Fourier Dönüşüm Tabanlı Betimleyici Kullanarak Görsel Nesne Sınıflandırma
Return of the King: The Fourier Transform Based Descriptor for Visual Object Classification

Hakan Çevikalp
Electrical and Electronics Engineering Dept.
Eskişehir Osmangazi University
hcevikalp@ogu.edu.tr

Zühal Kurt
Mathematics and Computer Dept.
Eskişehir Osmangazi University
zkurt@ogu.edu.tr

Ahmet Okan Onarcan
Porsuk Vocational School
Anadolu University
aoonarcan@anadolu.edu.tr

ÖZETÇE

Literatürdeki en başarılı görsel nesne sınıflandırma teknikleri imgeleri betimlemek için "kelimeler çantası" (bag of words) yöntemini kullanmaktadır. Bu yöntemde dijital imgelerden seçilen yamalar SIFT, LBP ve SURF gibi farklı şekil ve doku betimleyicilerle vektörel değerlere dönüştürülmektedir. Bu çalışmada biz Fourier dönüşümünün ağırlıklandırılmış açılarının histogramlarını kullanan yeni bir betimleme tekniği önerdik. Önerdiğimiz betimleme tekniğini kullanan kelimeler çantası tabanlı görsel nesne sınıflandırma yönteminin başarısını literatürdeki diğer betimleme tekniklerini kullanan nesne sınıflandırma yöntemleriyle Caltech-4 ve Coil-100 veri tabanları üzerinde karşılaştırdık. Deneysel sonuçlar Fourier dönüşüm tabanlı betimleyicinin oldukça iyi sonuçlar verdiğini göstermektedir. Ayrıca önerilen betimleyici ile literatürdeki betimleyicilerin beraber kullanılmasıyla birlikte sınıflandırma başarımının daha da arttığı gözlenmiştir. Bu da Fourier dönüşüm tabanlı betimleyicin diğer betimleyicilerden farklı bilgiler taşıdığını göstermektedir.

Anahtar Kelimeler: betimleyici, görsel nesne sınıflandırma, Fourier dönüşümü, kelimeler çantası modeli.

ABSTRACT

Most of the state-of-arts visual object classification methods use bag of words model for image representation. In this method, patches extracted from images are described by different shape and texture descriptors such as SIFT, LBP, SURF, etc. In this paper we introduce a new descriptor based on weighted histograms of phase angles of local Fourier transform (FT). We compare the classification accuracies obtained by using the proposed descriptor to the ones obtained by other well-known descriptors on Caltech-4 and Coil-100 data sets. Experimental results show that our proposed descriptor provides good accuracies indicating that FT based local descriptor captures important characteristics of images that are useful for classification. When we combined image representations obtained by FT descriptor with the representations obtained by other descriptors, results even get better suggesting that tested descriptors encode differential complementary information.

Keywords: descriptor, visual object classification, Fourier transfom, bag of words model .

1. INTRODUCTION

Visual object classification can be defined as the task of assigning an image one or multiple labels corresponding to the presence of a visual object class. It is an important task, and a successful visual object classification system may significantly enhance the performances of other major computer vision applications such as image retrieval and object localization. The major difficulty in object classification is due to the large intra-class variations and viewpoint changes in all of the categories. In addition to this, lighting-scale changes, complex backgrounds, occlusion and presence of noise in the images make the problem even harder.

Most of the recent state-of-art object classification methods use bag of words (BoW) model, which was first used for text classification. After extension of this model to the visual object classification by Csurka et al. [1], such representations have been widely used for both object classification and localization [3, 9, 19, 20, 21]. The BoW model treats each image as an ordered collection of representative patches. Therefore, it requires sampling a set of patches from the image, computing descriptor vectors for each patch, quantization of descriptors, and accumulating histograms or signatures of patch appearances based on this quantization to obtain the final image representation. Then, resulting image feature histograms are fed to a classifier (which is previously trained by using manually labeled image feature histograms) to determine the label(s) of the image category. Although BoW models ignore spatial relationships between the features, they surprisingly work well for object classification because of the high discriminative power of some words. They also have good resistance to occlusions, geometric deformations, and illumination variances.

There are basically three major implementation issues in BoW: how to sample patches from image, how to describe them (descriptor selection), and how to quantize the resulting descriptors. This is also known as codebook generation. Patches are typically sampled from the image at many different positions and scales, either densely [3,6], randomly [5], based on the output of some kind of salient region detector [1,20], or based on the output of segmentation algorithms [7,8]. Then chosen patches are described by using different descriptors. On one hand, the descriptors extracted from patches should be invariant to variations due to the image transformations, lighting variations and occlusions, which are irrelevant to the categorization. On the other, they must carry enough information to discriminate between the object categories.

978-1-4673-5563-6/13/$31.00 ©2013 IEEE

This work was funded by the Scientific and Technological Research Council of Turkey (TUBITAK) under Grant number EEEAG-109E279.

Among these, histogram based descriptors have become very popular owing to their good performance and efficiency. Many of these are based on oriented image gradients, including SIFT [4], SURF [15], Histograms of Oriented Gradients (HOG) [17], Generalized Shape Context [16]. Others are based on local patterns of qualitative gray level differences, including Local Binary Patterns (LBP) [11], and Local Ternary Patterns (LTP) [18]. The resulting descriptors are then clustered to obtain visual words (dictionary). Descriptors extracted from an image are assigned to visual words based on some similarity measure, and the final image vector representations are obtained by accumulating histograms of occurrences of each visual word. Therefore, identification of such visual words is important from two aspects: Firstly, it provides some robustness against descriptor variations since similar patch descriptors are assigned to the most similar visual word. Secondly, it provides a fixed length representation vector for images with different sizes. Different quantization algorithms are proposed to address the quantization process. Among these, quantization algorithms based on k-means clustering [1,19], mean shift [3], hierarchical clustering [21], randomized trees [2] are a few to name. Finally, the image feature vectors obtained from quantization algorithm are fed to classifiers such as Nearest Neighbor, Naïve Bayes or Support Vector Machines (SVM) to determine the label(s) of the visual object category.

In this study, we focus on descriptors that are used to represent the image patches, and propose a new descriptor based on weighted histograms of phase angles of Fourier Transform for BoW model based object classification. The remainder of the paper is organized as follows: In Section 2, we describe the proposed descriptor. In Section 3, experimental results are given. Lastly, our conclusions are presented in Section 4.

2. THE METHOD

2.1. Motivation

In order to motivate the proposed descriptor, we begin with a very well-known example illustrated in Fig. 1, which demonstrates the importance of the phase information of the FT for image representation. When we plot the magnitude and phase information of an image, the phase image looks like some kind of noise that does not include any important information regarding the image. However, when we reconstruct the image by inverse FT using only magnitude information, the image is mostly blank and does not carry any representative information. On the contrary, when the phase information is used for inverse FT, the resulting image is similar to the original one as illustrated in the figure. This clearly shows the phase information carries more characteristic information regarding the image compared to the magnitude of FT.

Using FT as descriptor is not new [10,12,13,14]. But, the most of the proposed descriptors use magnitude of FT to gain rotational invariance, or they apply the FT other image features such as gradients or LBP features. In [10,13], the authors use local 1-D FT histograms of gradient images for texture recognition. However, they ignore the phase information and use magnitude to obtain image representation which is invariant to image rotations (unlike magnitudes, phases of FT are sensitive to image rotations). In [12], the authors extract 1-D FT of 3×3 neighborhoods, and the image is represented by concatenating the histograms of magnitudes and phases. Finally, Ahonen et al. [14] introduce a rotation-invariant descriptor by using magnitudes of FT which is applied to LBP descriptors extracted from images.

In this paper we also propose a descriptor which uses FT of local patches as illustrated in Fig. 2. However, unlike the other methods described above, we use 2D FT of local patches and we apply it directly to the image gray-level values (not to gradients or LBP values) of patches extracted from the images during bag of words model construction. Our histogram construction also differs from the other methods in the sense that we obtain histograms of phase angles weighted by magnitude values as described below.

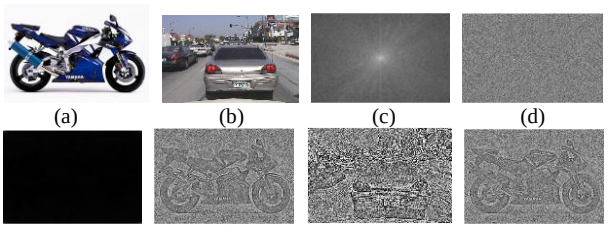

Figure 1. (a) motorbike image; (b) car image; (c) magnitude of the FT of image (a); (d) phase of the FT of image (a); (e) the image reconstructed by inverse Fourier transform using only magnitude of (a) with random phase information; (f) the image reconstructed by inverse FT using phase of (a) with random magnitude; (g) the image reconstructed by inverse FT using phase of (b) with random magnitude; (h) the image reconstructed by inverse FT using phase of (a) with magnitude of (b).

2.2. Bag of Words Based Object Classification With Fourier Transform Descriptor

BoW based visual object classification is illustrated in Fig. 3. In this method, patches are sampled from images at many different positions and scales by using different sampling techniques. This is followed by extracting fixed-size features from the patches by using various descriptors such as SIFT, SURF, LBP, LTP, etc. The resulting patches from all training images are then clustered to obtain visual words. During image representation, descriptors extracted from an image are assigned to visual words based on some similarity measure, and the final feature vector is obtained by accumulating histograms of occurrences of each visual word.

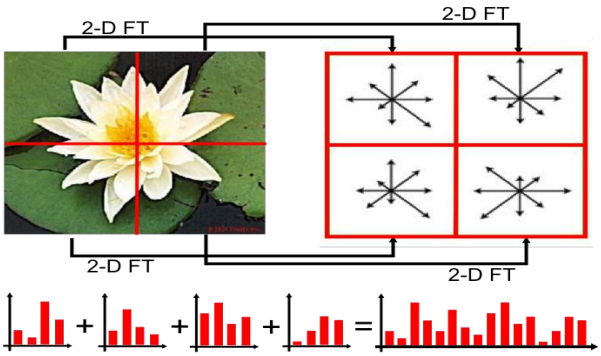

Figure 2. Computing FT descriptor. Image patch is first divided into 2×2 cells in this example. 2D FT is extracted for each cell, and the histogram of each cell is created by accumulating weighted votes to phase angle bins. The final descriptor is formed by concatenating the histograms of all cells.

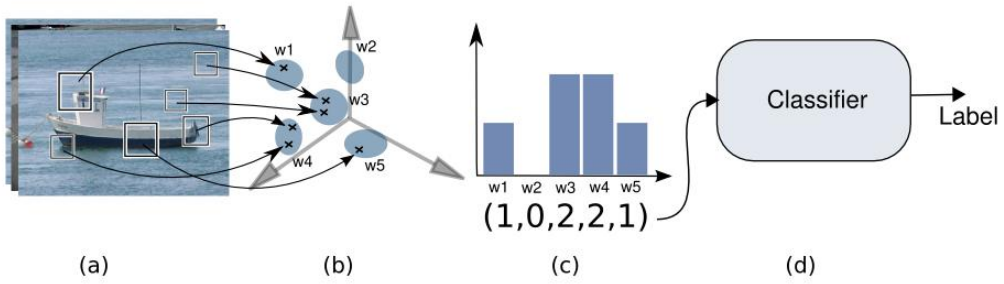

Figure 3. Visual object classification based on bag of words model.

Here we propose a new descriptor based on FT which may be an alternative to popular descriptors such as SIFT. To this end, we divide the patch region into $n \times n$ non-overlapping sub-regions (cells) as shown in Fig. 2. Then, we find the magnitudes and phase angles of 2D FT of each sub-region. Descriptor is formed by accumulating weighted votes to phase angle bins. The phase angle bins are evenly spaced into 19 intervals over $[0,2\pi]$. The vote may be a function of the FT magnitude, either the magnitude itself, its square, its square root, or a clipped form of the magnitude representing soft presence/absence of the phase angle bin. We used FT magnitude to vote the bins, and then the resulting histogram is L1 normalized. The dimensionality of the final descriptor vector is fixed for any patch and it is equal to $d = n^2 \times 19$.

It should be noted that the resulting descriptor is not rotational invariant similar to other popular descriptors such as SIFT, HOG, and LBP. We believe that the proposed descriptor provides richer information compared to other well-known descriptors. Consider SIFT for example: It uses weighted histograms of gradient angles. This is information is already contained in high-frequency components of FT. In addition to this, low-frequency components of FT capture information about the general appearance of patches.

Figure 4. Some image samples from Caltech-4.

3. EXPERIMENTS

We tested the proposed descriptor on two visual object classification databases: the Caltech-4 and Coil-100. BoW model is used for representation of images, and we compared the proposed FT descriptor to SIFT, LBP, LTP, and SURF. We also combined the image representations obtained by FT descriptor with the representations obtained by other descriptors to judge whether the descriptors complement each other in the sense that they capture different information which are useful for classification. We divided the patches into 6×6 cells, thus the dimensionality of the FT descriptor is 684. We used nonlinear Support Vector machine (SVM) classifier for classification in all experiments, and both one-against-rest and one-against-one procedures are used to extend the binary SVM classifier to multi-class classification. We only report the results for the one that yields the best accuracies. The Gaussian kernel $k(x,y) = \exp(-\|x-y\|^2 / \sigma)$ is used as a kernel in nonlinear SVM in all experiments.

3.1. Experiments on the Caltech-4 Database

The Caltech-4 database [22] includes images of objects belonging to 4 visual categories; airplanes, cars, faces, and motorbikes. There are respectively, 1074, 526, 450, and 826 images per category. Each class includes images of highly variable object poses under different lighting conditions with complex backgrounds as shown in Fig. 4. We used two techniques for sampling patches: dense sampling and sampling based on DoG (Difference of Gaussians) interest point detector [4]. All tested descriptors are computed by using the same patches. The descriptors extracted from training images are clustered by k-means clustering method and cluster centers are considered as visual words forming the visual vocabulary. The size of the visual vocabulary is set to 1,000. To build image representation, each extracted descriptor is compared to the visual words and associated to the closest word. The final image feature histogram vector is L1 normalized.

Table 1: Classification rates (%) for different descriptors on the Caltech-4 dataset.

Descriptors	Classification Rates	Classification Rates for Dense Sampling
FT	83.27 ±3.9	97.74 ±2.1
SIFT	81.36 ±5.9	96.94 ±0.9
SURF	75.09 ±3.4	96.35 ±1.5
LBP	86,54 ±2.8	94.74 ±3.8
LTP	84.35 ±2.5	95.30 ±3.2
FT+LBP	91.93 ±2.3	98.36 ±2.3
FT+LTP	90.96 ±1.4	97.35 ±2.5
FT+SURF	86.15 ±3.2	97.74 ±1.9
FT+LBP+SURF	**94.23** ±1.6	**98.78** ±2.1
FT+LTP+SURF	93.01 ±1.2	96.80 ±2.6

Figure 5. Forty objects chosen from Coil-100 dataset.

The classification rates obtained by descriptors are given in Table 1. Classification accuracies computed by using 5-fold cross validation. The second column of the table shows the

classification accuracies obtained by sampling patches using DoG interest point detector and the last column shows the accuracies for dense sampling. Our proposed FT descriptor achieves the best accuracy for dense patch sampling whereas LBP descriptor yields the best accuracy for detector based patch sampling. When we combine the image feature histograms returned by individual descriptors, classification accuracies get better indicating that the descriptors capture complementary information. Results also show that dense patch sampling gives much better results compared to the detector based patch sampling.

Table 2: Classification rates (%) on the Coil 40 dataset.

Descriptors	**Classification Rates for Dense Sampling**
FT	93.30 ±2.2
SIFT	89.45 ±2.5
SURF	87.88 ±1.9
LBP	88.36 ±1.3
LTP	89.26 ±3.3
FT+LBP	92.54 ±1.4
FT+LTP	93.37 ±2.3
FT+SURF	93.18 ±2.3
FT+LBP+SURF	93.21 ±1.5
FT+LTP+SURF	**94.26** ±2.0

3.2. Experiments on the Coil-100 Database

The Coil 100 database [23] includes images of objects belonging to 100 visual categories. There are 72 images for each category. The size of each image is 128×128 pixels. Each class includes highly variable poses of the same object under same lighting conditions. For our experiments, we chose 40 classes (shown in Fig. 5) from the database.

The images are easy in the sense that the background is uniform. Here we only used dense sampling since dense sampling produced better results on Caltech-4 images. As in the previous case, we set the visual vocabulary size (hence, the dimensionality of the image feature vectors) to 1000.

The classification rates are given in Table 2. Among all tested descriptors, our proposed FT descriptor provides the best accuracy followed by SIFT. Similar to the previous case, when we combine the image representations obtained by using different descriptors, accuracies get higher.

4. CONCLUSION AND FUTURE WORK

We have introduced a new descriptor using FT of an image patch for visual object classification based on BoW models. In contrast to the traditional FT based descriptors using the magnitude information, we emphasize on the phase information which is largely ignored in the literature mostly due to the concerns related to the rotational invariance. The proposed descriptor is formed by accumulating weighted histograms of phase angles of FT. Our initial results on small data sets are very encouraging, and as a future work we are planning to test our proposed descriptor on more challenging data sets including many classes. We will also try different overlapping/non-overlapping cell sizes, different bin sizes, and different normalizations of histograms during descriptor computing to improve the classification accuracies further. Lastly, we are planning to extend the descriptor to be able to use it in object detection tasks.

5. REFERENCES

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, C. Bray, *"Visual categorization with bags of keypoints"*, ECCV Workshop on Statistical Learning for Computer Vision, 2004.

[2] F. Moosman, E. Nowak, and F. Jurie, *"Randomized clustering forests for image classification"*, IEEE Transactions on PAMI, vol. 30, pp. 1632-1646, 2008.

[3] F. Jurie and B. Triggs, *"Creating efficient codebooks for visual recognition"*, ICCV, 2005.

[4] D. G. Lowe, *"Distinctive image features from scale-invariant keypoints"*, International Journal of Computer Vision, vol. 60, 2004.

[5] E. Nowak, F. Jurie, and B. Triggs, *"Sampling strategies for bag-of-features image classification"*, ECCV, 2006.

[6] T. Leung, J. Malik, *"Representing and recognizing the visual appearance of materials using three-dimensional textons"*, International Journal of Computer Vision, vol. 43, pp. 29-44, 2001.

[7] K. Barnard, P. Duygulu, R. Guru, P. Gabbur, and D. Forsyth, *"The effects of segmentation and feature choice in a translation model of object recognition"*, CVPR, 2003.

[8] P. Koniusz and K. Mikolajczyk, *"On a quest for image descriptors based on unsupervised segmentation maps"*, International Conference on Pattern Recognition, 2010.

[9] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, *"Discovering object categories in image collections*", CVPR, 2005.

[10] A. A. Ursani, K. Kpalma, and J. Ronsin, *"Texture features based on Fourier transform and Gabor filters: an empirical comparison"*, International Conference on Machine Vision, 2007.

[11] M. Heikkila, M.Pietikainen, and C. Schmid, *"Description of interest regions with local binary patterns"*, Pattern Recognition, vol. 42, pp. 425-436, 2009.

[12] F. Zhou, J.-F. Feng, and Q.-Y. Shi, *"Texture feature based on local Fourier transform"*, International Conference on Image Processing, 2001.

[13] A. A. Ursani, K. Kpalma, and J. Ronsin, *"Texture features based on local Fourier histogram: self-compensation against rotation"*, Journal of Electronic Imaging, 2008.

[14] T. Ahonen, J. Matas, C. He, and M. Pietikainen, *"Rotation invariant image description with local binary pattern histogram Fourier features"*, SCIA '09 Proceedings of the 16th Scandinavian Conference on Image Analysis, 2009.

[15] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, *"Surf: Speeded up robust features"*, Computer Vision and Image Understanding, vol. 110, pp. 346-359, 2008.

[16] S. Belongie, J. Malik, and J. Puzicha, *"Shape matching and object recognition using shape contexts"*, IEEE Transactions on PAMI, vol. 24, pp. 509-521, 2002.

[17] N. Dalal and B. Triggs, *"Histograms of oriented gradients for human detection"*, CVPR, 2005.

[18] X. Tan and B. Triggs, *"Enhanced local texture feature sets for face recognition under difficult lighting conditions"*, IEEE Transactions on Image Processing, vol. 19, pp. 1635-1650, 2010.

[19] H. Harzallah, F. Jurie, and C. Schmid, *"Combining efficient object localization and image classification"*, ICCV, 2009.

[20] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, *"Learning object categories from Google's image search"*, ICCV, 2005.

[21] D. Nister and H. Stewenius, *"Scalable recognition with a vocabulary tree"*, CVPR, 2006.

[22] Available at *http://www.vision.caltech.edu/html-files/archive.html.*

[23] Available at *http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php*