SOCIETY FOR SCIENCE AND EDUCATION
UNITED KINGDOM
TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
ISSN 2054-7390

Image Category Recognition using Bag of Visual Words Representation

[1]Suresh Kannaiyan, [2]Rajkumar Kannan and [3]Gheorghita Ghinea
[1,2]Departement of Computer Science, Bishop Heber College (Autonomous), Tiruchirappalli, India;
[3]Departement of Computer Science, Brunel University, Uxbridge, United Kingdom;
sureshk.naga@gmail.com; dr.rajkumarkannan@gmail.com; george.ghinea@brunel.ac.uk;

ABSTRACT

Image category recognition is one of the challenging tasks due to difference in image background, illumination, scale, clutter, rotation, etc. Bag-of-Visual-Words (BoVW) model is considered as the standard approach for image categorization. The performance of the BoVW is mainly depend on local features extracted from images. In this paper, a novel BoVW representation approach utilizing Compressed Local Retinal Features (CLRF) for image categorization is proposed. The CLRF uses interest point regions from images and transform them to log polar form. Then two dimensional Discrete Wavelet Transformation (2D DWT) is applied to compress the log polar form and the resultant are considered as features for the interest regions. These features are further used to build a visual vocabulary using k-means clustering algorithm. Then this visual vocabulary is used to form a histogram representation of each image where the images are further classified using Support Vector Machines (SVM) classifier. The performance of the proposed BoVW framework is evaluated using SIMPLIcity and butterflies datasets. The experimental results show that the proposed BoVW approach that uses CLRF is very competitive to the state-of-the-art methods.

Keywords: Bag-of-visual-words; Object recognition; Local image features; Interest point detector; Image descriptor.

1 Introduction

Image category recognition has become an important topic of research in the computer vision field. Classifying images is one of the challenging tasks due to large variations in terms of illumination, background, occlusion, clutter, size, etc. among images. The BoVW modeling [1, 2, 3] is a standard approach for image and object category recognition. This image representation technique is derived from the text representation approach called Bag-of-Words (BoW) [4]. The BoVW is very popular for image categorization because of its powerful local image region descriptors (feature vectors) [5] where these vector representations can be easily exploited for classification by using supervised learning techniques such as SVM [5].

The image categorization performance of BoVW modeling heavily depends on the local features extracted from images. Hence, the extraction of local features that are invariant to geometric and photometric changes is necessary. Use of local features such as Scale Invariant Feature Transform (SIFT) [6],

DOI: 10.14738/tmlai.45.2206
Publication Date: 28th September, 2016
URL: http://dx.doi.org/10.14738/tmlai.45.2223

Compressed Local Retinal Features (CLRF) [7], Speeded up Robust Features (SURF) [8] and Multisupport Region Order-Based Gradient Histogram (MROGH) [9] can help to overcome from these challenges. Recently we proposed the CLRF descriptor [7] utilizing log polar transformation and two dimensional Discrete Wavelet Transformation (2D DWT) and proved that the CLRF is very competitive to the state-of-the-art descriptors.

In this paper, BoVW modeling using CLRF descriptor is proposed for image category recognition. In contrast to the previous methods [1, 3, 10, 11], the proposed approach uses CLRF descriptor [7] to build local features. The CLRF descriptor has some advantages over other descriptors. It uses log polar transformation (LPT) that sufficiently preserves interest regions structural information and also has the invariance property to translation and rotation [12]. The CLRF descriptor uses the 2D DWT after applying LPT where the usage of 2D DWT makes the local features very compact and discriminative [7].

The remaining sections of this paper are organized as follows. Literature review on BoVW, local image region description and object recognition are presented in section 2. The proposed image category recognition approach is explained in detail in section 3. The experimental results and discussion of the proposed image categorization approach are presented in section 4. Finally, section 5 concludes the paper with future work.

2 Related work

This section reviews the related work on the BoVW modeling based object categorization and local image region descriptors.

2.1 Bag-of-Visual-Words

In the field of Computer Vision, the image category recognition has been active topic of research over decades. The BoVW modeling derived from the text retrieval approach called Bag-of-Words (BoW) [4] is often used for image representation in category recognition task. The simplistic and efficient histogram representation has made the BoVW very popular among the computer vision community. Following the BoVW, many approaches [13, 14, 15, 16] were proposed for image representation.

Fisher Vectors (FV) [17] is considered as an alternative to BoVW modeling. It uses Gaussian Mixture Model (GMM) to describe image patches. The gradient computation is restricted to the mixture weight parameters of the GMM. Incorporation of additional gradients to fisher vectors provides large improvements in terms of accuracy [18]. Since it can be computed from much smaller vocabularies, it is computationally very efficient [18]. The FV works well even with simple linear classifiers. It is noted in [18] that even though FV works better in image representation, it has an important disadvantage comparing to the BoVW. The FV image representation is almost dense while BoVW representation is sparse [18].

In [19], a new image representation framework called sparse vectors is proposed using local visual descriptors. It performs a nonlinear feature transformation on descriptors then aggregates the resultant features together to construct image-level representations, and then it applies the linear SVM for image classification [19]. This sparse vectors approach is scalable in computation [19]. The VLAD [20, 21] is proposed by considering classification accuracy, efficiency and memory utilization. It aggregates local image region descriptors into a vector of low dimensional feature that preserves the quality of vector comparison [21]. Even though several improvements to image representation are proposed, the BoVW is

highly preferred for image representation [22] because of its computation simplicity and effectiveness in image and object categorization applications.

2.2 Image region descriptors

To achieve desired accuracy in image category recognition, the BoVW model should use local image features that are invariant to geometric and photometric transformations. Scale Invariant Feature Transform (SIFT) [6] has been utilized in many object recognition approaches to extract invariant features from images. Though SIFT is a very popular keypoint (or interest region) description approach, its higher dimensionality makes the computation of object recognition system very complex. Following SIFT, Speeded up robust features [8] is proposed that is computationally very faster than the SIFT descriptor. Though SURF [8] is very fast in computing keypoint features, it is not much robust to geometric transformations such as rotation and view point changes. Multisupport Region Order-Based Gradient Histogram (MROGH) [9] is proposed using multiple support region around each interest point to improve feature discrimination. Since the MROGH use multiple support regions, the computation complexity is very high.

Local binary pattern (LBP) [23] based image region descriptors called Orthogonally Combined Local Binary Patterns (OCLBP) [22] and Center Symmetric Local Binary Patterns (CSLBP) [24] have shown better performance in image matching and object recognition tasks. The CSLBP and OCLBP are the dimensionality reduced versions of LBP operator that achieved competitive performance than the SIFT descriptor. Recently, the CLRF [7] has achieved better image matching accuracy than the SIFT descriptor in image matching even though it has less dimensional features than the SIFT features.

3 BoVW image representation using CLRF

In this section, the details of the proposed BoVW based image categorization approach is presented. The block diagram of the image categorization approach is depicted in figure 1. The details of the CLRF local features construction and BoVW modeling are explained in this section.

3.1 CLRF descriptor

First, interest points from images are detected using the Hessian keypoint detector [25] and a patch containing 41×41 pixels surrounding each keypoint is used as interest region. Then, the CLRF [7] descriptor is used to extract features from gray interest regions of an image. Here, the image patches are not rotated to their dominant orientation because this rotation invariance is useful for image matching but decreases the image categorization accuracy [22]. The construction of the CLRF has two steps: 1. Apply log polar transformation (LPT) on each interest region and 2. Apply 2D DWT after applying LPT on interest regions. The computation details of the LPT and 2D DWT are given below.

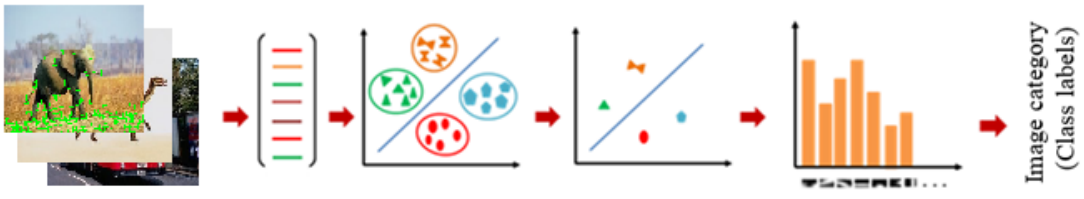


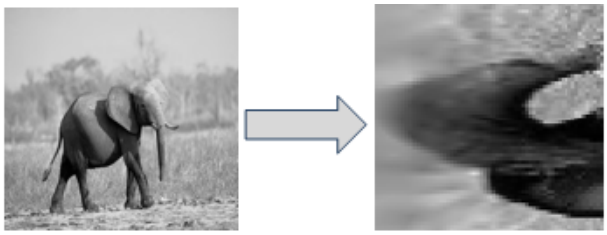
Figure 1. Block diagram of the BoVW based image categorization approach

The LPT converts each interest region from Cartesian coordinates $I(x, y)$ to the log-polar coordinates (ρ, θ) [7] using the equations given below,

$$\rho = \log_{\text{base}} \sqrt{(x - x_c)^2 + (y - y_c)^2} \quad (1)$$

$$\theta = \tan^{-1} \frac{y - y_c}{x - x_c} \quad (2)$$

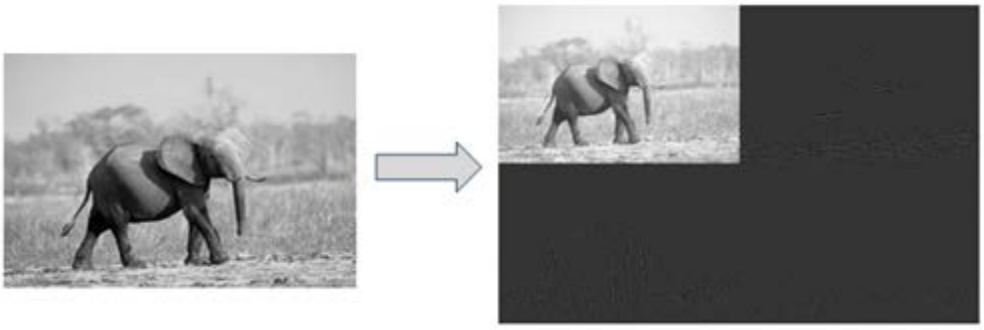
where (x, y) denotes the sampling pixels, (x_c, y_c) denotes the center of the Cartesian coordinates, and ρ and θ denote different radius and angular sampling positions in the polar geometric structure respectively [7]. The resultant of the LPT is a two dimensional matrix with entries (ρ, θ). Sample image after applying the log polar transformation is shown in figure 2.



(a) (b)

Figure 2. Applying Log polar transformation on image. (a) original image and (b) image after applying log polar transformation

High dimensional features make the system computationally expensive. Hence to reduce the feature dimension, the 2D DWT is used. The 2D DWT decomposes the matrix (image) and splits into an approximation image and three detailed images using a pair of low-pass and high-pass filters as shown in figure 3. The approximation image of each interest region are converted into vectors and kept as feature descriptors [7] which is called CLRF. These CLRF features are then used to build the BoVW as described in the following subsection.



(a) (b)

Figure 3. Applying 2D DWT (Level 0 decomposition). (a) Original image and (b) after applying 2D DWT on original image (images are shown in gray scale)

3.2 Bag-of-Visual-Words representation

Once the features are constructed using the CLRF, the BoVW modeling is used to represent each image in terms of a histogram of visual dictionary. The BoVW constructs an order-less collection of interest region descriptions called visual vocabulary (otherwise called visual dictionary) using K-means clustering algorithm. Here, the cluster centers are considered as visual words. Using the BoVW, each image is represented with standard dimension. The size of the visual dictionary needs to be set carefully, since a small size leads to less discrimination where a very large size leads to slower computation speed and less generalizable features. Thus, the visual vocabulary size is chosen appropriately to maintain a good trade-off between generalization and discrimination. Using the constructed visual words, each image feature is quantized into their closest visual word. It produces a histogram feature for each image according to the number of local descriptors assigned for the corresponding visual words. The resultant histogram features are very compact, informative and fixed–length representation [22] which are further exploited for image classification.

3.3 Classification

The popular image classification technique SVM [5] is exploited to identify each image category. The SVM training and prediction are performed using the following equation,

$$K_{\chi^2}(S,M) = e^{(-1/D)dist_{\chi^2}(S,M)} \tag{3}$$

where *S* and *M* are two BoVW image features, χ^2 is chi-square distance measure and it is calculated using the following equation,

$$\chi^2(f, f') = \sum_{i=1}^{n} \frac{(f_i - f_i')^2}{f_i + f_i'} \tag{4}$$

Here, *f* and *f'* are two features.

4 Experimental results

In this section, the details of the SIMPLIcity and butterflies datasets and experimental setup are presented. Then the performance of the proposed image categorization approach is presented and compared with the state-of-the-art methods.

Figure 4. Sample images from the SIMPLICity dataset [26] is given here. The image categories are African People, Beach, Building, Bus, Dinosaur, Elephant, Flower, Horse and Mountain and Food (from top left to bottom right).

4.1 Performance on SIMPLIcity dataset

The SIMPLIcity dataset [26] contains 10 different classes of images and each class contains 100 images. The 10 different categories are African people, beach, building, bus, elephant, flower, food, horse, dinosaur, and mountain. Each category of images contains images with the size of either 256×384 or 384×256 pixels. Sample images from the SIMPLIcity dataset are shown in figure 4. Local features applying image region descriptor (SIFT, CSLBP, OCLBP or CLRF) are extracted. Here the dimensions of the local feature descriptors SIFT, CSLBP, OCLBP and CLRF are 128, 256, 512 and 72 respectively. Each of these descriptors are used to construct a visual vocabulary of size 1000 visual words using the K-means clustering algorithm. Using these vocabularies, all the image features are quantized into a standard dimension. For evaluation, 50% of images from each image category is chosen for training and remaining for testing. The SVM classifier is used for training and to predict the class of each testing images. The results are presented in the table 1. From the average image classification accuracy in table 1, it can be seen the proposed BoVW model which uses the CLRF outperforms all the comparative approaches. It can also be understood that the CLRF based BoVW approach is competitive to other comparative approaches for individual image categories.

Table. 1. Image classification accuracy (%) on SIMPLIcity dataset

Methods	People	Beach	Building	Bus	Elephant	Flower	Food	Horse	Dinosaur	Mountain	Average
BoVW using SIFT	66	52	44	90	100	78	74	88	46	38	67.6
BoVW using CSLBP	66	36	50	76	100	36	56	70	42	48	58.0
BoVW using OCLBP	64	54	64	98	96	42	56	68	46	66	65.4
BoVW using CLRF	66	54	52	98	100	74	74	70	46	64	**69.8**

4.2 Performance on butterflies dataset

The Butterflies dataset [27] contains 619 images with 7 categories of butterflies such as Admiral, Black Swallow tail, Machaon, Monarch open, Monarch closed, Peacock and Zebra, where each category contains 111, 42, 83, 84, 74, 134 and 91 images respectively. Sample images from the butterflies dataset are given in figure 5. In this dataset, each class of butterfly contains geometric (rotation and translation) and photometric variations (blur, affine and illumination). A visual vocabulary is generated using first 10 images from each category. The K-means clustering algorithm is used to generate visual words for each descriptor separately. Here, K is set to 500, so the dimension of visual word is 500. The butterflies classification results are presented in table 2. Since LPT is robust for translation and rotational variations, the proposed approach shows better performance than other approaches.

Figure 5. Sample images from the butterflies dataset [27]. Top row contains categories of admiral, Machaon and Monarch_closed (two images per category). Bottom two rows contain Monarch_open, Peacock, Black_swallowtail and Zebra (three images per category)

Table 2. Classification performance on Butterflies dataset

Methods	Admiral	Black swallow tail	Machaon	Monarch open	Monarch closed	Peacock	Zebra	Average
BoVW using SIFT	51.79	33.33	52.38	56.76	54.76	61.19	77.78	55.43
BoVW using CSLBP	60.71	19.05	59.52	27.02	64.29	53.73	55.56	48.55
BoVW using OCLBP	37.50	52.38	50.00	32.43	59.52	53.73	64.44	50.00
BoVW using CLRF	60.00	70.00	53.49	54.78	86.49	74.63	68.89	**66.89**

5 Conclusion

In this paper, a novel BoVW modeling approach utilizing the CLRF image region descriptor is proposed for image category recognition. Exploitation of the CLRF descriptor has shown some important advantages. It sufficiently preserves interest regions structural information while constructing compact and discriminative features using the log polar transformation and 2D DWT. Performance of the proposed approach is evaluated using the popular SIMPLIcity and butterflies datasets. The experimental results showed that the proposed BoVW using CLRF descriptor achieved better accuracy than the comparative approaches for image categorization. Though, the proposed BoVW shows competitive performance in image categorization, the discriminative power can be further improved by extracting local features from double or multiple support regions around each interest point. That will be focused in the future work of this paper.

REFERENCES

[1]. Yang, J, et., Evaluating bag-of-visual-words representations in scene classification. In Proceedings of the international workshop on Workshop on multimedia information retrieval, 2007. p. 197-206.

[2]. Tirilly, P., Claveau, V., and Gros, P., Language modeling for bag-of-visual words image categorization. In *Proceedings of the 2008* international conference on Content-based image and video retrieval, 2008. p. 249-258.

[3]. Yang, Y., and Newsam, S., Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems, 2010. p. 270-279.

[4]. Sivic, J., and Zisserman, A., Video Google: A text retrieval approach to object matching in videos. In Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, 2003. p. 1470-1477.

[5]. Cortes, C., and Vapnik, V., Support-vector networks. *Machine learning*, 1995, 20(3): p. 273-297.

[6]. Lowe, D. G., Distinctive image features from scale-invariant keypoints. International journal of computer vision, 2004, 60(2): p. 91-110.

[7]. Kannan, R, et al., CLRF: Compressed Local Retinal Features for Image Description. In Advances in Pattern Recognition (ICAPR), Eighth International Conference on, 2015, p. 1-5.

[8]. Bay, H., Tuytelaars, T., and Van Gool, L. Surf: Speeded up robust features. In European conference on computer vision, 2006. p. 404-417.

[9]. Fan, B., Wu, F., and Hu, Z. Rotationally invariant descriptors using intensity order pooling. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012. 34(10): p.2031-2045.

[10]. Peng, X., et al., Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. Computer Vision and Image Understanding, 2016. p. 109-125.

[11]. Karakasis, E. G., et al., Image moment invariants as local features for content based image retrieval using the bag-of-visual-words model. Pattern Recognition Letters, 2015, 55, p. 22-27.

[12]. Pun, C. M., and Lee, M. C. Log-polar wavelet energy signatures for rotation and scale invariant texture classification. IEEE transactions on pattern analysis and machine intelligence, 2003, 25(5): p.590-603.

[13]. Kavukcuoglu, K., et al., Learning convolutional feature hierarchies for visual recognition. In Advances in neural information processing systems, 2010, p. 1090-1098.

[14]. Lazebnik, S., Schmid, C., and Ponce, J., Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006. 2: p. 2169-2178.

[15]. Goh, H., et al., Unsupervised and supervised visual codes with restricted boltzmann machines. In European Conference on Computer Vision, 2012. p. 298-311.

[16]. Vedaldi, A., et., Multiple kernels for object detection. In 2009 IEEE 12th international conference on computer vision, 2009. p. 606-613.

[17]. Perronnin, F., and Dance, C., Fisher kernels on visual vocabularies for image categorization. IEEE Conference on Computer Vision and Pattern Recognition, 2007. p. 1-8.

[18]. Sanchez, J., et al., Image classification with the fisher vector: Theory and practice. International journal of computer vision, 2013. 105(3): p. 222-245.

[19]. Zhou, X., et al., Image classification using super-vector coding of local image descriptors. In European conference on computer vision, 2010. p. 141-154.

[20]. Jegou, H., et., Aggregating local descriptors into a compact image representation. In Computer Vision and Pattern Recognition, 2010 IEEE Conference on, p. 3304-3311.

[21]. Arandjelovic, R., and Zisserman, A. All about VLAD. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013. p. 1578-1585.

[22]. Zhu, C., Bichot, C. E., and Chen, L. Image region description using orthogonal combination of local binary patterns enhanced with color information. Pattern Recognition, 2013. 46(7): p.1949-1963.

[23]. Ojala, T., Pietikainen, M., and Harwood, D. A comparative study of texture measures with classification based on featured distributions. Pattern recognition, 1996. 29(1): p. 51-59.

[24]. Heikkila, M., Pietikainen, M., and Schmid, C. Description of interest regions with local binary patterns. Pattern recognition, 2009. 42(3): p. 425-436.

[25]. Mikolajczyk, K., et al., A comparison of affine region detectors. International journal of computer vision, 2005. 65(1-2): p.43-72.

[26]. Wang, J. Z., Li, J., and Wiederhold, G. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. Pattern analysis and machine intelligence, IEEE Transactions on, 2001. 23(9): p.947-963.

[27]. Lazebnik, S., Schmid, C., and Ponce, J. Semi-local affine parts for object recognition. In British Machine Vision Conference, 2004. p. 779-788.

