Local Description Using Multi-Scale Complete Rank Transform for Improved Logo Recognition

Raluca Boia[1,2], Alessandra Bandrabur[1] and Corneliu Florea[1]
[1]*Image Processing and Analysis Laboratory (LAPI), University "Politehnica" Bucharest, Romania*
[2]*Department of Applied Electronics and Information Engineering, University "Politehnica"Bucharest, Romania*

***Abstract* - In this paper we address the challenging problem of logo recognition. As logos appear in various scales, colors and with illuminations coming from source lights of various intensities, positions or colors, their recognition is a tedious task. The typical approach stands within Bag-of-Words (BoW) framework with SIFT (Scale Invariant Feature Transform) based features. To increase the accuracy, we rely on the recently introduced feature of the complete rank transform, which we extend for a multi-scale description. The system's performance is verified on the FlickrLogos-32 database.**

***Keywords* - logo recognition, object recognition, local image descriptors.**

I. INTRODUCTION

Logo recognition is a challenging pattern recognition task that has been extensively investigated in the last few years. With a large area of applicability starting from modern marketing, advertising, and trademark registration to vehicle recognition, logo recognition has become an important issue to solve. It is also useful for e-business applications, in retrieving and classifying products according to their logos, inspection of industrial goods, identifying the source of documents etc.

While there exists a huge commercial interest to detect logos in images, the problem still remains unresolved. As the number of brands having personalized logos increases every day, such recognition systems must have a large processing capability and be able to support a high number of classes.

A logo is a graphic entity containing colors, shapes, textures, and perhaps text as well, organized in some spatial layout format. There are many challenges that make the building of a logo detection system of high accuracy a very difficult task. Some come from the perspective deformations, varying background and possible occlusions. Besides this, although the objects are almost planar, there are situations when the pattern suffers different deformations, such as warping. Also, there are large variations in size, from high resolution down to 20x20 pixels. Furthermore, the intra-class variability is high, as a certain brand logo can have variations in the colors used or even in the shape of the smaller details. Some examples that partly reveal the issues mentioned can be seen in the Fig. 1.

Previous work. Although the interest in the domain of object recognition is extremely high, the logo recognition has gained rather limited attention. The first approaches concerning it were not able to accurately handle large image collections, as the problem implies [1], [2]. A moment of breakthrough in scalable image object recognition was the introduction of the bag-of-words approach [3] that also lies at the basis of our system. In the past years, a number of works have addressed the logo detection using Bag-of-Words (BoW) techniques. Kleban et al. [4] do logo recognition by performing frequent item-set mining to discover association rules in spatial pyramids of visual words. Revaud et al. [5] use a bag-of-words-based approach coupled with learned weights that down-weight visual words that appear across different classes, while Romberg et al. [6] have developed a system inspired by the bag of visual words approach, through embedding spatial knowledge into the cascaded index. Romberg and Lienhart [7], later created a novel bundling technique on min-hashing of SIFT based visual words. The reasoning behind the choice of bag of words stands in the fact that logo images are designed to draw attention and thus contain sharp edges and contrasting colors that can be appropriately described by keypoint-based representations.

Paper structure. The purpose of our paper is to present a fast logo classification system, designed to efficiently distinguish between a large numbers of classes with high precision. The first phase of the technique proposed consists in extracting the keypoints of the image from hessian-affine interest points [8]. For these local extrema in the image, the descriptors are computed. At the basis of our system, we chose to use two types of local features: SIFT features [8], as they are robust to image tilt and perspective transformations than other features and CRT (Complete Rank Transform) [9] that enriches the amount of information about the shape of the objects, while being highly invariant to changes in the intensity values. The novelty of the proposed methods comes from the introduction of the multi-scale complete rank transform descriptor to complement the standard SIFT, bringing considerable improvements in the logo recognition system.

To exploit the local features depicting the image in an appropriate manner for an accurate classification there is need for a classification system able to cope with variable number of keypoints. The choice at hand is the bag of words technique [3], as it is capable of classifying objects irrespective of the number of features present. The method is able to depict the image with a fixed-size descriptor, overcoming the differences in scale, resolution and aspect ratio between the logos in the same class and it aids the recognition system in being robust to occlusions and viewpoint changes, while having a small computational complexity.

Thus, the paper structure is the following: in section II we present the used features, augmenting the properties of the

978-1-4799-2385-4/14/$31.00 ©2014 IEEE

Figure 1. Sample images from FlickrLogos-32 [6] containing logos from the classes Coca Cola, FedEx, Ferrari and Paulaner. Note the variability due to shadowing, color balance, warping, etc.

very recent Complete Rank Transform, and respectively we describe the classification system. In section III, we introduce the used database, the evaluation method and we present the achieved results. The paper ends with discussions and conclusions.

II. Features and Classification Methods

Given an image, we employ the standard Difference-of-Gausssians locator of points of interest [8]. The vicinity of each such point of interest is described by local features, namely SIFT and multi-scale complete rank transform. The image is thus represented by the concatenation of the key-points local descriptors. The detection and classification system is dealt in the Bag of Words framework.

A. Complete Rank Transform

The Complete Rank Transform (CRT) introduced in [9] is a form of non-parametric local transform, meaning it relies on the relative ordering of local intensity values, and not on the intensity values themselves. As part of a class of descriptors based on the grey value order, it is invariant to illumination changes and monotonic changes in the intensity values. It is an accurate shape descriptor based on ranking the luminance values of the pixels within a square window.

For each pixel, the CRT is obtained by the concatenation of the intensity orders corresponding to the pixels in the considered neighborhood. Consequently, it encodes much more local information than the classical rank transform [10], which retrieves only the rank of the reference pixel or Census signature [10] which consisted in expressing the relationship between the central pixel and each of its neighbors.

Basically, the construction of the complete rank signature implies that for each element of the patch will be allocated its

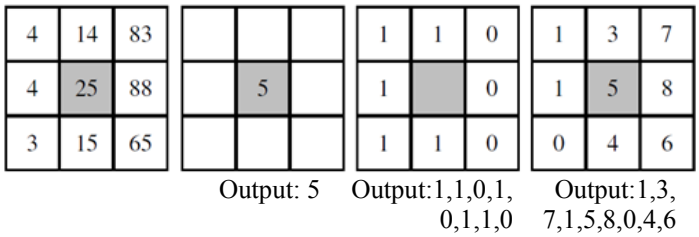


Figure 2. Computation techniques for the non-parametric ordering transforms. (a)Intensity values, (b) Rank Transform, (c) Census Transform, (d) Complete Rank Transform. Figure from [9].

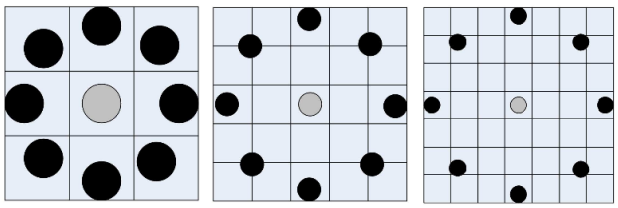


Figure 3. Positions taken into account for the computation of multi-scale CRT.

rank, meaning *the number of pixels with smaller intensity*. The outcome for that window will be the concatenated values of the ranks in the same manner as for census transform. A visual exemplification of the procedure of complete rank transform computation is presented in Fig. 2, while compared to the techniques for simple rank transform and Census transform.

The CRT is usually computed on 3x3 window sizes and for the studied problem it did not lead to improvement. Thus, we investigated several extensions, namely larger windows and multi-scale.

First, we tested the performance of the system when using larger window sizes and the results revealed higher accuracies than when using the classical approach of calculating the CRT. Two methods were used for the multi-scale complete rank transform. For a window size of *dim x dim*, the first method outputs a vector of dim^2 values, representing the ordering of the intensity values for all the positions in the patch. On the other hand, the second technique extracts a vector of 9 values obtained by ranking of the locations in the window that are equally distanced on a circle of radius dim, such as for multi-scale LBP [11]. Fig. 3 shows the positions taken into account for this computation method.

The complete rank transform as presented was introduced in our system in conjunction with the SIFT descriptor. The experiments demonstrated an improvement brought by enlarging the image descriptor with the histogram of complete rank transform.

B. SIFT (Scale Invariant Feature Transform)

Scale-invariant feature transform (SIFT) is an algorithm for extracting features in certain points of interest in the image, that are robust to changes in scale, orientation, shear, position, camera viewpoint and illumination [8]. It is one of the best feature computation techniques regarding robustness and thus widely used for image feature extraction.

The first step of the technique consists in detecting the most important locations in the image. They are named points of extreme or keypoints. The image is processed at various scales

and octaves in order to highlight the points in the image that carry more information than their surroundings. These are the potential candidates for image features computation.

Next, once the initial list of points of interest is established, it has to be refined for more accurate results. More precisely, in this phase the low-contrast keypoints and edge keypoints are eliminated, as these are most likely to be misleading in the final image representation.

In the third phase of the algorithm, orientation assignment is performed. This manages to convert each keypoint and its neighborhood into a set of vectors by computing a magnitude and a direction for them. If this process reveals other peaks in the gradient that were not considered to be starting points so far, they will be added to the series of existing candidates, thus completing the description of the image.

On the final set of interest points, the descriptors are computed. For each location, its neighborhood is divided into sub-blocks in which orientation histograms are computed, composing this way the final descriptor. Each of them is converted into a feature by computing a normalized sum of these vectors.

The choice to use SIFT descriptors for the system is based on the fact that they are more robust than other features to deformations introduced by perspective change and other mentioned image transformations that lead to difficulties in the classification.

C. Bag of Words

Bag of visual words method is a representation technique inspired from text classification [3]. In this model, a text is described as the multiset of its words, disregarding grammar and even word order, but keeping count of the multiplicity. In the same manner, for image classification, local image features are treated as words and their distribution will characterize the entire image.

The first step in classification using Bag of Words model is to extract for the examples in the training database the local features. These are called the *visual words* and form the *visual vocabulary* of the dataset that must be further quantized according to the statistics of the words. To do so, k-means clustering is performed on the data.

The next stage consists in representing each image by frequencies of visual words. For this, every feature will be labeled as the closest visual word, using k-dimensional trees. Then, each image will be represented by a bag of words, meaning a histogram that counts the number of features in the image assigned to each visual word existent. Finally, a classifier is trained to separate the images into the corresponding classes, the descriptor being in this case the histogram of visual words.

III. Implementation and Results

A. Database

For a realistic evaluation of our proposed method, we chose the FlickrLogos-32 database [6], which was formed by careful selecting images from collection of photos in a real word environment, depicting brand logos. For the actual test, we selected a subset consisting from a relatively small number of logo instances per class (i.e. 10 images for training and 10 for testing), but a large number of classes, more exactly, 29: Adidas, Aldi, Becks, BMW, Carlsberg, Chimay, Coca-Cola, Corona, DHL, Esso Erdinger, Fedex, Ferrari, Ford, Google, Guiness, Heineken, HP, Milka, Nvidia, Paulaner, Pepsi, Ritter Sport, Shell, Singha, Starbucks, Stella Artois, Texaco, Tsingtao.

We chose FlickrLogos-32 over another popular database for logo analysis, BelgaLogos dataset [3], as the latter was originally used for logo retrieval rather than for classification and it only defines a small number of images per class.

Compared to other databases for object detection, FlickrLogos-32 can be considered a small-object dataset, by taking in to account the average object size. This adds up to the other challenges brought by the database: the great variance of object sizes - from tiny logos in the background to image-filling views, perspective tilt, important rotations, color and shape changes inside the same class of objects, different occlusions and variable background. Moreover, the difficulty arises from the need to do a multi-class recognition on this large number of classes.

B. Evaluation method

The database used in the experiments conducted is a subset of the Flickr Logo Database, holding 29 classes, each comprised of 10 images for training and 10 for testing. The first stage in the workflow of logo classification consists in extracting the interest points in the train images by using the Difference-of-Gausssians detector [8]. Given the keypoints locations, the image is represented by two local features: the 128-dimensional SIFT descriptor [8] and the multi-scale CRT, used to increase the accuracy of the shape description. Tests were performed with the purpose of finding the optimal parameters for the CRT, the window size and the output vector size.

To quantize the descriptor vectors to visual words, k-means clustering is employed, thus forming the visual vocabulary. Unlike other existing object recognition methods based on bag of words that use large vocabularies with several thousand up to millions visual words [12][13][14], we have used a vocabulary with a rather small size, of 600 words. The reasons lie in the fact that we have to deal with high intra-class variance, thus by choosing a smaller sized vocabulary, we can reduce the quantization errors.

Once the vocabulary is established, all the images in the train and test database are depicted by frequencies of visual words. For this, each feature is labeled as the closest visual word, using k-dimensional trees.

The result is that each image will be represented by a histogram that counts the number of features in the image assigned to each existent visual word. For the classification part, a linear *SVM* is trained to separate the classes of logos, the descriptor being in this case the histograms of visual words. For the implementation of SIFT and of the classifier, we have used the toolbox *vl_feat* presented in [15].

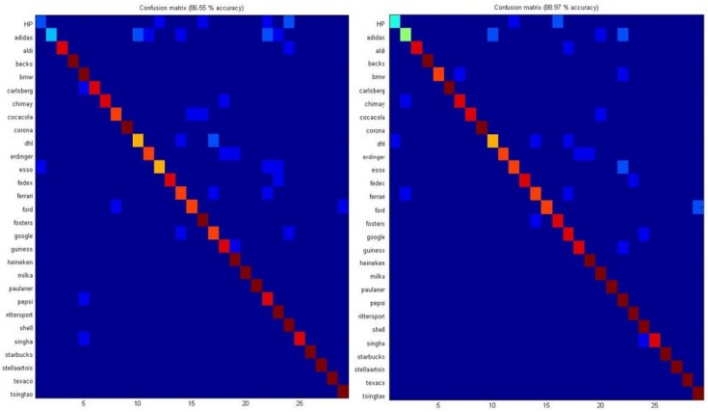
C. Results

The results of the tests revealed increases in accuracy when introducing the CRT descriptor. Experiments proved that the system based only on the SIFT descriptor can classify the logos

Table I

COMPARATIVE RESULTS OF THE THREE METHODS ON THE FLICKR LOGO IMAGES SUBSET. THE CASE RETRIEVING THE BEST ACCURACY IS HIGHLIGHTED.

Local descriptor used	CRT window size	CRT output size	Accuracy [%]
SIFT	0	0	86.55
SIFT+multi-valued CRT	3	9	86.90
	5	25	85.86
	7	**49**	**88.97**
	9	81	86.90
SIFT+multi-scale CRT	5	9	86.21
	7	9	86.90
	9	9	87.59



(a) Accuracy 86.55% (b) Accuracy 88.97%

Figure 4. Confusion matrices for: (a) the system based only on SIFT descriptor and (b) the improved system with the CRT of 7x7 window size added.

with an overall accuracy of 86.55%. When adding the CRT of window size 3x3, the accuracy increases up to 86.90%. Even if this is a small accuracy increase, it indicates that the features based on ranking the intensities complement the edge based SIFT descriptor, improving its performance.

When considering a larger scale CRT, the classification performance of the system is further *increased*, up to *87.59%* for a window size of 9x9, as depicted in Table 1. The best results, showing an accuracy of **88.97%,** were obtained for a window size of 7x7 in the scenario when using all the values in the neighborhood of the central pixel. The confusion matrix for this situation is presented in Fig. 4, in comparison to the case when using only the SIFT descriptor.

IV. CONCLUSIONS

In this paper, we have presented a system for automatic logo recognition based on the bag of words paradigm. Our specific contribution lies in introduction of the multi-scale Complete Rank Transform and using it to complement the standard SIFT local descriptor, leading to increased performance more precisely to improve the recognition rate for the tedious task of logo recognition with 3%.

This is only the first step in our work, as we aim for developing an efficient logo recognition system capable of functioning real-time for a vast number of classes and handling more issues related to the real-world environment. As continuation paths, there is need to extended the system performance for very low resolutions and to accelerate it so bring it closer to real-time.

ACKNOWLEDGMENT

This work was supported by the Romanian Sectoral Operational Programme Human Resources Development 2007-2013 through the European Social Fund Financial Agreements POSDRU/ 107/1.5/S/134398 (KNOWLEDGE), POSDRU/ 89/1.5/S/132397 (ExcelDOC) and POSDRU/ 89/1.5/S/132395 (InnoRESEARCH).

REFERENCES

[1] A. Bagdanov, L. Ballan, M. Bertini, and A. Del Bimbo. "Trademark matching and retrieval in sports video databases". In *Proc. of the ACM International Workshop on Multimedia Information Retrieval*, MIR, 2007, pp. 79-86.

[2] A. Joly and O. Buisson. "Logo retrieval with a contrario visual query expansion". In *Proc. of the ACM International Conf. on Multimedia, ACM MM*, 2009, pp. 581-584.

[3] J. Sivic and A. Zisserman. "Video Google: a text retrieval approach to object matching in videos". *In Proc. of International Conf. on Computer Vision*, *ICCV*, 2003, pp. 1470 - 1477.

[4] J. Kleban, X. Xie, and Wei-Ying Ma. "Spatial pyramid mining for logo detection in natural scenes". In *Proc. of IEEE International Conf. on Multimedia and Expo (ICME)*, 2008.

[5] J. Revaud, C. Schmid, M. Douze, and C. Schmid. "Correlation-Based Burstiness for Logo Retrieval". In *Proc. of the ACM International Conference on Multimedia, ACM MM*, 2012, 965-968.

[6] S. Romberg, L. Garcia Pueyo, R. Lienhart, and R. van Zwol. "Scalable Logo Recognition in Real-World Images". In *Proc. of International Conference on Multimedia Retrieval*, ICMR, 2011.

[7] S. Romberg and R. Lienhart. "Bundle Min-Hashing for Logo Recognition". In *Proc. of International Conf. on Multimedia Retrieval*, ICMR 2013, 113-120.

[8] D. Lowe. "Distinctive image features from scale-invariant keypoints". *International Journal of Computer Vision*, 60(2):91–110, 2004

[9] O. Demetz, D. Hafner and J. Weickert. "The Complete Rank Transform: A Tool for Accurate and Morphologically Invariant Matching of Structures". In *Proc. of British Machine Vision Conf., BMVC*, 2013.

[10] R. Zabih and J. Woodfill. "Non-parametric local transforms for computing visual correspondence". In *Proc. of European Conf. on Computer Vision*, ECCV, 1994, pp. 151–158.

[11] T. Maenpaa and M. Pietikainen, "Multi-Scale Binary Patterns for Texture Analysis", In *Proc. of Scandinavian Conf. on Image Analysis, SCIA*, 2003, pp. 885-892.

[12] O. Chum, M. Perdoch and J. Matas. "Geometric min-Hashing: Finding a (thick) needle in a haystack" In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR*, 2009, pp. 17–24.

[13] D. Nistér and H. Stewénius. "Scalable Recognition with a Vocabulary Tree". In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, *CVPR*, 2006, pp. 2161-2168.

[14] Z. Wu, Q. Ke, M. Isard and J. Sun. Bundling features for large scale partial-duplicate web image search. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, CVPR,* 2009, pp. 25-32.

[15] A. Vedaldi, B. Fulkerson. "Vlfeat: an open and portable library of computer vision algorithms". In Proc. of *ACM International Conf. on Multimedia, ACM MM*, 2010, 1469-1472.