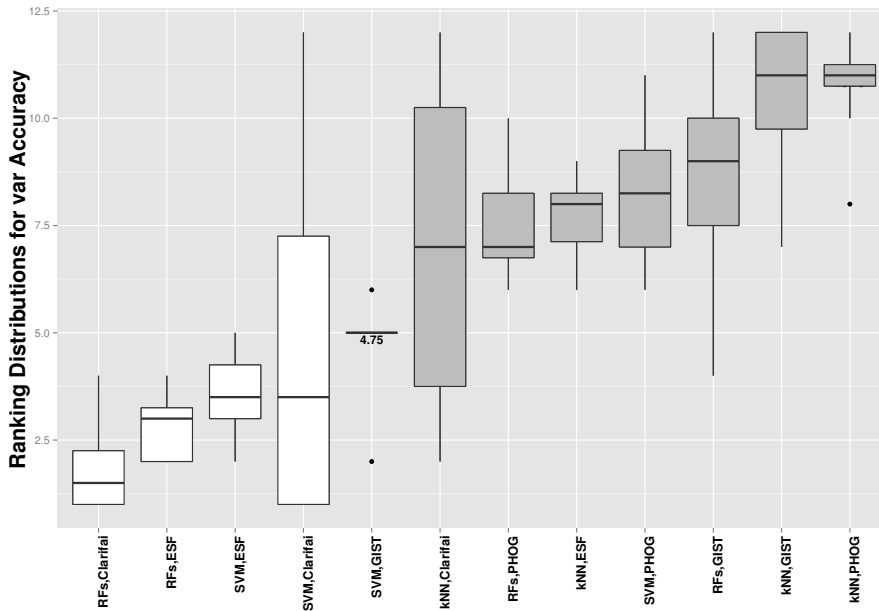
Figure 10. Average ranking for all classifiers and descriptors combinations.

in the images. The frequency distribution shows that 50% of the annotations correspond to just the 0.026% of the labels (21/793). **This fact makes these labels to play a very important role in the generation of the descriptor, in detriment of the rest of labels provided by the Clarifai annotation system.**
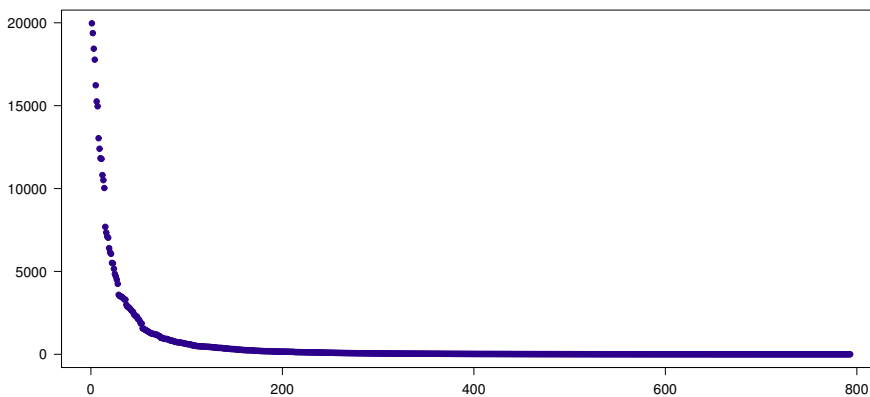

Figure 11. Frequency obtained by the labels in ViDRILO .

A second point to be taken into account is the discrimination capabilities of this set of labels. To do so, we firstly selected only the 10 most frequent labels. From these labels, and taking advantage of another Clarifai capability, we obtained some representative images of these concepts. Both the labels and the representative images are shown in Fig. 12, and we can observe how these labels are too general-ist. Concretely, these labels may be helpful for other computer vision tasks, such as to determine whether an image represents an indoor or outdoor environment. However, these labels are not discriminative when facing the scene classification

problem. That is, the presence of label "floor" does not help to resolve the type of scene where the image was acquired.

Indoors   Room   Window   House   Door

Dwelling   Nobody   Architecture   Contemporary   Floor

Figure 12. Exemplar visual images for the tags in the Clarifai Descriptor.

In summary, the Clarifai-based proposed descriptor provides a very high **balance** between simplicity and performance against well-known complex image descriptors in the context of scene classification problems.

Also, the lexical nature of this descriptor, namely a set of labels describing the scene, allows for its direct human understanding. This interpretation of the descriptor can be used to integrate expert knowledge in the scene classification pipeline, such as expert-driven feature/label selection techniques, or high-order linguistic feature combinations using NLP techniques among others.

6.   Conclusions and future work

We have proposed and evaluated the use of Clarifai labels as a valid descriptor for the scene classification problem. These descriptors are generated from the semantic annotations obtained thought an external API. The trial version of this Clarifai API obtains the most feasible 20 labels representing the input visual image. Thanks to this approach, researchers can focus on selecting the appropriate classification model as the image descriptors are automatically generated.

In view of the results obtained, we can conclude that Clarifai descriptors are as competitive as state-of-the-art ones, which are computed with computer vision (GIST), or 3D processing techniques (ESF). Moreover, Clarifai is exposed as the most outstanding descriptor when generalization capabilities are requested. This situation has been evaluated by training scene classifiers from sequences acquired in a building, and then test these classifiers in sequences acquired in a different building.

We have also reviewed the distribution, as well as the description, of the semantic labels obtained with Clarifai from ViDRILO images. In this review, it has been exposed that we are facing a specific problem (scene classification) from generalist information. That is, Clarifai annotations are not focused on describing scenes but, general concepts. Despite this fact, a competitive image descriptor has been proposed, developed and evaluated.

As future work, we plan to select a set of relevant semantic labels in a preliminary step, and then perform a problem-oriented labeling by asking Clarifai about the probabilities of this labels in the image. We also have in mind the use of classifiers capable of working with missing values.

Besides, we are using only the Clarifai responses, namely the final layer in the CNN architecture. This is made because Clarifai does not provide internal CNN values. We pretend to change our system to other open framework like Caffe with which we are able to use internal layers. Thus we plan to use the last two layers for classification purposes.

Acknowledgments.

References

[1] Maron O, Ratan AL. Multiple-instance learning for natural scene classification. In: Proceedings of the fifteenth international conference on machine learning. ICML'98. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.. 1998. p. 341–349.

[2] Martínez-Gómez J, Fernández-Caballero A, García-Varea I, Rodríguez L, Romero-González C. A taxonomy of vision systems for ground mobile robots. International Journal of Advanced Robotic Systems. 2014;11:1–11.

[3] Wu J, Christensen H, Rehg JM, et al.. Visual place categorization: Problem, dataset, and algorithm. In: IEEE/RSJ international conference on intelligent robots and systems. IEEE. 2009. p. 4763–4770.

[4] Csurka G, Dance CR, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. 2004. p. 1–22.

[5] Martínez-Gómez J, Morell V, Cazorla M, García-Varea I. Semantic localization in the PCL library. Robotics and Autonomous Systems. 2016;75, Part B:641 – 648.

[6] 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 - volume 2. CVPR '06. Washington, DC, USA: IEEE Computer Society. 2006. p. 2169–2178.

[7] Li L, Su H, Lim Y, Li F. Objects as attributes for scene classification. In: Trends and topics in computer vision - ECCV 2010 workshops. Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I. 2010. p. 57–69.

[8] Lampert CH, Nickisch H, Harmeling S. Attribute-based classification for zero-shot visual object categorization. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2014;36(3):453–465.

[9] Li L, Su H, Lim Y, Li F. Object bank: An object-level image representation for high-level visual recognition. International Journal of Computer Vision. 2014;107(1):20–39.

[10] LeCun Y, Kavukcuoglu K, Farabet C. Convolutional networks and applications in vision. In: International symposium on circuits and systems (ISCAS 2010). Paris, France. 2010 Jun. p. 253–256.

[11] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. Imagenet large scale visual recognition challenge. International Journal of Computer Vision (IJCV). 2015;115(3):211–252.

[12] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097–1105.

[13] Martinez-Gomez J, Garcia-Varea I, Cazorla M, Morell V. ViDRILO: The visual and depth robot indoor localization with objects information dataset. The International Journal of Robotics Research. 2015;34(14):1681–1687.

[14] Bosch A, Zisserman A, Munoz X. Representing shape with a spatial pyramid kernel. In: Proceedings of the 6th acm international conference on image and video retrieval. CIVR '07. Amsterdam, The Netherlands. New York, NY, USA: ACM. 2007. p. 401–408.

[15] Oliva A, Torralba A. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision. 2001;42(3):145–175.

[16] Wohlkinger W, Vincze M. Ensemble of shape functions for 3d object classification. In: 2011 IEEE international conference on robotics and biomimetics (ROBIO). IEEE. 2011. p. 2987–2992.

[17] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: Proceedings of the 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) - volume 1. CVPR '05. Washington, DC, USA: IEEE Computer Society. 2005. p. 886–893.

[18] Clarifai. Clarifai: Amplifying Intelligence. 2015. Available from: `http://www.clarifai.com/`.

[19] Martínez-Gómez J, Caputo B, Cazorla M, Christensen HI, Fornoni M, García-Varea I, Pronobis A. Where are we after five editions?: Robot vision challenge, a competition that evaluates solutions for the visual place classification problem. IEEE Robotics and Automation Magazine. 2015 12;2(4):147–156.

[20] Breiman L. Random forests. Machine Learning. 2001;45(1):5–32.

[21] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Transactions on Information Theory. 1967;13(1):21–27.

[22] Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011 May;2(3):27:1–27:27.

[23] Friedman M. A comparison of alternative tests of significance for the problem of m rankings. The Annals of Mathematical Statistics. 1940;11(1):86–92.

[24] Demšar J. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research. 2006;7:1–30.