Computing Image Descriptors from Annotations Acquired from External Tools

Jose Carlos Rangel, Miguel Cazorla, Ismael García-Varea, Jesús Martínez-Gómez, Elisa Fromont, Marc Sebban

Computing Image Descriptors from Annotations Acquired from External Tools

Jose Carlos Rangel[1], Miguel Cazorla[1], Ismael García-Varea[2], Jesús Martínez-Gómez[2], Élisa Fromont[3], and Marc Sebban[3]

[1] Computer Science Research Institute. University of Alicante
P.O. Box 99. E-03080. Alicante. Spain
[2] University of Castilla-La Mancha
Albacete. Spain
[3] Jean Monnet University
Saint Etienne.France

Abstract. Visual descriptors are widely used in several recognition and classification tasks in robotics. The main challenge for these tasks is to find a descriptor that could represent the image content without losing representative information of the image. Nowadays, there exists a wide range of visual descriptors computed with computer vision techniques and different pooling strategies. This paper proposes a novel way for building image descriptors using an external tool, namely: Clarifai. This is a remote web tool that allows to automatically describe an input image using semantic tags, and these tags are used to generate our descriptor. The descriptor generation procedure has been tested in the ViDRILO dataset, where it has been compared and merged with some well-known descriptors. Moreover, subset variable selection techniques have been evaluated. The experimental results show that our descriptor is competitive in classification tasks with the results obtained with other kind of descriptors.

Keywords: Descriptor generation, Computer Vision, Semantic Localization, Robotics

1 Introduction

Representing images in an appropriate way is essential for tasks like image reconstruction, image search or place recognition [17]. Comparisons between image descriptors can be used to determine the similarity between pairs of images [16]. Moreover, they can also be used for generalization capabilities. This is usually done by learning classification models, where the class corresponds to a desired image category. We can find binary categorization [14], and also multi-nominal proposals [9].

The main goal of an image descriptor is to find a proper representation minimizing the loss of information. Besides, some well-known approaches like histograms of gradients (HoG [4]) or Centrist [19], we can find some novel and

interesting approaches. Among these alternative representations, Fei et al. [10] propose the use of an Object Filter Bank (OFB) for scene recognition. OFB contains the responses produced for objects detectors formerly trained. The work presented in [15] includes the generation of a model that simultaneously classifies and obtains a list of annotations from images. Zhou et al. [20] suggest the application of the Super Vector (SV) coding, a non linear method to compute image descriptors. Lampert et al. [8] employ an attribute classification based object recognizer. The proposal relies on semantic attributes like shape or color of an object to perform a high-level description. Banerji et al. [1] construct a descriptor based on the color, shape and texture, through the fusion of two different descriptors using an feature representation technique.

Nowadays, it is very common the use of an external and/or remote tools that provides some functionalities or information. We can find traffic or meteorology information systems [13], but also some technologies offering processing capabilities, like grid and cloud computing [5]. These systems allow the access to a wide range of services and novel capabilities. In this sense, the Clarifai system [4] provides the technology to analyze images and identify descriptive annotations (i.e. tags) related to them. Clarifai offers an Application Programming Interface (API) that obtains the 20 most descriptive annotations from a submitted image.

This article proposes (and analyzes) the use of Clarifai to build an image descriptor based on the labels got by means of this system. To carry out the experimentation, we extract Clarifai descriptors from the visual images included in the ViDRILO dataset [11]. These descriptors are then evaluated and compared with two well-known visual and depth descriptors (GIST [12], and the Ensemble of Shape Functions (ESF [18])) in the scene classification problem, using Support Vector Machines(SVMs [2]) as classifier. Clarifai descriptors are also tested when combined with ESF and GIST. Furthermore, a subset variable selection algorithm is applied to the Clarifai descriptors.

The rest of the paper is organized as follows: Section 2 exposes how the Clarifai annotation system works. Section 3 describes our proposal to build the Clarifai descriptor. The computed descriptors used in this paper are explained in Section 4. Section 5 shows the experimentation and the obtained results. Finally, in Section 6 the conclusions obtained for this work and the future work are presented.

2 Remote Image Annotation: Clarifai technology

Clarifai technology relies on the use of Convolutional Neural Networks(CNN [6] to process an image, and then generates a list of tags describing the image. CNNs are defined as hierarchical machine learning models, which learn complex images representations from large volumes of annotated data. They use multiple layers of basic transformations that finally generate a highly sophisticated representation of the image [3]. The Clarifai approach was firstly proposed to the ImageNet classification challenge [7] in 2013, where the system produced the top 5 results.

[4] `http://www.clarifai.com/`

Clarifai works through the analysis of images to produce a list of descriptive tags representative of a given image. For each tag in this list, the system also provides a probability value. This probability represents the likelihood of describing the image using the specific tag. The Clarifai API can be accessed as a remote web service. The working scheme of the Clarifai technology is shown in Fig. 1. In this scheme, Clarifai uses a ViDRILO image as input and uses the CNNs to analyse it and produce the list of labels and probabilities.

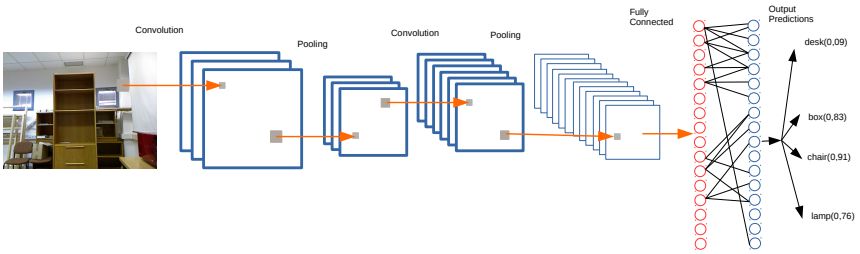


Fig. 1. Processing Scheme of the Clarifai API.

The Clarifai API can be accessed using a trial option or used under a payment mode. In this work, the trial mode has been used to obtain the tags of the images. This mode allows to get the tags for a restricted number of images (10000), and it limits the number of calls per hour to 1000. Using the trial mode, we can only obtain 20 tags per image with their associated probability. In Fig. 2, we show the image sent to the API and the returned labels/probabilities. From these labels, Fig. 3 shows some examples of representative images produced by the API. These images are generated using the annotations from ImageNet.

Regarding processing time, Clarifai has a latency time of almost 1 minute. Clearly it is not a real-time option, but new approaches for deep learning (using Caffe with Googlenet, for example) will provide a local solution (speeding-up with GPU architectures) which good time responses.

3 Image descriptors from Clarifai annotations

The process to generate Clarifai descriptors starts with an early step where we discover all the labels included in our problem domain. This is done by processing all the available images with the Clarifai API, and then removing the duplicated label values. This step can be seen as a codebook or dictionary generation process. Each input image is then encoded as a list of probability values whose length is determined by the size of the dictionary. The i-th value in this descriptor will correspond to the probability returned by Clarifai for the i-th label, or zero otherwise. This process is shown in Table 1.

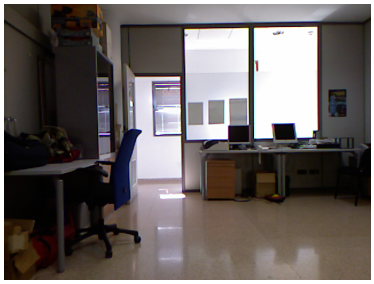
Image	Label	Probability
	Indoors	0.9936
	Seat	0.9892
	Contemporary	0.9787
	Chair	0.9779
	Furniture	0.9744
	Room	0.9634
	Interior design	0.9627
	Window	0.9505
	Table	0.9428
	Computer Technology	0.9417

Fig. 2. Labels and probabilities obtained with the Clarifai API for a ViDRILO image.

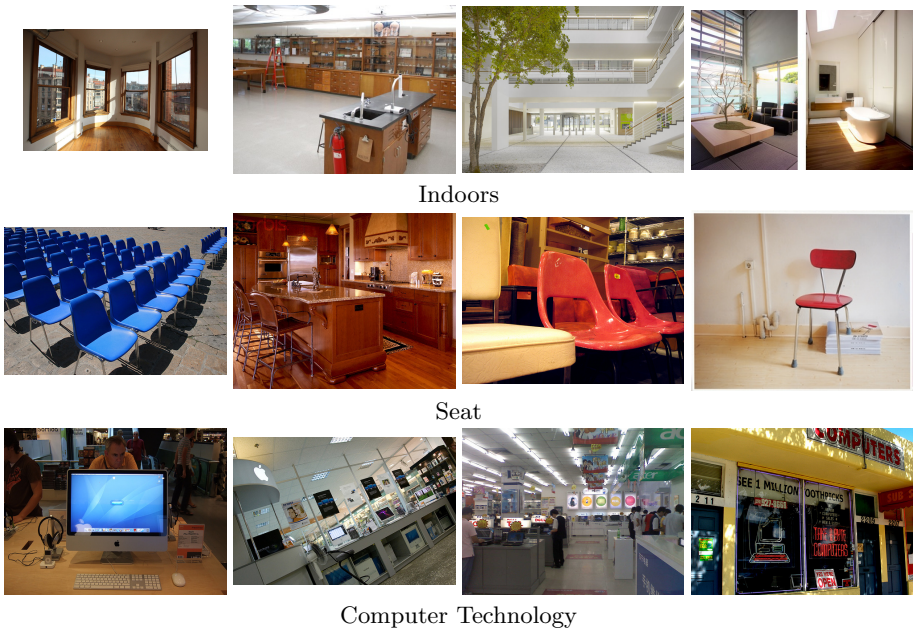


Fig. 3. Representative images for the labels extracted in Fig. 2.

4 Computing ViDRILO descriptors

To make a comparison among existing descriptors and Clarifai ones, the first step is to generate all these descriptors for the images from the chosen ViDRILO dataset. We select GIST [12], and the Ensemble of Shape Functions (ESF [18]) as they obtain the highest baseline results using visual and depth information respectively. Moreover, they are released in conjunction with the ViDRILO dataset

Table 1. Generation of Clarifai descriptors

	$label_1$	$label_2$	$label_3$	$label_4$	$label_5$	$label_6$	...	$label_N$
$Image_1$	0.99	0.98	0.97	0	0	0	...	0
$Image_2$	0	0.94	0	0.93	0.94	0	...	0
...	...	...	...	...	...	...	...	...
$Image_M$	0	0	0	0	0	0	...	0.91

and its processing toolbox[5]. We briefly describe these two descriptors and the proposed one.

4.1 GIST

The purpose of the GIST descriptor is to represent the shape of the image using a holistic representation of the space envelope. To generate the descriptor, an image is split into $N \times N$ patches. A low-dimensional vector is then generated from each of these patches. The size of these vectors depends on the number of orientations and scale variations initially selected. Using 16 patches ($N = 4$), 4 scales and 8 orientations (default values) we generate GIST descriptors whose dimensionality is 512.

4.2 Ensemble of Shape Functions(ESF)

This depth descriptor produces 10 different histograms as a result of three shape functions. It codifies the relation existing among the 3D points of a depth image. It uses 64 bins for histogram, which result in a descriptor with size 640.

4.3 Clarifai Descriptor

The Clarifai descriptor is generated as explained in Section 3. To compute the descriptor for the ViDRILO dataset, we firstly tagged all the images of the dataset. The next step was to count the unique produced labels, and we obtained 793 different values. This is the final dimensionality of the Clarifai descriptors. Hence, the Clarifai descriptors are built for every image placing the probability of the label under the respective label in the descriptor. Table 2 contains the 10 most frequent labels in the ViDRILO dataset and their appearance ratio by sequence. This table shows that the most common identified labels of the dataset are the most common in each sequence too with small variations. It is worthy of note the fact that most common labels neither show nor describe a specific scene or object name.

[5] `http://www.rovit.ua.es/dataset/vidrilo/`

Table 2. Most frequent labels in the ViDRILO sequences and their appearance ratio by sequence.

Labels	Sequence 1	Sequence 2	Sequence 3	Sequence 4	Sequence 5
Indoors	91%	94%	81%	79%	94%
Room	90%	89%	80%	79%	90%
Window	85%	86%	91%	74%	82%
House	83%	80%	88%	73%	79%
Door	76%	77%	74%	68%	70%
Dwelling	69%	71%	60%	64%	70%
Nobody	62%	75%	67%	83%	54%
Architecture	62%	65%	63%	63%	49%
Contemporary	58%	61%	40%	44%	61%
Floor	59%	50%	38%	42%	62%

5 Experiments and Results

All the experiments included in this paper have been performed using ViDRILO [11] as benchmark. The main characteristics of this dataset, as well as the distribution of its sequences, are shown in Table 3. The RGB-D images from this dataset have been acquired with a mobile robot in an indoor office environment.

Table 3. Overall ViDRILO characteristics and sequences distribution.

Sequence	Number of Frames	Floors imaged	Dark Rooms
Sequence 1	2389	1st,2nd	0/18
Sequence 2	4579	1st,2nd	0/18
Sequence 3	2248	2nd	4/13
Sequence 4	4826	1st,2nd	6/18
Sequence 5	8412	1st,2nd	0/20

In ViDRILO , each image is annotated with the category of the scene it was acquired, from a set of 10 room categories. Fig. 4 shows some representative images for the categories.

5.1 Experimentation

In order to carry out the evaluation of the proposed descriptor, the experiments consist of training a classifier using a sequence of the ViDRILO dataset and then use another sequence to test the trained model. From the 5 available sequences, we trained five different classifiers and all of them were evaluated tested against five different sequences. This resulted into 25 possible scenarios where each descriptor (GIST, ESF and Clarifai) was evaluated. To determinate the effectiveness of the Clarifai descriptor, we measure the accuracy of the results

Fig. 4. Exemplar visual images for the 10 room categories in ViDRILO .

produced by the classifier. This is defined as the percentage of well classified images in the test sequence of the dataset. As the experimentation produces 75 accuracy values (3 descriptor and 25 scenarios), we averaged the results by training and test sequence to better visualize these results. All the experiments were performed using a χ^2 SVM classifier.

5.2 Baseline results

The baseline results are obtained by comparing the accuracy values obtained with Clarifai, ESF and GIST descriptors. The Table 4 contains the average results obtained by the descriptors grouped by the train sequence. Fig. 5 shows the results obtained by descriptor and averaged over the sequence used to train (left) and test (right) the classifier. Here we see that our proposal obtains results similar to those obtained with the rest of descriptors. Moreover, Clarifai descriptor outperforms ESF and GIST when using Sequence 5 for training. This indicates a capability of generalization, as Sequence 5 was taken in another building where structure and color was different.

Table 4. Accuracy values obtained with the 3 evaluated descriptors and averaged by training sequence.

Sequence	GIST	ESF	Clarifai Tags
Sequence 1	65,38	56,59	58,32
Sequence 2	66,80	60,32	56,10
Sequence 3	60,55	60,34	59,43
Sequence 4	66,01	62,24	58,69
Sequence 5	57,94	54,82	63,17

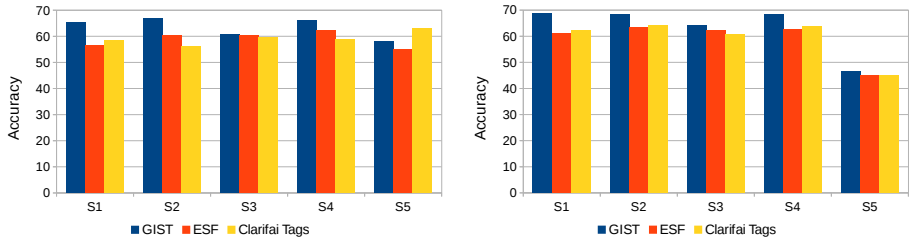


Fig. 5. Clarifai Descriptor averaged by Train (left) and Test Sequence(right).

5.3 Descriptor combination

Here we present the results training the classifier using a combination of 2 descriptors. The Table 5 contains the average results obtained using a merge of descriptors grouped by the train sequence. Fig. 6 shows the results obtained by the merge of descriptors and averaged over the sequence used to train (left) and test (right) the classifier. Fig.7 shows how the accuracy of the Clarifai descriptors improves by using a merge of descriptors. These results are grouped by the training sequence.

Table 5. Merge of Descriptors averaged by Train Sequence

Sequence	ESF+GIST	ESF+Clarifai	GIST+Clarifai
Sequence 1	65,77	58,38	58,36
Sequence 2	66,82	58,60	58,51
Sequence 3	61,76	58,98	59,34
Sequence 4	66,38	58,67	58,65
Sequence 5	53,04	63,18	63,21

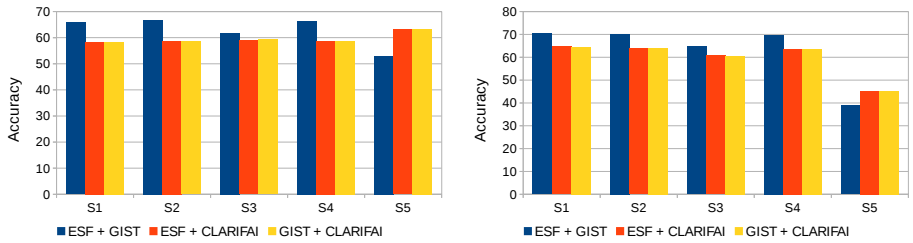


Fig. 6. Results with merge of descriptors averaged by Train Sequence (left) and by Test Sequence (right).

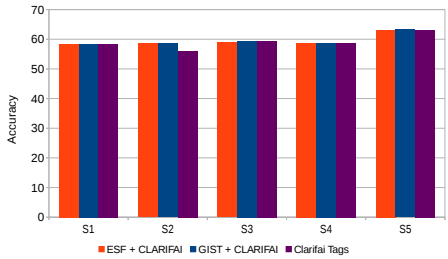


Fig. 7. Comparison of baseline Clarifai against merge of descriptors.

5.4 Label Subset Selection

The experiments for this section included an additional step before the classification. This step consists in the application of a subset variable selection process in the Clarifai descriptors to reduce the number of attributes used to describe the image. The descriptor still gets good results with this process, obtaining even better results using a reduced amount of labels to train the classifiers. Fig. 8 shows the result of selecting 10, 20, 50, 75, 100, 150, 200 variables and then training the SVM classifier. It shows that the use of a few labels reduces the accuracy results, but when the number increase the accuracy outperforms the one obtained using the 793 labels of the descriptor. Hence the use of 75 labels gets higher accuracy than the use of all the labels.

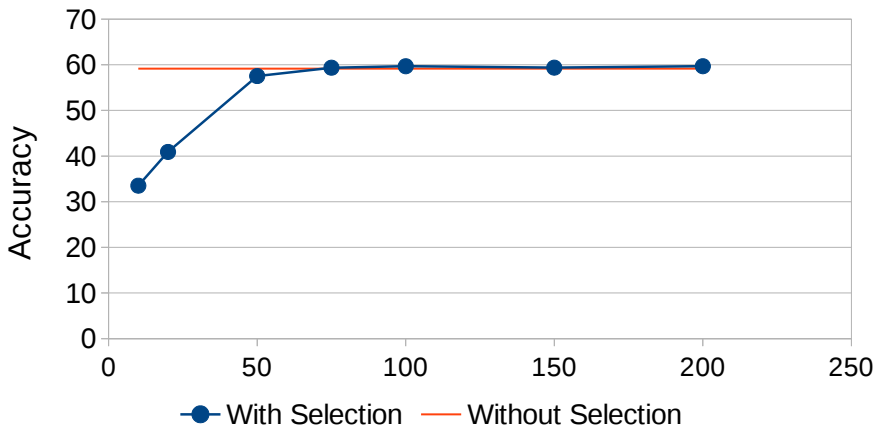


Fig. 8. Accuracy of the Clarifai Descriptors using a Subset Variable Selection Process

6 Conclusions and future work

We have presented a new way to build an image descriptor using Clarifai, an external image labeling tool. We have compared this descriptor with other two descriptors, achieving competitive results in the different scenarios that we have tested. The Clarifai descriptor shows that even with a reduced dimensionality, outperforms its own results in classifications tasks. As future work, we want to test this system using all the possible tags for a given image.

Acknowledgments.

This work was supported by the Ministerio de Economia y Competitividad of the Spanish Government, supported with Feder funds under grant DPI2013-40534-R; Consejería de Educación, Cultura y Deportes of the JCCM regional government under project PPII-2014-015-P. Jesus Martínez-Gómez is also funded by the JCCM grant POST2014/8171.

References

1. Banerji, S., Sinha, A., Liu, C.: Novel color, shape and texture-based scene image descriptors. In: Intelligent Computer Communication and Processing (ICCP), 2012 IEEE International Conference on. pp. 245–248 (Aug 2012)
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011)
3. Clarifai: Clarifai: Amplifying Intelligence (2015), `http://www.clarifai.com/`
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Int. Conf. on CVPR. vol. 1, pp. 886–893. IEEE (2005)
5. Foster, I., Zhao, Y., Raicu, I., Lu, S.: Cloud computing and grid computing 360-degree compared. In: Grid Computing Environments Workshop, 2008. GCE'08. pp. 1–10. Ieee (2008)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
8. Lampert, C., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36(3), 453–465 (March 2014)
9. Li, L.J., Fei-Fei, L.: What, where and who? classifying events by scene and object recognition. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. pp. 1–8. IEEE (2007)
10. Li, L.J., Su, H., Lim, Y., Fei-Fei, L.: Objects as attributes for scene classification. In: Kutulakos, K. (ed.) Trends and Topics in Computer Vision, Lecture Notes in Computer Science, vol. 6553, pp. 57–69. Springer Berlin Heidelberg (2012)

11. Martinez-Gomez, J., Cazorla, M., Garcia-Varea, I., Morell, V.: Vidrilo: The visual and depth robot indoor localization with objects information dataset. International Journal of Robotics Research (2015)
12. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42(3), 145–175 (2001)
13. Petty, K.F., Moylan, A.J., Kwon, J., Mewes, J.J.: Traffic state estimation with integration of traffic, weather, incident, pavement condition, and roadway operations data (Feb 5 2014), uS Patent App. 14/173,611
14. Szummer, M., Picard, R.W.: Indoor-outdoor image classification. In: Content-Based Access of Image and Video Database, 1998. Proceedings., 1998 IEEE International Workshop on. pp. 42–51. IEEE (1998)
15. Wang, C., Blei, D., Li, F.F.: Simultaneous image classification and annotation. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. pp. 1903–1910 (June 2009)
16. Wang, G., Hoiem, D., Forsyth, D.: Learning image similarity from flickr groups using fast kernel machines. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(11), 2177–2188 (2012)
17. Winder, S., Brown, M.: Learning local image descriptors. In: Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on. pp. 1–8 (June 2007)
18. Wohlkinger, W., Vincze, M.: Ensemble of shape functions for 3d object classification. In: Robotics and Biomimetics (ROBIO), 2011 IEEE International Conference on. pp. 2987–2992. IEEE (2011)
19. Wu, J., Rehg, J.M.: Centrist: A visual descriptor for scene categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33(8), 1489–1501 (2011)
20. Zhou, X., Yu, K., Zhang, T., Huang, T.: Image classification using super-vector coding of local image descriptors. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) Computer Vision ECCV 2010, Lecture Notes in Computer Science, vol. 6315, pp. 141–154. Springer Berlin Heidelberg (2010)