Anadolu Üniversitesi Bilim ve Teknoloji Dergisi A- Uygulamalı Bilimler ve Mühendislik
Anadolu University Journal of Science and Technology A- Applied Sciences and Engineering

2017 - Volume: 18 Number: 1
Page: 247 - 261
DOI: 10.18038/aubtda.300419
Received: 21 October 2016 Revised: 08 March 2017 Accepted: 08 March 2017

THE FOURIER TRANSFORM BASED DESCRIPTOR FOR VISUAL OBJECT CLASSIFICATION

Hakan ÇEVİKALP [1, *], Zuhal KURT [2]

[1] Department of Electrical and Electronics Engineering, Faculty of Engineering, Eskişehir Osmangazi University, 26480 Eskişehir, Turkey
[2] Department of Mathematics and Computer Sciences, Faculty of Science and Arts, Eskişehir Osmangazi University, 26480 Eskişehir, Turkey

ABSTRACT

Image representation models such as bag of words (BoW) or Fisher vector (FV), which are built depending on encoding local features, are commonly used in visual object classification tasks. In this context, the local patches sampled from images are represented by different texture and shape descriptors such as Scale Invariant Feature Transform (SIFT), Local Binary Pattern (LBP), Speed up Robust Features (SURF), etc. In this study, we propose a new descriptor using weighted histograms of phase angles of local 2-D discrete Fourier transform (FT). We make comparisons with the classification accuracies achieved by using the proposed descriptor to the ones achieved by other commonly used descriptors on Caltech-101, Coil-100 and PASCAL VOC 2007 databases. Experimental results show that our proposed descriptor yields good classification accuracies (the best results on Coil-100, and the second best result on Caltech-101 and PASCAL VOC 2007 datasets) indicating that FT based local descriptors obtain important properties of images that are valuable for visual object classification. By combining image representations resulting from FT descriptor with the representations resulting from other descriptors, accuracies even get better demonstrating that tested descriptors encode different supplementary knowledge.

Keywords: Visual object classification, Fourier transform, Descriptor, Bag of words, Fisher vector

1. INTRODUCTION

Visual object classification can be described as the task of labeling an image with one or multiple labels corresponding to the existence of a visual object class. This task is treated as a standard supervised learning problem in which classifiers that predict the labels of images are learned by using a training set of labeled images. It is a significantly important task, and a useful visual object categorization system may considerably improve the productivity of other essential computer vision applications such as image retrieval and object localization. The main problem in visual object classification stems from the large intra-class differences and the variations of viewpoints in all of the object categories. Other difficulties of visual object classification task are changes in illumination and scale, complex backgrounds, occlusion and existence of noise in the images.

Most of the state of art research on visual object classification has been devoted to the design of successful image representations. A very successful representation paradigm is to extract a set of local patch descriptors from an image and encode them into a high-dimensional visual feature vector. Bag of words (BoW) model was applied in most of the state-of-art object classification techniques for this purpose. BoW was firstly used for text classification and it was applied to the visual object classification by Csurka et al. [1]. After this, these kinds of representations have been extensively used for both visual object classification and localization [2,3,4,5,6]. The BoW model illustrated in Figure 1 treats each image as a regular collection of local patches. Hence, one has to sample a set of patches from the image, extract descriptor vectors from each patch, quantize the descriptors, and accumulate histograms of patch appearances using this quantization to obtain the final image representation.

[*] Corresponding Author: hcevikalp@ogu.edu.tr

There are essentially three important design issues in BoW model: (1) to extract patches from image, (2) to select descriptor type, and (3) to quantize the final descriptors. Patches are typically extracted from the image at many various positions and scales by using different sampling techniques including dense sampling [3,7], random sampling [8], sampling based on the output of some sort of salient region detector [1,5], or sampling built on the output of segmentation techniques [9,10]. After patches are chosen, they are described by using different descriptors. The final descriptors are then clustered to acquire visual words (also called visual words dictionary or vocabulary). Several clustering algorithms including k-means clustering [1,4], mean shift [2], hierarchical clustering [6], randomized trees [11] or other clustering methods were used for this purpose. Finally, the image feature histogram is obtained by averaging the presence of counts, that is commonly referred to as the average pooling. More recent methods adopted better coding techniques by using the soft-assignment of the local descriptors. There are various methods for this, e.g., to use sparse coding [12,13], local coordinate coding [14], or locally-constrained linear coding [15], which enforce a descriptor to be assigned to several number of visual words in the dictionary. Using soft-assignment techniques in BoW models also gave rise to new aggregation techniques such as max-pooling [12,16] or geometric l_p-norm pooling [17] beside the most common average pooling strategy.

In contrast to the BoW modeling which uses a simple first-order statistic of visual word occurrence, more recent image representation models have utilized higher-order statistics. Among these, the Fisher Vector (FV) representation of Perronnin and Dance [18] encodes the deviation of a set of local descriptors from a Gaussian mixture model (GMM) obtained from training images. Despite its inherent limitation of returning dense representations, FV models significantly outperformed BoW models when they are used with linear SVMs [18,19]. In a similar manner, the Super Vector (SV) [16] representations of local descriptors and the Vector of Locally Aggregated Descriptors (VLAD) [20] also used higher-order statistics and successful accuracies have been reported for both retrieval and classification tasks owing to the fact that these methods perform an explicit embedding of local descriptors in very high-dimensional discriminative spaces, and consequently linear classifiers work well on these.

Once the image feature vectors are obtained, these are fed to classifiers that learn to predict the image labels. Various classification methods including Nearest Neighbor, Naïve Bayes, Support Vector Machines (SVM) or other kernel methods were used for visual object classification for this purpose, but currently linear/kernelized SVMs dominate the field.

In this paper, we concentrate on the descriptors, which are used for describing image patches and we introduce a new descriptor using weighted histograms of phase angles of Fourier Transform to be used in BoW or FV models for visual object classification. A preliminary version of this paper has appeared in [21]. This paper extends our previous work with (1) a more elaborated analysis of the current related work on image representation, (2) a more detailed description of the proposed descriptor and modifications on the descriptor by changing the histogram construction, (3) more experiments on larger image classification data sets, and (4) testing different design methodologies such as the sizes of cells, the sizes of the phase angle bins, and different normalizations of histograms on the classification performance.

1.1 Related Work

In this paper we are interested in the descriptors which are used for representing image patches. We want the descriptors to be invariant to the variations due to the image transformations, illuminating variations and occlusions, because these properties are irrelevant to the image categorization. Also, descriptors must include discriminative information to separate the object categories. Among all descriptors, histogram based descriptors became very popular since they perform well and they are easily computed. Most of the histogram based descriptors use oriented image gradients, containing

SIFT [22], SURF [23], Histograms of Oriented Gradients (HOG) [24], Generalized Shape Context [25], Multi-support Region Order-Based Gradient Histogram (MRHOG) [26]. Some of them use color histograms [27, 28] whereas some use local patterns of qualitative gray level differences such as Local Binary Patterns (LBP) [29] and Local Ternary Patterns (LTP) [30]. In contrast, some descriptors may use filter outputs including Gabor descriptors [31] or Zernike-Moment Phase descriptor of Chen and Sun [32]. Empirically, the best descriptor set depends on the implementation and new ones are being developed constantly. Here we briefly describe the most popular SIFT, SURF, LBP and LTP descriptors.

SIFT: SIFT descriptor is constructed by accumulating gradient norms around some key points and the final descriptor is a 128-dimensional vector. To this end, we first calculate the gradient magnitude and orientation at each image sample point in a region around the key point. Then, these samples are weighted and accumulated into orientation histograms. Finally, the descriptors are built by combining histograms of 4×4 cells with eight orientation bins in each [22].

SURF: The SURF descriptor represents a distribution of Haar-wavelet responses within some interest point neighborhood. The interest region is first split up into 4×4 square sub-regions. The Haar wavelet responses in horizontal and vertical directions are computed for each sub-region. These responses are weighted with a Gaussian kernel function to make the descriptor more robust to geometrical deformations. Then, wavelet responses are summed and these values together with their absolute values are combined to produce a descriptor vector with dimension 64 [23].

LBP: The LBP descriptor defines a small neighborhood around every pixel, thresholds the pixels of the neighborhood at the value of the central pixel and uses the resulting binary valued image patch as a local image descriptor. Initially, it was defined for 3×3 neighborhoods giving 8 bit integer LBP codes depending on the 8 neighboring pixels. More precisely, the LBP operator takes the form

$$LBP(x_c, y_c) = \sum_{n=0}^{7} 2^n f(I_n - I_c), \qquad (1)$$

where in this case n runs over the 8 neighbors of the central pixel c, I_c and I_n are the gray level values at c and n, and $f(u)$ is 1 if $u \geq 0$ and 0 otherwise. LBP descriptors are robust to most lighting changes, but they are sensitive to noise in near-uniform regions [29].

LTP: LBP descriptors are discriminative features for texture classification and they are resistant to illumination changes in a sense that they are invariant to monotonic gray level transformations. On the other hand, since they threshold at exactly the value of the central pixel I_c they tend to be sensitive to noise, particularly in near-uniform image regions. The LTP descriptor extends LBP to 3-valued descriptors, *Local Ternary Patterns* (LTP), in which gray-levels in an area of width $\mp t$ around I_c are quantized to zero, ones above this are quantized to $+1$ and ones below it to -1, *i.e.* the indicator $f(u)$ is replaced with a 3-valued function and the binary LBP descriptors is replaced by a ternary LTP descriptors as

$$f(u, I_c, t) = \begin{cases} 1, & u \geq I_c + t \\ 0, & |u - I_c| < t, \\ -1, & u \leq I_c - t \end{cases} \qquad (2)$$

where t is a user specified threshold. As a result, LTP descriptors are more resistant to noise, but no longer strictly invariant to gray level transformations [29].

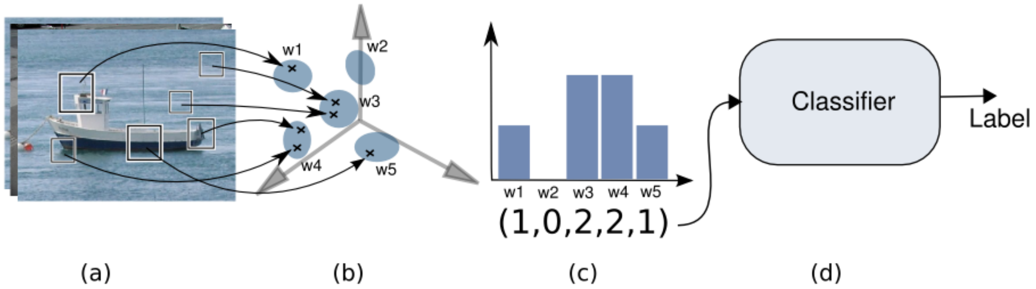

Figure 1. Visual object classification using bag of words model: a) Sampling of the patches and extraction of the descriptor values b) Clustering and construction of the visual vocabulary c) Building image histograms d) Classification of images based on histograms.

2. METHOD

2.1. Motivation

In order to motivate the proposed descriptor, we start with a widely known example shown in Figure 2 illustrating the significance of the phase information of the 2-D FT for image description. When we draw the magnitude and phase information of an image, the phase image looks similar to some sort of noise that does not contain any significant information related the original image. However, the reconstructing image by inverse FT using only magnitude information produces a largely blank image that does not include any important information for classification. Conversely, when we use the phase information for inverse FT to reconstruct image, the final image is more similar to the input image as shown in the figure. This undoubtedly demonstrates that the phase information includes more characteristic knowledge useful for image representation compared to the magnitude of FT. In fact, recent studies also report the similar findings and show the importance of phase information in other image representations such as image gradients [24] or Zernike moments [32]. Therefore, in the proposed descriptor, we put more emphasize on the phase of FT during encoding FT values.

Using FT as descriptor is not new [31,33,34,35,36,37,38]. However, the commonly known FT based descriptors use magnitude of FT to obtain rotational invariance property, or they use the FT with other image features (such as gradients or LBP features) together. The local 1-D FT histograms of gradient images were used for texture recognition in [31, 34]. However, they disregard the phase information and use only the magnitude to acquire image representations that are invariant to image rotations (in contrast with magnitudes of FT, phases of FT are sensitive to rotational transformations). In [33], the authors extract 1-D FT of 3 × 3 neighborhoods, and the image is described by concatenating the histograms of magnitudes and phases. Recently, Ahonen et al. [35] propose a rotation-invariant descriptor by using magnitudes of FT applied to the LBP descriptors extracted from images. In [36], object boundaries are obtained by using edge detection first and then each object boundary is transformed into a 1-D signal and finally FT of this signal is used to describe the objects. [36,37] both use Local Phase quantization, in which 2D-FT is aplied to difference of neighboring image pixel values only on 4 selected frequencies.

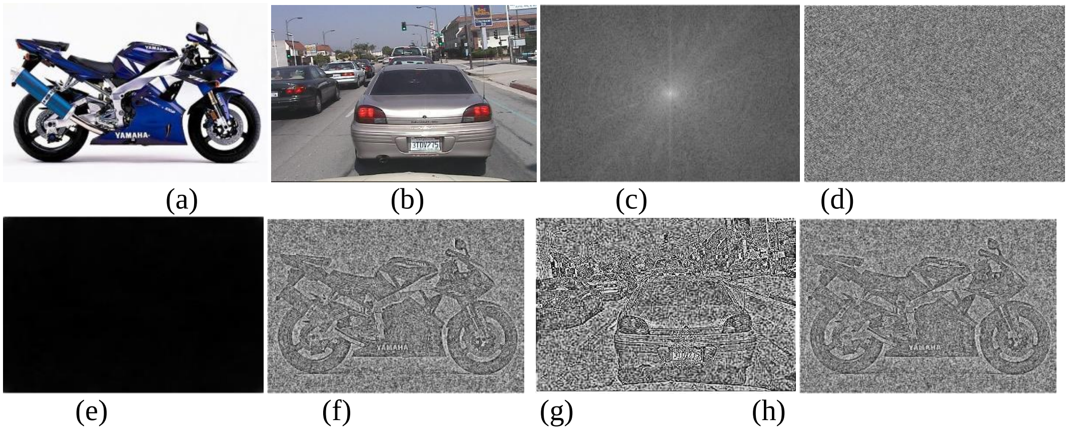

Figure 2. (a) motorbike image; (b) car image; (c) magnitude of the FT of image (a); (d) phase of the FT of image (a); (e) the resulting image by inverse Fourier transform applying just magnitude of (a) with random phase information; (f) the resulting image by inverse FT applying phase of (a) with random magnitude; (g) the resulting image by inverse FT using phase of (b) with random magnitude; (h) the resulting image by inverse FT using phase of (a) with magnitude of (b).

In this study we also introduce a descriptor that utilizes the discrete FT of local patches as shown in Figure 3. In contrast with the recent studies defined above, we use 2-D discrete FT of local patches and apply it directly to the image gray-level values (not to gradients or LBP features) of patches extracted from the images. The construction of our histogram also diverges from the other methods in a way that we use histograms of phase angles weighted by magnitude values as defined below.

2.2. D Fourier Transform Based Descriptor

In image representations using BoW or FV models, patches are sampled from images at many various positions and scales by applying different sampling methods. Then, the fixed-size features are extracted from the patches by applying different descriptors such as SIFT, SURF, LBP, LTP, etc. We use two-dimensional (2-D) discrete Fourier transform of image patches to compute the proposed descriptor.

The discrete Fourier transform of an image region $I(x,y)$ of size $M \times N$ is given by the equation

$$F(u,v) = \sum_{x=0}^{M-1} \sum_{y=0}^{N-1} I(x,y)\, e^{-j2\pi(\frac{ux}{M}+\frac{vy}{N})},\ u = 0,\ldots,M-1, v = 0,\ldots,N-1. \quad (3)$$

The original image $I(x,y)$ can be obtained by using the inverse Fourier transform, given by the expression

$$I(x,y) = \frac{1}{MN} \sum_{u=0}^{M-1} \sum_{v=0}^{N-1} F(u,v)\, e^{j2\pi(\frac{ux}{M}+\frac{vy}{N})},\ x = 0,\ldots,M-1, y = 0,\ldots,N-1. \quad (4)$$

Equations (3) and (4) comprise the 2-D discrete Fourier transform (DFT) pair. The variables u and v are referred to as the frequency variables. Fourier spectrum, phase angle, and power spectrum of an image are defined respectively as follows:

$$|F(u,v)| = [Re^2\{F(u,v)\} + Im^2\{F(u,v)\}]^{1/2},$$
$$\phi(u,v) = tan^{-1}[Im\{F(u,v)\}/Re\{F(u,v)\}],$$
$$P(u,v) = |F(u,v)|^2,$$

where $Re\{F(u,v)\}$ and $Im\{F(u,v)\}$ are the real and imaginary parts of $F(u,v)$, respectively.

The value of the 2-D FT transform at the origin becomes $F(0,0) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} I(x,y)$, which gives the sum of values of $I(x,y)$. In other words, if $I(x,y)$ is an image region, the value of the Fourier transform at the origin is equal to the sum of all pixel values in that region. Because both frequencies are zero at the origin, $F(0,0)$ is also called the dc-component of the spectrum.

In our proposed method, we partition the patch region into $n \times n$ non-overlapping sub-regions (cells) as illustrated in Figure 3. After that, the magnitudes and phase angles of 2-D FT of each cell are found. The descriptor is then created by accumulating weighted votes to phase angle bins. The phase angle bins are evenly spaced into b (we tried 9 and 19 bins) intervals over $[0,2\pi]$. The vote may be a transformation of the FT magnitude, either the magnitude itself, its square, its square root, or a clipped form of the magnitude demonstrating soft existence/absence of the phase angle bin. In our study, the magnitude of FT is used to vote the bins followed by a normalization. The size of the resulting descriptor vector is fixed for any patch size and it is equal to $d = n^2 \times b$.

As we mentioned above, the dc-component term at the origin is just equal to the sum of the pixel values and its magnitude is the highest compared to the other Fourier spectrum values. Its magnitude always contributes to the first bin of the phase angle histogram. As a result, the first histogram bin typically dominates other bin values, and this may negatively affect the classification performance. Therefore, we ignore the contribution of dc-component term during computing values of histogram bins. The number of pixels in each cell may vary based on the sizes of the patches, thus cell histograms must be normalized. To this end, we tried different normalization types as given below;

- L_1: $V_k = V_k/(\sum_{i=1}^{b} V_i)$ (histograms are normalized so that the sum is equal to 1)
- $L_1 - sqrt$: $V_k = \sqrt{V_k/(\sum_{i=1}^{b} V_i)}$ (histograms are normalized so that the sum is equal to 1, and then they are square rooted.)
- L_2 : $V_k = V_k / \sqrt{(\sum_{i=1}^{b} V_i^2)}$, (this nonlinear normalization emphasizes cells whose counts are distributed into many bins.)
- Log_2: $V_k = log_2(V_k + 1)$ (a simple log_2 compression of the histogram bins.)

Here V_k represents *k*-th bin of the histogram.

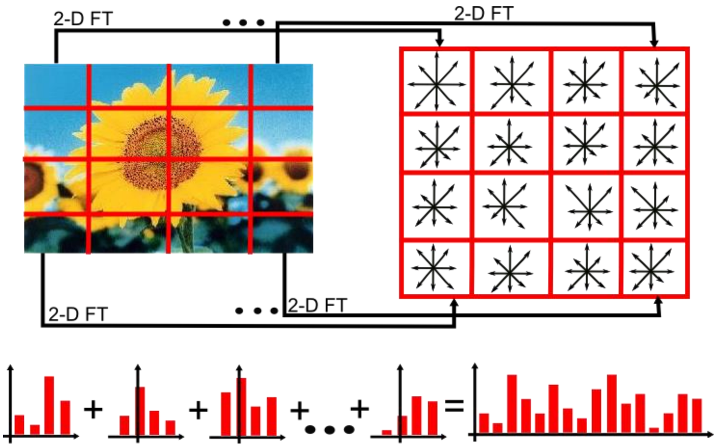

Figure 3. Computation of FT descriptor: Firstly, image patch is split into 4×4 cells in this example. 2D FT is extracted from every cell, and the histogram of every cell is constituted by accumulating weighted votes to phase angle bins followed by normalization. The resulting descriptor is constructed by concatenating the histograms of all cells.

We use these normalization types for different number of cell sizes ranging from 4×4 to 7×7. The phase angle bins are evenly spaced into 9 or 19 intervals over $[0,2\pi]$. The best classification accuracies are obtained by using 5×5 cells, 9 bins and log_2 normalization. Therefore the default size of the resulting descriptor vector is fixed for any patch, and it is equal to $d = 225(5^2 \times 9)$.

It should be noted that the final descriptor is not rotational invariant similar to the most commonly used descriptors such as SIFT, HOG, and LBP. We believe that the introduced descriptor encodes more important information for image representation in comparison with other commonly known descriptors. When we consider the SIFT descriptor for instance, it uses weighted histograms of gradient angles. This knowledge is already included in high-frequency components of FT. In addition to this, low-frequency components of the proposed FT descriptor capture information related to the general appearance of patches. But, SIFT does not contain this infromation.

3. EXPERIMENTS

The proposed descriptor is tested on three challenging visual object classification databases: the Coil-100, Caltech-101, and PASCAL VOC 2007. We also conducted experiments to determine the best design parameters, such as bin size, cell size and cell normalization type. We used BoW or FV models to represent images. The proposed FT descriptor is compared with the most commonly used descriptors, such as SIFT, LBP, LTP, and SURF. Moreover, we conducted some tests to verify that whether the combining the image representations resulting from FT descriptor with the representations resulting from other descriptors achieve better accuracies rather than using these descriptors alone. The linear and nonlinear Support Vector machine (SVM) classifier were used for classification in experiments, and either one-against-rest or one-against-one techniques is utilized to extend the binary SVM classifier to multi-class classification. We just give the accuracy for the one that performs best.

3.1. Experiments for Setting The Best Design Parameters

There are design parameters of the proposed FT based descriptor that seriously affect the overall classification performance. These are the number of phase angle bins, the number of cells, and the type of the cell normalization. We conducted some experiments to estimate the best values of these parameters. As we mentioned at the Introduction section, the patches can be chosen from the images at numerous positions and scales densely or based on the output of some sort of salient region detector. Many studies [2,8] show that dense sampling significantly outperforms other sampling types in the context of object classification, so we use dense sampling in our experiments. To determine the best parameters, we used the images selected from 10 categories of Caltech-101 database (available at http://www.vision.caltech.edu/ImageDatasets/ Caltech101/). To set bin size, we compared the descriptors using 9 and 19 phase angle bins. For cell sizes, we tried 4×4, 5×5, 6×6, and 7×7 cells. Finally, for normalization we compared the descriptors using L1, L2, L1-sqrt, and Log$_2$ normalizations. The classification accuracies for different design parameters are given in Table 1. Linear SVM is used as classifier. In general, Log$_2$ normalization significantly outperforms other normalization techniques and the best accuracy is obtained by FT descriptor using 5×5 cells, 9 bins and Log$_2$ normalization. Therefore, we used these parameters for the rest of the experiments.

Table 1. Classification rates (%) for different design parameters

Bin Size	**9 bins**				**19 bins**			
Cell Size	**4×4**	**5×5**	**6×6**	**7×7**	**4×4**	**5×5**	**6×6**	**7×7**
L1	55.9	53.7	63.4	51.2	57.4	56.7	56.2	45.1
L2	57.1	49.1	60.9	55.6	52.1	49.6	55.8	49.5
L1-sqrt	57.5	53.2	57.5	52.3	50.5	49.0	51.7	50.0
Log$_2$	68.9	71.9	64.4	64.2	67.1	71.4	67.4	64.2

3.2. Experiments on the Coil-100 Database

The Coil 100 dataset contains images of hundred different object classes (available at http://www.cs.columbia.edu/CAVE/software/ softlib/coil-100.php). Each class includes 72 images for every object category and the size of images is 128×128 pixels. Images include several poses of the same objects under same illuminating conditions and background. We selected 40 object categories (illustrated in Figure 4) from the dataset for our experiments.

The background of images in this database is uniform, therefore the object can be easily separated from the background. Firstly, all images were converted to gray-scale and BoW model is used to represent images. To this end, we firstly extracted patches densely from the images. For each chosen patch, we computed the proposed FT and other tested descriptors. Then, we used k-means clustering method to determine the visual words of BoW model. We set the dimensionality of the visual vocabulary to 1,000. To construct feature histograms of test images, the extracted descriptors are compared to the visual words based on their Euclidean distances and the final histogram of "visual word" occurrences is built based on average pooling (i.e., we count the number of occurrences for every visual word and divide by the total number of descriptors). Therefore, the size of the image feature vectors is also 1000.

The classification accuracies are given in Table 2. Linear SVM is used as classifier as before. The 5-fold cross-validation method is used for computing the classification accuracies. Our proposed FT descriptor achieved the best accuracy for all cases and it is followed by SIFT. SURF descriptor is the worst performing descriptor. The classification accuracies typically get better after combination of the image representations by using different descriptors (we just concatenate image feature histograms to form an extended feature vector without any normalization). More precisely, the best classification accuracy is achieved by combining the representations resulting from the best performing 3 descriptors (FT, SIFT, and LTP), which indicates that the tested descriptors capture supplementary knowledge.

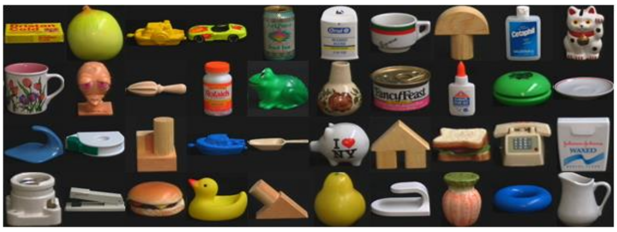

Figure 4. Forty object classes is selected from Coil-100 dataset

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

Descriptors	Classification Rates
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+SIFT	92.83 ±1.6
FT+SURF	93.18 ±2.3
FT+LBP	92.54 ±1.4
FT+LTP	93.37 ±2.3
FT+SIFT+LTP	**95.09** ±1.6

3.3. E xperiments on the Caltech-101 Database

Caltech-101 is an image database that contains images belonging to 101 object categories. There is also an additional background class, making the total number of classes 102. There are 40 - 800 images each category, and a lot of categories include approximately 50 images. Some examples are given in Figure 5. The images of object categories have background clutter and significant intra-class and scale variability, so we again use the "bag of features" representation using the spatial pyramid approach of Lazebnik et al. [39]. We set the visual vocabulary size to 600, which yields image feature histograms of size 12600. We use the same experimental setup as in [40,41]. We randomly pick 30 images from every class and divide them into 15 for training part and 15 for test part. Then, we invert the role of training and test part. The final classification rates are the averages.

We used both linear and nonlinear (kernelized) SVMs to assess the classification performance. The Gaussian kernel, $k(x,y)=\exp(-\|x-y\|^2/\sigma)$ $k(x,y)=\exp(-\|x-y\|^2/\sigma)$ is used for nonlinear SVM classifier, as a kernel in all experiments. The results are given in Table 3. It should be noted that lower accuracies are reported for SIFT descriptors compared to our results for linear SVMs in the literature. This is due to the normalization of histograms. We standardized the image feature histograms so that standard deviation of each feature histogram bin is equal to 1. This yielded significant improvements (up to 10%) in linear classification accuracies. This also explains that why kernels using Chi-squared distances perform much better compared to the Euclidean distance based linear kernels. The best classification accuracies are obtained by using SIFT descriptors for both linear and nonlinear cases. The proposed FT based descriptor yield to the second best result after SIFTs. LBP descriptor is the worst performing descriptor. Moreover, we combined the feature extracting techniques with using the best performing three descriptors (SIFT, FT and SURF) to see if any further improvement can be achieved. Approximately 2.5-3% improvement is obtained over the best performing SIFT descriptor when 3 different descriptors are used together.

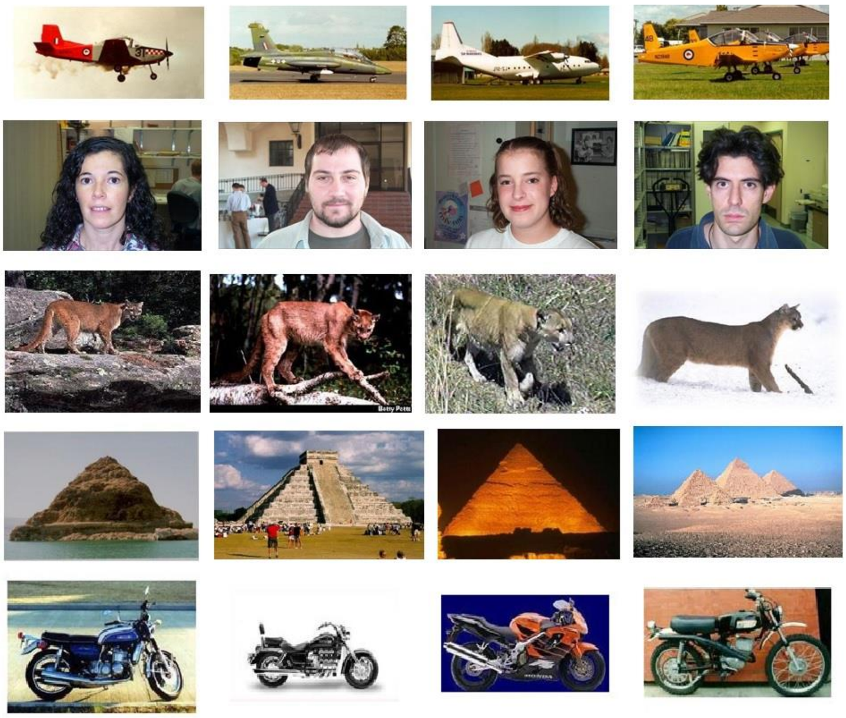

Figure 5. Some image samples belonging to 5 object classes (airplanes, faces, cougar_body, pyramid and motorbikes) in Caltech-101. Images in each class have significant intra-class and scale variability, which make the classification problem more difficult.

Table 3. Classification rates (%) on the Caltech 101 dataset

Descriptors	**Linear Kernel**	**Gaussian Kernel**
FT	48.33	51.05
SIFT	60.37	62.12
SURF	42.84	44.58
LBP	33.47	36.01
LTP	42.78	43.89
FT+SIFT+SURF	**63.37**	**64.51**

We also visually compare the image descriptor histograms on some images chosen from this dataset in Fig. 6. These are simple 600-dimensional BoW feature histograms extracted without using spatial pyramid approach. As can be seen in the figure, the feature histograms obtained using FTT for the objects in the same category are very similar whereas histograms for different objects are quite different. Histograms obtained using SIFT also have this property but the LBP histograms fail in the sense that the histograms of objects in the same category are quite different for LBP descriptors.

3.4. Experiments on the PASCAL VOC 2007 Database

The PASCAL VOC 2007 database includes images of daily scenes with annotations for all entirely or partial samples of 20 common visual object classes: aeroplane, bicycle, bird, boat, bottle, bus, cat, car, chair, cow, dog, dining table, horse, motorbike, person, potted plant, sofa, sheep, train, and tv-monitor. Its training/validation subset contains 5011 images with 12608 annotated instances, and its test set contains 4952 images with 12032 annotated instances. We use the standard protocol for training and test, i.e., we train and set the parameters of the classifiers by using the provided “train” and “val” sets, and test the classifiers on the “test” set. The accuracies are computed by applying the standard measure on this database, that is the Average Precision (AP) scores obtained from Precision-Recall curves.

We used FV representation for this dataset since this model significantly outperforms BoW when linear SVM classifier is used. For this purpose, we implemented the same structure as in [19]. More precisely, we extracted roughly 10K descriptors for each image from 24×24 patches on a regular grid each four pixels at 5 scales. The size of the tested descriptors is reduced to 80 by applying Principal Component Analysis (PCA) except for the SURF descriptor, for which the dimensionality is reduced to 60 since the original dimensionality of the descriptor is 64. We used 6×10^6 descriptors to learn the projections of PCA and the components of Gaussian mixture model (GMM). We used 256 components for GMM model. Also as in [19], we used the same spatial pyramid structure and extracted 4 FVs per image: one FV for the entire image and one in three horizontal stripes according to the top, middle, and bottom regions of the image. We used only linear SVMs since the dimensionality of the final FV representation is too high (the final dimensionality is around 164K).

The accuracies for the tested descriptors are given in Table 4. We also give the best result of official winner of the original VOC 2007 challenge (available at: http://pascallin.ecs.soton.ac.uk/ challenges/ VOC/voc2007/results/index.shtml) for each class for comparison. When single descriptors are used, SIFT descriptor significantly outperforms others yielding an average AP score of 58.3%. It is followed by the proposed FT descriptor and LTP. SURF is the worst performing descriptor. When we combine the best performing three descriptors SIFT, FT and LTP, the accuracies improve over the best performing SIFT descriptor. In general, we obtain better results than the best results reported in the official PASCAL VOC contest for 11 object categories, whereas our results are slightly behind for 9 object categories.

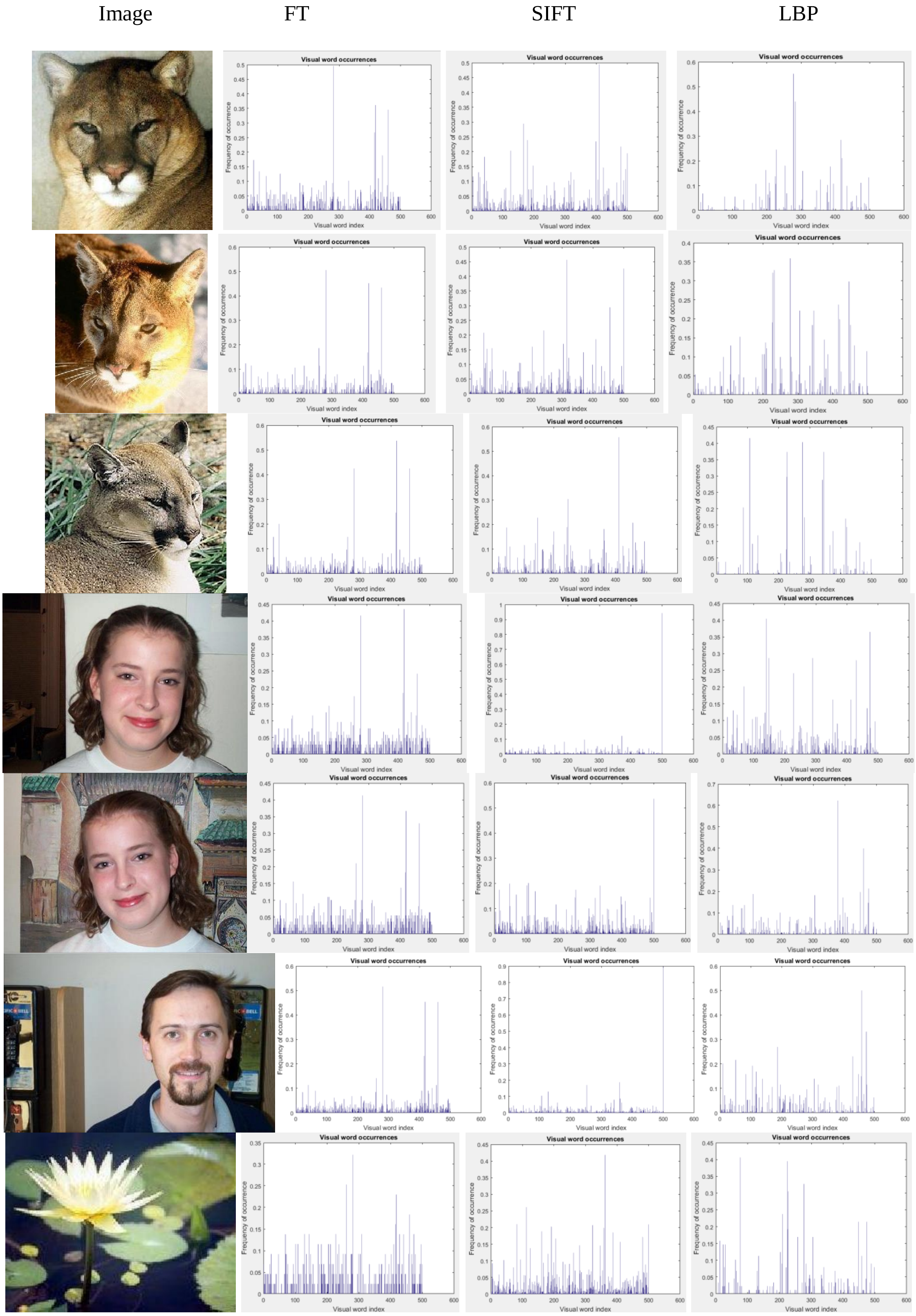

Figure 6. Visual comparision of feature histograms for different descriptors

Table 4. Classification Accuracies (AP Scores) for PASCAL VOC 2007 dataset

Descriptors	Aeroplane	Bicycle	Bird	Boat	Bottle	Bus	Car	Cat	Chair	Cow	Diningtable	Dog	Horse	Motorbike	Person	Pottedplant	Sheep	Sofa	Train	Tvmonitor
VOC 2007 Winner	77.5	63.6	**56.1**	**71.9**	**33.1**	60.6	78.0	**58.8**	**53.5**	42.6	54.9	**45.8**	77.5	64.0	**85.9**	**36.3**	44.7	50.9	79.2	**53.2**
FT	73.2	49.1	39.1	61.7	22.9	48.0	67.6	43.5	45.9	33.8	42.7	35.9	69.1	50.4	77.7	22.5	35.2	41.1	69.8	35.8
SIFT	77.6	65.2	51.7	67.4	29.9	65.3	77.2	57.0	51.1	**45.0**	53.5	40.4	78.6	67.7	82.0	28.6	47.6	49.4	78.7	51.9
SURF	60.9	39.7	32.8	37.9	10.9	33.0	60.7	29.1	34.8	20.9	27.3	19.8	62.1	43.5	67.7	11.4	18.7	37.4	47.8	29.7
LBP	69.6	45.4	39.5	54.4	22.6	47.2	66.7	43.7	42.9	26.3	38.9	35.7	70.6	55.1	75.4	21.8	26.5	38.4	66.6	38.9
LTP	68.1	50.0	37.6	63.4	21.4	46.5	68.1	47.2	45.0	36.1	43.6	32.8	69.2	49.1	76.2	24.3	33.9	38.6	71.9	37.9
FT+SIFT	77.8	65.2	53.8	69.6	29.6	65.0	77.1	55.6	50.8	45.3	54.4	40.2	**78.9**	**68.6**	83.2	28.3	46.7	50.6	78.6	50.8
FT+SURF	70.1	51.5	46.2	55.3	21.3	49.5	70.0	44.7	42.4	26.9	39.7	29.8	70.5	56.1	75.8	20.8	31.9	44.4	68.8	38.8
FT+LBP	75.8	53.9	45.1	66.5	20.6	53.5	71.6	49.7	47.1	32.9	44.0	38.8	73.0	58.5	79.9	24.8	35.0	48.7	74.6	41.9
FT+LTP	72.0	52.6	44.0	64.5	24.5	52.7	71.1	51.2	46.5	36.9	45.0	37.5	72.5	56.5	78.8	24.9	35.9	43.2	72.1	39.1
FT+SIFT+LTP	**78.2**	**66.3**	55.1	71.1	29.5	**66.4**	**78.1**	57.3	52.0	44.6	**56.7**	42.3	78.6	**68.6**	83.3	28.5	**47.9**	**51.5**	**80.2**	51.8

4. CONCLUSION

We have proposed a new descriptor using 2-D discrete FT of an image patch for categorization of different object classes to be used with BoW and FV image representation models. In the proposed descriptor, we focus on the phase information of FT that is disregarded in the most of the state of art descriptor methods to gain rotational invariance. We build our descriptor by accumulating weighted histograms of phase angles of FT of local patches. We tested different cell sizes, bin sizes, and different normalizations to estimate the best design parameters that work best for the proposed descriptor. Although we obtained the best classification accuracy with our FT descriptor for small Coil-40 dataset including more image samples per-class, our proposed descriptor yielded the second best accuracy on more challenging Caltech 101 and PASCAL VOC 2007 datasets. However, combining image representations obtained by the proposed descriptor and other tested descriptors improved the results over the best performing single descriptor, which shows that our proposed descriptor encodes complementary information for classification of images.

REFERENCES

[1] Csurka G, Dance CR, Fan L, Willamowski J, Bray C. Visual categorization with bags of keypoints. In: 8th European Conference on Computer Vision (ECCV) ; 11-14 May 2004; Prague, Czech Republic: Springer. pp. 59-74.

[2] Jurie F, Triggs B. Creating efficient codebooks for visual recognition. In: 10th IEEE International Conference on Computer Vision (ICCV 2005); 17-20 October 2005; Beijing, China: IEEE. pp. 1103-1110.

[3] Sivic J, Russell BC, Efros AA, Zisserman A, Freeman WT. Discovering object categories in image collections. In: 10th IEEE International Conference on Computer Vision (ICCV); 17-20 October 2005; Beijing, China: IEEE. pp.370-377.

[4] Harzallah H, Jurie F, Schmid C. Combining efficient object localization and image classification. In: IEEE International Conference on Computer Vision (ICCV); 29Sept. -2 Oct. 2009; Kyoto, Japan: IEEE. pp.237-244.

[5] Fergus R, Fei-Fei L, Perona P, Zisserman A. Learning object categories from Google's image search. In: IEEE International Conference on Computer Vision (ICCV); 17-20 October 2005; Beijing, China: IEEE. pp. 1816 - 1823.

[6] Nister D, Stewenius H. Scalable recognition with a vocabulary tree. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 17-22 June 2006; New York, NY, USA: IEEE. pp. 2161-2168.

[7] Leung T, Malik J. Representing and recognizing the visual appearance of materials using three-dimensional textons. International Journal of Computer Vision 2001; 43: 29-44.

[8] Nowak E, Jurie F, Triggs B. Sampling strategies for bag-of-features image classification. In: 9th European Conference on Computer Vision (ECCV); 7-13 May 2006; Graz, Austria: Springer. pp. 490-503.

[9] Barnard K, Duygulu P, Guru R, Gabbur P, Forsyth D. The effects of segmentation and feature choice in a translation model of object recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 16-22 June 2003; Madison, WI, USA: IEEE. pp. 675-682.

[10] Koniusz P, Mikolajczyk K. On a quest for image descriptors based on unsupervised segmentation maps. In: IEEE International Conference on Pattern Recognition (ICPR); 23-26 Aug. 2010; Istanbul, Turkey: IEEE. pp.762-765.

[11] Moosman F, Nowak E, Jurie F. Randomized clustering forests for image classification. IEEE Transactions on PAMI 2008; 30: 1632-1646.

[12] Bourreau YL, Bach F, LeCun Y, Ponce J. Learning mid-level features for recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 13-18 June 2010; San Francisco, CA, USA: IEEE. pp. 2559 - 2566.

[13] Yang J, Yu K, Gong Y, Huang T. S. Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 20-25 June 2009; Miami, Florida, USA: IEEE. pp. 1794 - 1801.

[14] Yu K, Zhang T, Gong Y. Nonlinear learning using local coordinate coding. In: 23rd Annual Conference on Neural Information Processing Systems (NIPS); 7-10 December 2009; Vancouver, British Columbia, Canada: pp. 2223-2231.

[15] Wang J, Yang J, Yu K, Lv F, Huang T, Gong Y. Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 13-18 June 2010; San Francisco, CA, USA: IEEE. pp. 3360-3367.

[16] Zhou Z, Yu K, Zhang T, Huang T. Image classification using super-vector coding of local image descriptors. In: 11th European Conference on Computer Vision (ECCV); 5-11September 2010; Heraklion, Crete, Greece: Springer. pp. 141-154.

[17] Feng J, Ni B, Tian Q, Yan S. Geometric l_p-norm feature pooling for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 21-23 June 2011; Colorado Springs, CO, USA: IEEE. 2609 - 2704.

[18] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 18-23 June 2007; Minneapolis, Minnesota, USA: IEEE. pp. 1 - 8.

[19] Sanchez J, Perronnin F, Mensink T, Verbeek J. Image classification with the Fisher vector: Theory and practice. International Journal of Computer Vision 2013; 105: 222-245.

[20] Jegou H, Perronnin F, Douze M, Sanchez J, Perez P, Schmid C. Aggregating local image descriptors into compact codes. IEEE Transactions on PAMI 2012; 34: 1704-1716.

[21] Cevikalp H, Kurt Z, Onarcan A. O. Return of the King: The Fourier Transform Based Descriptor for Visual Object Classification, IEEE Signal Processing and Communications Applications Conference, 2013.

[22] Lowe DG. Distinctive image features from scale-invariant keypoint. International Journal of Computer Vision 2004; 60: 91-110 .

[23] Bay H, Ess A, Tuytelaars T, Gool LV. Surf: Speeded up robust features. Computer Vision and Image Understanding 2008; 110: 346-359.

[24] Dalal N, Triggs B. Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 20-25 June 2005; San Diego, CA, USA: IEEE. pp. 886-893.

[25] Belongie S, Malik J, Puzicha J. Shape matching and object recognition using shape contexts IEEE Transactions on PAMI 2002; 24: 509-521.

[26] Fan B, Wu F, Hu Z. Rotationally invariant descriptors using intensity order pooling. IEEE Transactions on PAMI 2012; 34: 2031-2045.

[27] Van de Sande K, Gevers T, Snoek C. Evaluating color descriptors for object and scene recognition. IEEE Transactions on PAMI 2010; 32: 1582-1596.

[28] Van de Weijer J, Schmid C. Coloring local feature extraction. In: 9th European Conference on Computer Vision (ECCV); 7-13 May 2006; Graz, Austria: Springer. pp. 334-348.

[29] Heikkila M, Pietikainen M, Schmid C. Description of interest regions with local binary patterns. Pattern Recognition 2009; 42: 425-436.

[30] Tan X, Triggs B. Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Transactions on Image Processing 2010; 19: 1635-1650.

[31] Ursani A. A, Kpalma K, Ronsin J. Texture features based on Fourier transform and Gabor filters: an empirical comparison. In: IEEE International Conference on Machine Vision; 28-29 Dec. 2007; Tokyo, Japan: IEEE. pp.67-72.

[32] Chen Z, Sun S. K. A Zernike moment-phase based descriptor for local image representation and matching. IEEE Transactions on Image Processing 2010; 19: 205-219.

[33] Zhou F, Feng J-F, Shi Q.-Y. Texture feature based on local Fourier transform. In: IEEE International Conference on Image Processing; 7-10 Oct 2001; Thessaloniki, Greece: IEEE. pp. 610 - 613.

[34] Ursani A. A, Kpalma K, Ronsin J. Texture features based on local Fourier histogram: self-compensation against rotation. Journal of Electronic Imaging 2008; 17(3): 030503.

[35] Ahonen T, Matas J, He C, Pietikainen M. Rotation invariant image description with local binary pattern histogram Fourier features. In: SCIA '09 Proceedings of the 16th Scandinavian Conference on Image Analysis; 15-18June 2009; Oslo, Norway: Springer. pp. 61-70.

[36] Bartoloni I, Ciaccia P, Patella M. WARP: Accurate retrieval of shapes using phase of fourier descriptors and time warping distance, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005; vol .27: 142-147.

[37] Ojansivu V, Heikkila J. Blur insensitive texture classification using local phase quantization, Image and Signal Processing, Volume 5099 of the series Lecture Notes in Computer Science, 2009: pp. 236-243.

[38] Ahonen T, Rahtu E, Ojansivu V, Heikkila J. Recongition of blurred faces using local phase quantization. International Conference on Pattern Recognition, 2008.

[39] Lazebnik S, Schmid C, Ponce J. Beyond bag of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 17-22 June 2006; New York, NY, USA: IEEE. pp. 2169 – 2178

[40] Zhang H, Berg AC, Maire M, Malik J. Svmknn: Discriminative nearest neighbor classification for visual recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 17-22 June 2006; New York, NY, USA: IEEE. pp. 2126 - 2136.

[41] Berg AC, Berg TL, Malik J. Shape matching and object recognition using low distortion correspondence. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 20-25 June 2005; San Diego, CA, USA: IEEE. pp. 26 - 33.