

Object Recognition Approach based on Color Distribution

Nada Farhani^{1,2}, Naim Terbeh², Mounir Zrigui²

¹ University of Sousse, Hammam Sousse,
Tunisia

² University of Monastir, Monastir,
Tunisia

farhaninada@yahoo.fr, naim.terbeh@gmail.com,
mounir.zrigui@fsm.rnu.tn

Abstract. Until now, object recognition presents a difficult task and still requires research to improve the recognition results. The observation of an object can be very different depending on the field of activity as well as the reason to recognize and classify, perhaps simple or complex, and the tools developed are adapted to each use. It is in this context that this paper intervenes. We propose a recognition approach based on the color distribution using the histogram and spatiogram calculation, in the RGB space. With the used database, our approach gave good results. The main purpose of our work is to use the concepts from recognition to generate sentences in Arabic that summarize the content of the image.

Keywords: Object recognition, color distribution, histogram, spatiogram, learning.

1 Introduction

Recognition of objects is a delicate problem that takes place at the top level in the hierarchy of vision tasks and constitutes the most difficult computational part. To overcome this difficulty, a vision system must be able to combine its internal representational capacities in order to make successful decisions. Many difficulties appear in the recognition of objects, in particular those linked to the variability of appearance linked to light, orientation, etc. Most theories of object recognition only deal with the geometric aspect. Today we can count two large groups of theories, which diverge, on the format of the representation according to whether this one is independent or dependent on the views of the object to be represented.

The first group of these theories considers that the representation of an object is conceived as a set of characteristics (invariants) of the object which are independent of

the views of this same object [1]. This is a structural description of the object. One of the most interesting approaches is that of Biederman: "Recognition by Components" [2] which consists in representing the object by breaking it down into structures (primitives) according to a scheme proposed by de Marr and Nishihara [3].

The second group of these theories considers that the representation of an object is linked to views specific to the object and that any other view can be deduced using these views [4,5]. The models of this type of representation consider a view as a collection of characteristics (2D information, 3D information...). Recognition is expressed as a function of the images already seen. The model claimed by the first group can appear attractive thanks to a compact and robust representation for the objects.

However, experience has shown that it is very difficult to detect invariants in images on the one hand and that this structure of invariants necessarily leads to categorization and not to identification on the other. To remedy these drawbacks, new approaches have surfaced according to the theory of the second group. These approaches model objects by their images themselves, thus abandoning models of geometric type or based on invariant structures. Thus, an object is represented by a collection of images and the recognition is based on the matching of a new image of the object with the images in the collection.

Several works [4] have evolved in this direction. This paper is organized as follows. In the second section, an overview of the image description methods that help in object recognition tasks. In the third section, a state of art is presented. The fourth section presents the proposed approach based on the histogram and spatiogram calculation. Some experimental results as well as the discussion will be presented in section five. Finally, section six concludes this paper.

2 Image Description Methods

After doing image preprocessing, we are interested in the task of object recognition. In fact, it is necessary to extract areas of interest from information relating to what is contained in the image; for this, there are image descriptors that characterize the information available. Two ways of proceeding are possible:

- Either the image is described at specific and significant points: it is a local descriptor, and these interesting points are points of interest.
- Or all the pixels of the image corresponding to the area of interest are considered in the description: it is a global descriptor.

2.1 Local Descriptors

For local descriptors present methods that are based on correlation measures. They make it possible to quantify the resemblance between two pixels and their neighborhoods. There are methods to detect points of interest. For example, the Harris point detector [6], which is certainly the most widely used; its principle is to detect

sudden changes in intensity in the image, thereby highlighting corners. Then a descriptor characterizes each point of interest by its neighborhood. Most often, this is the information the detector has calculated. Local methods are the most used today for the object recognition thanks to their good management of occultation, charged background and changes of point of view as well as their speed.

However, unfortunately, local methods do not take the entire image into account and are fraught with ambiguities. This results in many matching errors.

2.2 Global Descriptors

In object recognition, a global descriptor is easier to use because it processes the image in its entirety. The descriptor is thus less sensitive to distortions from one image to another. Two close images must therefore lead to two close descriptors. A classifier can thus be trained on this data during training, and it will be able to associate these two images with the same class thanks to a notion of distance between the descriptors (for example Euclidean distance).

The same will be true for recognition: by using the same descriptor as during training, the classifier will be able to associate a new object with a class of learned objects. Many methods exist. The choice depends above all on the intended application and the calculation time available to carry it out. Indeed, in the case of real-time recognition and detection, there is often a trade-off between the quality of the response and the execution time.

However, if the information provided by the descriptor is not precise enough, the classifier will be difficult to train during learning and the results it will provide in recognition will therefore be unreliable. Global methods take the entire image into account. They are based on the following principle: if the calculated disparity map is correct and if an image is constructed from the reference image and the disparity map, then the resulting image should resemble the other image. We then seek to find the disparity map which maximizes a global similarity function.

An example of a method is the method presented in [7]. This method is one of the best classified in the protocol of Scharstein and Szeliski of [8]. It is divided into four stages:

- Color segmentation.
- Use of an adaptive correlation score that maximizes the number of reliable matches.
- Assigning a disparity value to each region.
- Search for the optimal disparity using a belief propagation based on the Markov field model.

3 State of the Art: Objects Recognition

The authors, in [9], presented a residual learning framework in order to simplify the formation of deeper networks than previously used networks. In an explicit way, Zhang

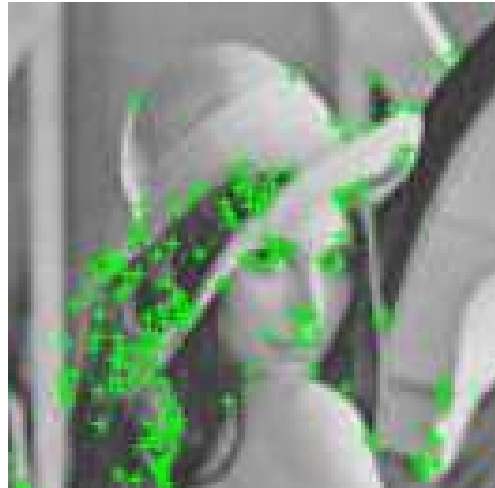


Fig. 1. Harris Detector - In order to find points of interest, the Harris Detector calculates, for each pixel, the autocorrelation matrix from the two components of the image's gradient vectors. Then, the detector response matrix is obtained from these matrices. Finally, the points of interest, here marked with a green cross, are located from this answer.

et al. did not choose to learn unreferenced functions, but they reformulated the layers as residual learning functions with reference to the layer inputs. In their work [10], Chabot & al. proposed a system that is divided into two stages. In fact, this system is used to transmit the input image via the Deep MANTA network which produces the visibility properties of the parts, the 2D bounding boxes and the associated vehicle geometry.

OverFeat [11] is an example of the first modern objects detectors, which is a one-stage detector based on deep networks. One of the newest detectors is SSD (Single Shot MultiBox Detector) [12], where its approach is based on a convolutional feed-forward network generating scores for the presence of object class instances. In fixed size bounding boxes that were generated before, then a non-maximal deletion step to produce the final detections. YOLO9000 [13] also a real-time framework is designed to detect more than 9000 categories of objects while optimizing the classification and detection. In order to detect faces, in their works [14], the authors exploited detectors of boosted objects.

The use of integral channels [15] and HOG functions [16] has led to the emergence of effective methods for pedestrian detection. In the classical computer vision, the sliding window approach was the main model of detection, with the appearance of deep learning [17]. With regard to work based on two-stage detectors, in [18], the first phase is the generation of an isolated set of candidate propositions in the obligation to include all the objects as well as the filtering of the majority negative locations. The second phase classifies proposals into foreground / background.

For the purpose of improving the second-stage classifier, R-CNN [19] has made modifications to this level to have a convolutional network that offers significant gains

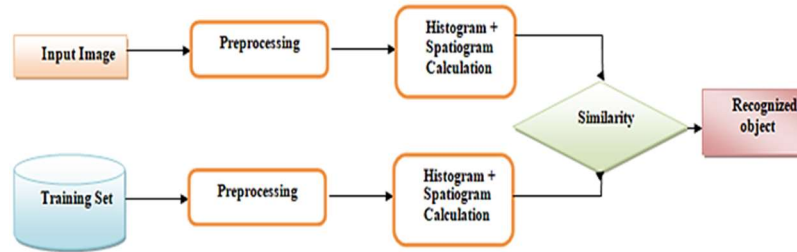


Fig. 2. Recognition steps. This figure shows the details of the steps to follow in order to recognize object (s) in an image.

in accuracy. In turn, R-CNN has also been improved over the years, at the same time at the speed level [20] and by exploiting proposals for learned objects [21]. To improve timeliness and to have a faster classifier than that used by the RCNN, Region Proposal Networks (RPPs) incorporated proposal generation with the second-stage classifier into a single convolutional network [21]. As well as several extensions have been proposed in this framework [22, 23, 24]. In [25], Karpathy & al. proposed a model generating natural language descriptions of images and their regions.

To learn more about the correspondence between images and language, the approach proposed the authors exploits the dataset of text descriptions of images. The base of the alignment model is a new combination of two-way recursive neural networks on sentences, Convolutional neural networks on image regions, and a structured objective that serves to align the two modalities via multimodal integration.

The approach proposed in [26] by Lin & al. is the focal loss applying a modulating term to the loss of cross entropy in order to weight the many easy negatives and to focus learning on concrete examples. Vaillant & al. applied in their work [27] convolutional neural networks to the recognition of handwritten figures. Many researchers work on the detection in different fields [28, 29, 30, 31, 32], but the objects detection and recognition still a challenge in the field of research because of several difficulties that the researcher can envisage because of the variability of shape, position, contrast of objects.

4 The Proposed Approach

Since we want to build an image description system that can be used by humans, and it is the human being who gives meaning to what he sees, it may be interesting to draw inspiration from of the human perception system to choose the visual spaces in order to come as close as possible to the human being's understanding of the image, and thus reduce the semantic gap. Low-level descriptors are used at the level of global methods, also called feature vectors, such as color, texture, and shape.

In our approach we will use the color distribution in order to make the objects recognition task which is based on the histogram and spatiogram calculation. After

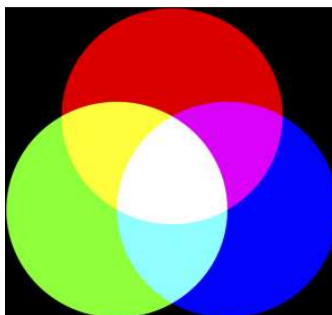


Fig. 3. Additive synthesis of colors.

building our database, we will do the preprocessing step for the images, after this the training step is done.

For the test phase, also we do a pretreatment for the image, calculate the histogram and spatiogram. And for the recognition task we will opt for a similarity calculation to find the recognized object. The following figure presents the system of objects recognition.

4.1 Color Image

A color image is actually made up of three images, in order to represent red, green, and blue. Each of these three images is called a channel. This representation in red, green and blue reflects the functioning of the human visual system.

Each pixel in the color image thus contains three numbers (r, g, b) , each being an integer between 0 and 255. If the pixel is equal to $(r, g, b) = (255, 0, 0)$, it contains only red information, and is displayed as red. Similarly, pixels of $(0, 255, 0)$ and $(0, 0, 255)$ are displayed green and blue, respectively. A color image can be displayed on the screen from its three channels (r, g, b) using the rules of additive color synthesis. The following figure shows the composition rules for this additive synthesis of colors. A pixel with the values $(r, v, b) = (255, 0, 255)$ is a mixture of red and green, so it is displayed as yellow. Figure 4 shows the decomposition of a color image into its three constituent channels.

4.1.1 Color Image Histogram

For a monochrome image, that is to say with a single component, the histogram is defined as a discrete function which associates with each intensity value the number of pixels taking this value. The histogram is therefore determined by counting the number of pixels for each intensity of the image. Sometimes a quantization is carried out, which groups together several intensity values in a single class, which can make it possible to better visualize the distribution of the intensities of the image. Histograms are usually normalized by dividing the values of each class by the total number of pixels in the image. The value of a class then varies between 0 and 1 and can be interpreted as the probability of occurrence of the class in the image. The histogram can then be seen as

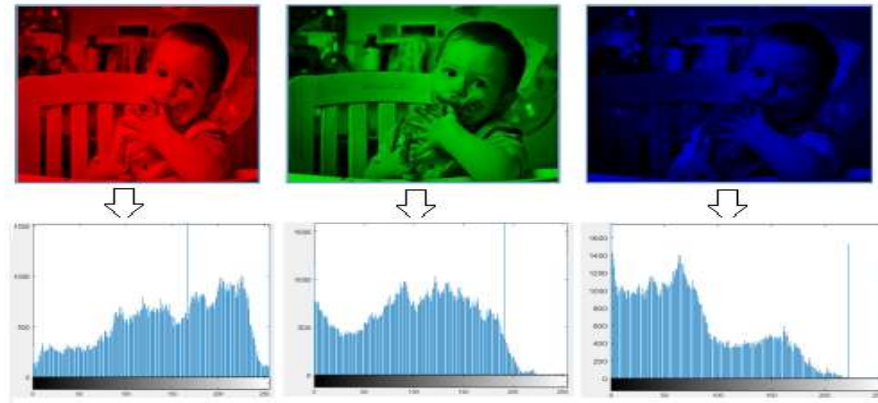


Fig. 4. Decomposition of a color image into its three channels and their corresponding Histograms.

a probability density. For color images, we can consider the histograms of the 3 components independently, but this is generally not efficient.

Rather, we construct a histogram directly in the color space. Histogram classes now correspond to a color, rather than intensity. This is called a color histogram.

4.1.2 Color Image Spatiogram

The lack of spatial information, in standard color histograms, has led us to the use of Spatiogram(s). Spatiogram(s) are generalized histograms describing more than the occurrence of the pixels in each color box, the average, and the covariance of the coordinates of the pixels, which makes it possible to capture the spatial distributions of the different image colors.








5 Tests and Results

Some experimental results are presented to evaluate our approach based on Histogram(s) and Spatiogram(s) to recognize objects in an image.

5.1 Image Database

For the database we used images from ImageNet and Pascal VOC and also, we have collected images from the Internet. All these images are divided according to their categories into test images and others into images forming the learning base. We tried to collect a large number of images according to their category so that each one can cover the maximum of images in different positions and different contrasts. We have used in our approach to do the recognition part the Support vector machines (SVM) which are a set of supervised learning techniques designed to solve discrimination and regression problems. SVMs are a generalization of linear classifiers.

Table 1. Comparison between the test image and some images from database with circular form.

Image	Histogram Similarity	Spatio Similarity
 تفاحة	0.94	0.92
 تفاحة	0.97	0.93
 تفاحة	0.22	0.15
 كرة	0.05	0.03
 كرة	0.02	0.01
 كرة	0.03	0.02
 كرة	0.06	0.05

5.2 Experimental Results

From the image database collected. We will test our algorithms on some categories of objects. The objects in the learning database are tagged to use the tags after. Subsequently, we will associate the object to recognize its histogram and its spatiogram to compare them later to those of the objects of the learning base by calculating a similarity factor. The table below shows the comparison results with different objects. From the results, we find that the values close to 1 are those associated with the images which are the most similar to the test image, despite having several objects of the same form which is circular.

5.3 Discussion

This approach has a part that serves to preserve the color information and that is consists on the use of histograms and spatiograms to help the recognition and this by a comparison between the histograms and the spatiograms with those of the unknown object. The results obtained show that the comparison between the approaches gave a precision rate of 96%. Compared with another works, [25] where the precision rate was

93% and also with [26], where the precision rate for histograms was 95% and spatiograms was also 95%, we have obtained good results that allow us to use it to generate Arabic sentences from the recognized objects.

6 Conclusion

We have presented an approach for the object recognition which is based on the color distribution using histograms and spatiograms for images in the RGB space. We can conclude that we have important results, but we will work in the future to ameliorate the results by introducing other features. The main goal of our works is to generate arabic description for an image using the recognized objects in the image.

References

1. Wallis, G., Rous, E.: Invariant face and object recognition in the visual system. *Progress in Neurobiology*, vol. 51, pp. 167–194 (1997)
2. Biederman, I.: Recognition-by-components: A theory of human image understanding. *Psychological Review*, vol. 94, 115–147 (1987)
3. Marr, D., Nishihara, H. K.: Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society B: Biological Sciences*. vol. 200, 269–294 (1978)
4. Edelman, S.: *Representation and Recognition in Vision*. MIT Press, Cambridge, MA. (1999)
5. Tait, M., Williams, P., Hayward, W.: Three-dimensional object recognition is viewpoint-dependent. *Nature Neuroscience*, vol. 1, pp. 275–277 (1998)
6. Harris, C., Stephens, M.: A combined corner and edge detector. In: *Alvey Vision Conference, Manchester - United Kingdom*, pp. 147–151 (1988)
7. Klaus, A., Sormann, M., Karner, K.: Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. In: *Dans International Conference on Pattern Recognition, Graz, Autriche, août*, vol. 3, pp. 15–18, (2006)
8. Scharstein, D., Szeliski, R.: A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms. *International Journal of Computer Vision*, vol. 47, no. 1, pp. 7–42 (2002)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778 (2016)
10. Chabot, F., Chaouch, M., Rabarisoa, J., Teulière, C., Chateau, T.: Deep MANTA: A Coarse-to-fine Many-Task Network for joint 2D and 3D vehicle analysis from monocular image (2017)
11. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun Y.: Integrated recognition, localization and detection using convolutional networks. In: *International Conference on Learning Representations* (2014)
12. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S.: Single shot multibox detector. In: *European Conference on Computer Vision* (2016)

13. Redmon, J., Farhadi, A.: YOLO9000: Better, faster, stronger. In: Computer Vision and Pattern Recognition, pp. 7263–7271 (2017)
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Computer Vision and Pattern Recognition (2001)
15. Dollar, P., Tu, Z., Perona, P., Belongie, S.: Integral channel features (2009)
16. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition (2005)
17. Krizhevsky, A., Sutskever, I., Hinton, G.: ImageNet classification with deep convolutional neural networks. In: Neural Information Processing Systems (2012)
18. Uijlings, J. R., van de Sande, K. E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. *International Journal of Computer Vision* (2013)
19. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask R- CNN. In: International Conference on Computer Vision, pp. 2961–2969 (2017)
20. He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. In: European Conference on Computer Vision (2014)
21. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Neural Information Processing Systems (2015)
22. Lin, T.Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Computer Vision and Pattern Recognition vol. 37, no. 9, pp. 1904–1916 (2017)
23. Shrivastava, A., Gupta, A., Girshick, R.: Training region-based object detectors with online hard example mining. In: Computer Vision and Pattern Recognition (2016)
24. Shrivastava, A., Sukthankar, R., Malik, J., Gupta, A.: Beyond skip connections: Top-down modulation for object detection (2016)
25. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
26. Lin, T. Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection (2017)
27. Vaillant, R., Monroca, C., Le-Cun, Y.: Original approach for the localisation of objects in images. *IEEE Proceedings on Vision, Image, and Signal Processing*, vol. 141, no. 4, pp. 245–250 (1994)
28. Farhani, N., Naim, T., Zrigui, M.: Image to text conversion: state of the art and extended work. 2017 In: IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA), pp. 937–943, (2017)
29. Terbeh, N., Zrigui, M.: Vocal pathologies detection and mispronounced phonemes identification: Case of Arabic continuous speech. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pp. 2108–2113 (2016)
30. Ayadi, R., Maraoui, M., Zrigui, M.: A Survey of Arabic Text Representation and Classification Methods. *Research in Computing Science*. vol. 117, pp. 51–62 (2016)
31. Terbeh, N., Achraf, M., Ben, M., Zrigui, M.: Probabilistic approach to Arabic speech correction for peoples with language disabilities. *International Journal of Information Retrieval Research (IJIRR)* vol. 5, no. 4, pp. 1–18 (2015)

32. Mansouri, S., Charhad, M., Zrigui, M.: Arabic text detection in news video based on line segment detector. *Research in Computing Science*. vol. 132, pp. 97–106 (2017)