A Pyramid of Concentric Circular Regions to Improve Rotation Invariance in Bag-of-Words Approach for Object Categorization

Arnaldo Câmara Lara and Roberto Hirata Jr.
Institute of Mathematics and Statistics, University of Sao Paulo, Sao Paulo, Brazil

Keywords: Spatial Pyramids, Rotation-Invariant, Bag-of-Words, Object Categorization.

Abstract: The bag-of-words (BoW) approach has shown to be effective in image categorization. Spatial pyramids in conjunction to the original BoW approach improve overall performance in the categorization process. This work proposes a new way of partitioning an image in concentric circular regions and calculating histograms of codewords for each circular region. The histogram of the entire image is concatenated forming the image descriptor. This slight and simple modification preserves the performance of the original spatial information and adds robustness to image rotation. The pyramid of concentric circular regions showed to be almost 78% more robust to rotation of images in our tests compared to the traditional rectangular spatial pyramids.

1 INTRODUCTION

Object recognition is a very active research area in Computer Vision (CV). Several approaches are used to deal with this problem. One of the most used and successful ones is the bag-of-words (BoW) approach. It consists of clustering the local features in a predetermined number of centers, the *codewords* or *visual words*. The collection of visual words forms the visual dictionary. The final descriptor of an image is a frequency histogram of these visual words (Csurka et al., 2004). Spatial pyramid is a technique that adds some spatial information in the BoW approach and it improves the final results (Lazebnik et al., 2006).

A big drawback in the use of spatial pyramids in the BoW approach is that it is not rotation-invariant. Rotation invariance is not important for a lot of objects like mountains, cars, houses, buildings, etc. These objects are usually presented in a scene without variations in rotation. However cans, bottles, shoes, pens or books are examples of objects that can be presented in a scene in different angles of rotation. For a human being, it is easy to identify a bottle or a can independent of its rotation in the scene. However, algorithms for object categorization do not treat these differences in rotation appearance without a cost.

In this work, we propose a pyramid of concentric circular regions in conjunction with a visual dictionary to improve the rotation invariance of object categorization algorithms. The piramid is formed by a collection of concentric circles where the frequency histograms of visual words are computed and the final descriptor is the concatenation of the frequency histogram of each circular region and the histogram of the entire image.

This text is organized in the following way. In this section, we introduced the problem and the motivation of our proposed modification in the spatial pyramids. The next section describes the main approaches used to deal with the object recognition problem. The main contribution of this work, the pyramid of concentric circular regions, is described in Section 3. Section 4 gives details about the implementation of the techniques described in this work: descriptor, classifier and database used. Section 5 describes the experiments and shows the results. The results of experiments are analysed in Section 6. Finally, Section 7 concludes the work and gives some directions for future works.

2 OBJECT RECOGNITION

From all possible tasks of CV area, object recognition is among the most challenging ones and it is also a very active research area (Szeliski, 2011). Even the state-of-art algorithms are unable to recognize all objects in all type of scenes in all possible conditions. Objects can be presented in different viewpoints, scales or sizes. They can be in a different positions, rotated or ocluded in the scenes. Changes in

Câmara Lara A. and Hirata Jr. R..
A Pyramid of Concentric Circular Regions to Improve Rotation Invariance in Bag-of-Words Approach for Object Categorization.
DOI: 10.5220/0004298806870692
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2013), pages 687-692
ISBN: 978-989-8565-47-1

the lightning conditions can vary the appearance of objects. Therefore, many different situations can produce a big variability in the aspects of objects.

The object recognition problem can be divided into some smaller sub-problems. If one wants to classify an object already detected in a scene as belonging to a class of a set of possible classes, the problem is named *categorization*. Categorization is an important task. When we know the category of an object in a scene, it is known what can be done with this object. That is very important to the semantic understanding of a scene.

Categorization of objects is an extremely difficult task and, up to date, the algorithms did not reach a 2-years child level of accuracy in the results of object categorization (Szeliski, 2011). Categorization algorithms only have good performance classifying an object as one of a set of known classes in a closed world collection. An adult person can recognize tens of thousands of different object categories. As the number of known classes increase, the accuracy of classification process decreases. For 10000 image categories, the best accuracy is less than 5% (Deng et al., 2010).

The first successful approach used to handle this task was the BoW. It appeared in text classification area and it was adapted to be used to classify objects detected in a scene (Cula and Dana, 2001). It is robust to variability of appearance of objects in the scene, to cluttered background and to changes in the lightning conditions. It showed computational efficiency and it has become widely used since then (Csurka et al., 2004). The main steps of BoW approach are:

- *Extraction of local features*. Local features are known to be very efficient in the task of object detection and categorization. They are computed using a spatial area near the pixel of interest.
- *Construction of the visual dictionary*. The set of local features are clustered in some number of centers, generally using *k-means* (Duda et al., 2001). The center of each cluster is a representative feature of all features in its neighbourhood. The collection of all centers of clusters forms the visual dictionary.
- *Representation*. Local features extracted from training and testing image sets are translated to visual words. Each local feature is represented by its nearest visual word. The final descriptor is the frequency histogram of the visual words from each image.
- *Learning*. A classifier is trained using labeled descriptors to build a model for each class of the image collection.

Spatial Pyramids is an extension of the BoW approach that adds spatial information in the orderless features of the BoW significantly improving the overall performance (Lazebnik et al., 2006).

Spatial pyramids consist of partitioning an image in increasingly smaller sub-regions and computing BoW technique in each sub-region. Spatial pyramids have a sequence of L levels. Each level $l \in 0,\ldots,L-1$ has 2^{2*l} sub-regions. Spatial pyramids are just an extension of the traditional BoW approach. In the case of $L = 1$, the spatial pyramid technique is reduced to the standard BoW approach.

3 PYRAMID OF CONCENTRIC CIRCULAR REGIONS

A pyramid of concentric circular regions (PCCR) is proposed in this work to substitute spatial pyramids in object categorization when rotation-invariance is a desirable property. A PCCR is built using L levels, where the last level, level L, is the whole image and the other levels, level l where $l \in 1..L-1$, are concentric circular regions with increasing radii centered in a point P passed as a parameter. The first level has a radius equals a factor R of the smaller axis of the image (height or width). The next level has a radius of $2 \times R$, and so on. The last level, level L, is the entire image. The area of a level l includes the area of level $l-1$. For each region, a set of local descriptors are extracted and we translate these descriptors into codewords. So, each region is represented as a normalized frequency histogram of codewords, following the BoW approach. A set of weights ω_i, with $\sum_{i=1}^{L} \omega_i = 1$, is applied to the histograms of each level. The final descriptor is the concatenation of weighted frequency histogram of visual words of each region. We excluded from the final descriptor the frequency of the codeword used to map the regions outside the circular region. Therefore, for a visual dictionary of C codewords, the size of the final descriptor is $L \times (C-1)$ dimensions.

Figure 1 shows an image of class BACKGROUND_Google and the same image rotated in 196°. The second row shows the plots of the descriptor of both images in the BoW approach using a visual dictionary of 600 codewords. The next row shows a plot of descriptors of both images in BoW + spatial pyramids in 2 levels. The first level has one region and next level has 4 regions. Descriptors of each region are translated to codewords of the same visual dictionary. Final descriptors of each image totalize 3000 dimensions. The last row shows the plots of descriptors in PCCR approach. It was used 4 levels and the

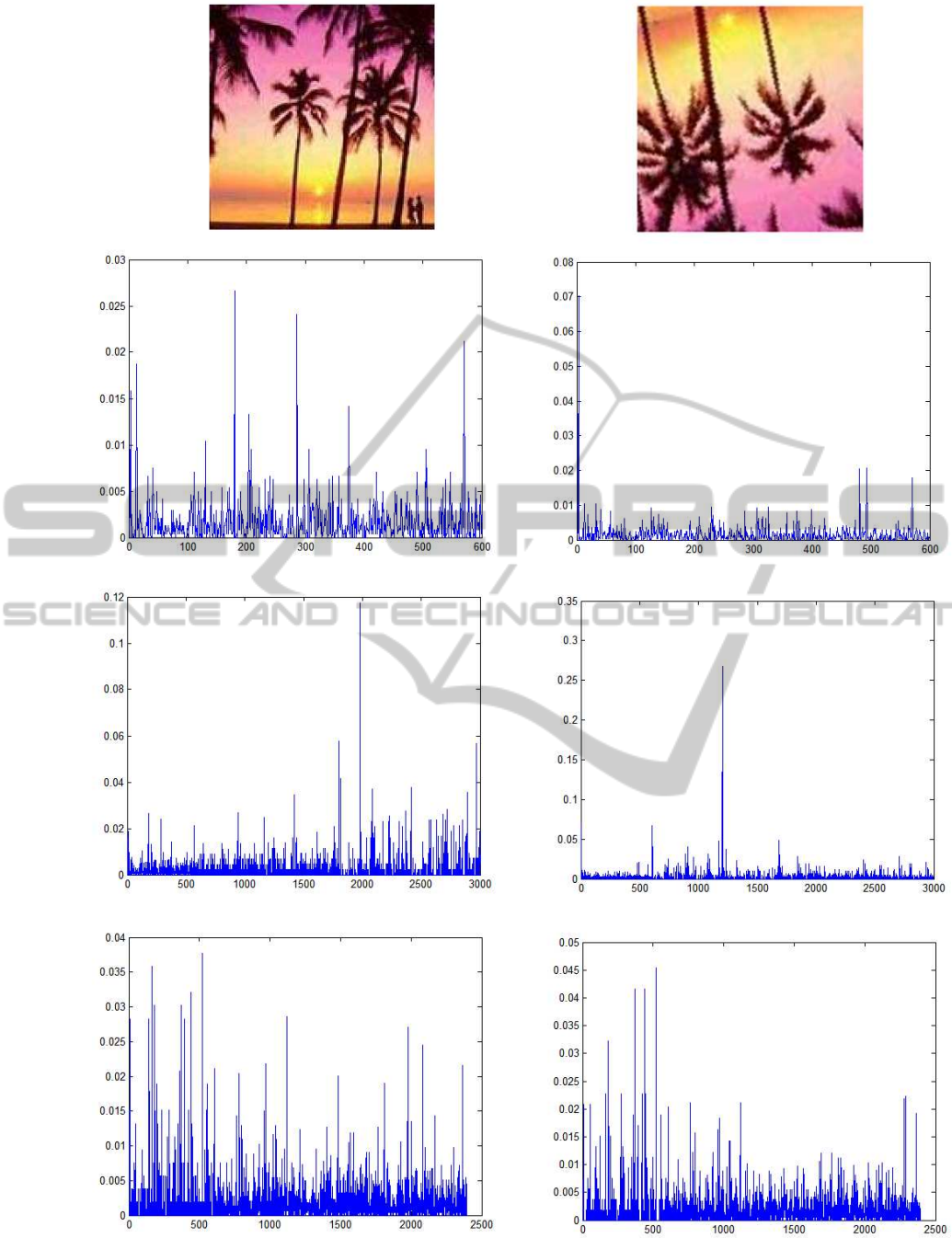

Figure 1: Original image of BACKGROUND_Google and a rotated image (196° of rotation). The second row shows the descriptor of each image in the traditional BoW with 600 codewords. The third row shows the descriptors of both images in BoW + spatial pyramid of 2 levels totalizing 3000 dimensions and the last row shows the descriptors of the images using the PCCR with 2396 dimensions.

same 600 codewords visual dictionary. The final descriptor in PCCR approach has 2396 dimensions for each image ($= 4$ levels $\times (600 - 1)$ codewords).

To calculate the difference between the descriptors computed using the BoW + spatial pyramid and the PCCR, we applied the chi-squared distance between the two final descriptors of original and rotated image. For BoW + spatial pyramid approach, the difference between the two descriptors is 3.4055 and for the PCCR the difference is 1.2146. Therefore,

Figure 2: This figure shows the same images of Figure 1. The original image of the BACKGROUND_Google class of Caltech-101 database and the image rotated in 196°. It shows the 4 regions of the PCCR approach for each image. The first column shows the first circular region with radius of 15% of image's height, the next column shows the second region and so on, until the last column showing the entire image that is the fourth region used to calculate the final descriptor.

for the same rotation, a PCCR produced a smaller difference between the two descriptors. As the proposed spatial approach has 2396 dimensions and the original rectangular spatial pyramids has 3000 dimensions, we calculated the average difference for each dimension. The result for spatial pyramid is $3.4055 \div 3000 = 1.1352 \times 10^{-3}$ and the result for PCCR is $1.2146 \div 2396 = 5.0693 \times 10^{-4}$. So, the PCCR has also a smaller average difference for each dimension.

Figure 2 shows the same images of Figure 1 and each region used to calculate the final descriptor.

4 IMPLEMENTATION

In this section, we describe some of the implementation details. We tested the experiments in Caltech-101 database (Fei-Fei et al., 2006). It has 101 different categories of images. Most of the classes have a relative small number of images, but 5 classes have hundreds of images. This subset of the database is named "Tiny" subset and some papers report results in this subset (Sivic et al., 2005). There is not occlusion in the images, there is just one object in each image and the database is considered easy for object categorization. Table 1 shows the 5 classes of the Tiny subset of Caltech-101 database and the number of samples for each class. The class *BACKGROUND_Google* is composed by random images. There is no specific pattern in the images there. *Faces* is composed by faces and some background area. The class *Faces_easy* shows faces of different people in the same way as the class Faces, but less background area is visible in the images. The class *Leopards* shows different images of leopards in different situations in nature and the class *Motorbikes* shows different motorbikes in lateral view.

Table 1: Classes of tiny subset of Caltech-101 database and number of samples per class.

Class	# of samples
BACKGROUND_Google	467
Faces	435
Faces_easy	435
Leopards	200
Motorbikes	798

The descriptor used is Pyramidal Histogram of Visual Words (PHOW) that is a SIFT-derived descriptor (Bosch et al., 2007). It does not use the SIFT keypoint detector but it computes the SIFT descriptor in a dense grid of points. The SIFT descriptor is computed using different neighbourhood areas. It is often used in conjunction with the BoW approach and the spatial pyramids. The PHOW descriptor, BoW approach and spatial pyramids are reported as the state-of-art in object categorization (Wang et al., 2010).

We use the Support Vector Machines (SVM) classifier to model the classes (Cortes and Vapnik, 1995). It is the most used classifier in object recognition problems (Szeliski, 2011). Training is done using data composed by ordered pairs (x_i, y_i), $i = 1, ..., L$, where $x_i \in R^D$, L is the size of training set and D is the number of dimensions of the training points and $y_i \in \{-1, +1\}$. For the classification, it is necessary

to solve the optimization problem of minimizing:

$$\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{L}\xi_i, \quad (1)$$

where $y_i(x_i.w+b)-1+\xi_i \geq 0 \; \forall \; i$, that is a decision boundary more flexible than a pure linear classification. If the data points are not linearly separable, it is necessary to use the *"kernel trick"* that is a special function, the kernel function, that maps the training points into a higher dimension space where data can be linearly separable. The kernel function depends on the nature of the problem.

5 EXPERIMENTS AND RESULTS

This section reports the experiments done in this work. The experiments used PHOW descriptor and SVM classifier using chi-squared kernel (Section 4). It has used 100 images of each class of the Tiny subset of Caltech-101 database for the training phase. A total of 250 images were used in the test phase, 50 images of each class. A multiclass approach was used, each image of the test set has one object belonging to each one of the 5 classes of the subset. The algorithm responses the most likely class of each image of the test set. PHOW descriptor calculates the SIFT descriptor using different neighbourhood areas. We used neighbourhood areas with radius of 2, 4, 6 and 10 pixels. PHOW descriptor uses a dense grid of points. The distance between two consecutive points is 5 pixels. It was used 3 approaches to deal the problem: BoW, BoW + Spatial Pyramids and the proposed PCCR. For PCCR approach, we used $L = 4$ levels, a visual dictionary of PHOW descriptor with $C = 600$ codewords and the same weight for each level ($\omega_i = \frac{1}{L}$). These values were empirically chosen. The point P, center of concentric circular regions, used were the center of each image.

For each tested approach, two experiments were done: one using the original images of the test set and other one using random angles of rotation applied in the images. Figure 3 shows some of images used and their rotated version. The images are resized by a factor F computed by the following formula:

$$F = \frac{\sqrt{(\frac{height}{2})^2 + (\frac{width}{2})^2}}{height}, \quad (2)$$

where height and width are the original measurements of the image. After the image is resized by the factor F, it is rotated by a random angle and a crop centered keep the final image with the same size of the original image. This procedure avoid to show background color in the final image. Table 2 shows the accuracy obtained by the 6 experiments and Table 3 shows the accuracy of the 6 experiments of each tested class.

Table 2: Accuracy of the experiments using 3 different approaches in rotated and non-rotated test images.

Expt.	Rotation	Approach	Acc.(%)
1	No	BoW	96.4
2	Yes	BoW	42.0
3	No	BoW + SPM	96.8
4	Yes	BoW + SPM	43.2
5	No	PCCR	98.4
6	Yes	PCCR	76.8

Table 3: Accuracy (%) for each class of tested subset of Caltech-101 database in each performed experiment.

Classes	Experiments					
	1	2	3	4	5	6
BCK_Google	84	74	86	90	94	72
Faces	98	2	98	10	100	86
Faces_easy	100	96	100	90	100	84
Leopards	100	22	100	10	98	92
Motorbikes	100	16	100	16	100	50

6 ANALYSIS OF RESULTS

We can see in Experiments 1 and 3 that using the BoW approach and spatial pyramids the problem of object categorization using these 5 classes of Caltech-101 is practically solved. Less than 10 misses in 250 testing images, an accuracy of above 96%. Three classes, Faces_easy, Leopards and Motorbikes, reached 100% of correctness. The two faces classes, Faces and Faces_easy, have a big intraclass similarity but, still, they reached 98% and 100% of accuracy respectively. Other complex class is the BACKGROUND_Google class. It tries to model any other kind of category in the world what is practically impossible. However, in this closed world problem, this class got 84% and 86% of accuracy in the two experiments.

The results are completely different when the testing set is rotated. The overall accuracy of the object categorization dropped to 42% without the use of spatial pyramids (Experiment 2) and 43.2% using spatial pyramids (Experiment 4). A completely different results when compared to Experiments 1 and 3.

Experiment 5 showed our proposed approach in the normal dataset without rotation in the test set. We got a surprisingly 98.4% of accuracy. The overall accuracy did not dropped using this alternative spatial pyramid and there is a significant improvement of results where rotations are presented in the tests.

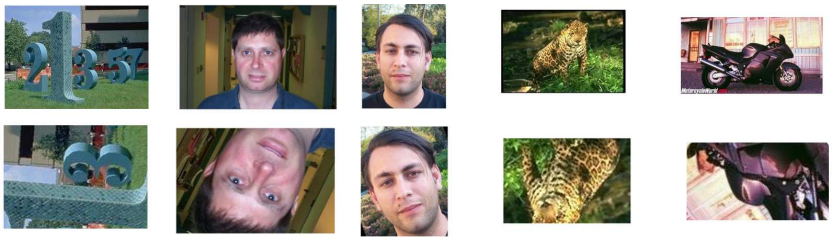

Figure 3: It shows a sample of image for each class of the testing set. In the first line, it shows the original images and in the second line, it shows the rotated image. From left to right column, the classes are: BACKGROUND_Google, Faces, Faces_Easy, Leopards and Motorbikes.

The proposed approach got 76.8% of accuracy against 43.2% of accuracy using the BoW + spatial pyramids. Almost 78% better when compared to the traditional approach.

7 CONCLUSIONS AND FUTURE WORK

In this work, we treat the rotation invariance in the object categorization problem. The state-of-art methods used to deal the problem do not treat the presence of rotation objects in the scene. This property is important for some kinds of objects and scenes. We propose, for spatial quantization, a pyramid that uses a collection of circular concentric areas and showed to be more robust to rotation of objects in the scene. When compared to the traditional spatial pyramids, the results showed very promising results improving the accuracy in almost 78% in a scenary of presence of rotation of objects. As future work, we are planning to test some other parameters for the proposed approach like a different number of levels for the pyramid, different size of radius in each level of the pyramid, different descriptors like SURF that is reported to be more robust to rotation than SIFT and PHOW and to test in a different and more challenging database. In a more realistic scenario, just some objects were rotated while background remains unchanged. Still, rotation can occur around different points in 3D space.

ACKNOWLEDGEMENTS

We would like to thanks the reviewers for the valuable contributions. The authors are partially supported by CNPq, the Brazilian National Research Council.

REFERENCES

Bosch, A., Zisserman, A., and Munoz, X. (2007). Image classification using random forests and ferns. In *11th International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *ECCV International Workshop on Statistical Learning in Computer Vision*, Prague, Czech Republic.

Cula, G. and Dana, J. (2001). Compact representation of bidirectional texture functions. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1041–1047, Kauai, USA. IEEE Computer Society.

Deng, J., Berg, A., Li, K., and Fei-Fei, L. (2010). What does classifying more than 10,000 image categories tell us? *Computer Vision–ECCV 2010*, pages 71–84.

Duda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification*. John Wiley and Sons.

Fei-Fei, L., Fergus, R., and Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions On Pattern Analysis and Machine Intelligence*, 28(4):594–611.

Lazebnik, S., Schmid, C., and Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, pages 2169–2178, New York, USA.

Sivic, J., Russell, R., Efros, A., Zisserman, A., and Freeman, W. (2005). Discovering objects and their location in images. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, pages 370–377, San Diego, USA. IEEE Computer Society.

Szeliski, R. (2011). *Computer Vision: Algorithms and Applications*. Springer-Verlag.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., and Gong, Y. (2010). Locality-constrained linear coding for image classification. In *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, pages 3360–3367, San Francisco, USA. IEEE Computer Society.