Multiclass Object Recognition Using Object-based and Local Image Features Extractors

Menna Maged Kamel
College of Computing and Information Technology
Arab Academy for Science and Technology
Cairo, Egypt
mennakamel@aast.edu

Mohamed Waleed Fakhr
College of Computing and Information Technology
Arab Academy of Science and Technology
Cairo, Egypt
waleedf@aast.edu

H. I. Saleh
Radiation Engineering Department
Atomic Energy Authority
Cairo, Egypt
h_i_saleh@hotmail.com

M. B. Abdelhalim
College of Computing and Information Technology
Arab Academy of Science and Technology
Cairo, Egypt
mbakr@ieee.org

***Abstract*- This paper examines a combined feature extraction method for visual object recognition. The method is based on applying Bag of words (BoW) using the object-based Zernike Moment (ZM) shape descriptor and the Speeded up Robust Features (SURF) local descriptor on the detected objects in an image. Support Vector Machine (SVM) classifier is trained on the extracted features using a one versus all method. The experiments are tested on two benchmark data sets for object recognition like COREL 1000, and Caltech 101. Using this combined feature extraction method, the derived results outperform the other published methods applied on the same databases.**

Keywords— Object Recognition; Speeded up Robust Features; Bag of Words; Zernike moment

I. INTRODUCTION

The field of visual object recognition has been growing in the recent years and became a challenging problem; it didn't reach the human level performance yet. Image recognition and classification is an important research area for manipulating large databases. As the features of an image have strong relationship with its semantic meaning; object recognition systems seeks to recognize the content of images automatically using descriptors to classify the objects in an image. Some approaches rely on the local features of an image and don't capture the global features in it, although, shape is important and gives a powerful clue to identify an object.

Recent approaches, like speeded up robust features (SURF), use local descriptors to extract local information from an image [1]. They tend to detect the local interest areas and represent them in numeric vectors to quantify the image invariant descriptors.

Other studies use complex moments as a global feature extractor. An effective shape descriptor is the Zernike moment (ZM) which is based on a set of orthogonal complex moments. It was first introduced to image analysis by Teague who constructed rotation invariants by ZM [2].

ZMs represent the image by a set of descriptors with a minimal amount of information redundancy. As well, ZMs are rotation and scale invariant so they can deal with the shape problems in images.

Practically, neither global nor local features are enough alone to recognize scene images. As global ZM works on the whole scenery image which is poor as the image contains multiple objects in it and extracting features for the whole image won't represent them. As well, local SURF descriptor ignores the object shape which might give a good clue to identify it.

This paper presents a combined approach of local and object-based feature extractors. The approach is composed of the following steps;

First, all objects in an image are annotated and detected. So an image is treated as a set of objects. Second, to extract object-based feature for each image, ZM features are calculated for the detected objects. Third, interest points detection and features extraction; local interest points are detected from each object in an image and features are calculated accordingly using the SURF technique.

For each of the feature extraction techniques in second and third steps; a dictionary of visual words is built from a set of extracted random patches. The key points of these patches are clustered using k-means to form the visual word dictionary. By mapping the key points to the visual words each image is then represented by a bag of words (BoW). A visual word vector is then constructed with the frequency of the presence or absence of each visual word in an image.

Finally, the resulted visual word vectors of the ZM extractor and the SURF extractor are concatenated. Images are then classified using non-linear support vector machine (SVM) classifier.

The method was tested on popular benchmark data sets. Experimental results obtained on the Caltech-101 and COREL 1000 datasets, presented later in this paper, achieve accuracies that are superior to the best published results to date on the same databases.

After discussing the background and related work in the next section, section III explains the proposed approach to extract features from an image. Experiments and results are presented in section IV. Finally, conclusion and future research are discussed in section V.

II. Background and Related Work

A. Background

1) Zernike Moments

Zernike moments consist of a set of complex polynomials that form a complete orthogonal set over the interior of the unit circle [4]. The computation of ZMs from an input image consists of three steps; computation of radial polynomials, computation of Zernike basis functions, and computation of ZMs by projecting the image on to the basis functions.

- First the Zernike radial polynomials are computed. The real-valued 1-D radial polynomial *Rnm(ρ)* is defined as:

$$Rnm\,(\rho) = \sum_{s=0}^{(n-|m|)/2} c\,(n,m,s)\,\rho^{n-2s} \qquad (1)$$

Where, $c\,(n,m,s) = (-1)^s \frac{(n-s)!}{s!((n+|m|)/2-s)!((n-|m|)/2-s)!}$

n is the non-negative integer that represents the order, and *m* is an integer represents the repetition satisfying $n - |m| =$ (even) and $|m| \le n$.

- Using the radial polynomial, complex-valued 2-D Zernike basis functions, which are defined within a unit circle, are formed by:

$$Vnm\,(\rho, \Theta) = Rnm\,(\rho)\exp(j\,m\,\Theta) \qquad (2)$$

Where, $|\rho| \le 1$, $j = \sqrt{-1}$

Zernike basis functions are orthogonal and imply no redundancy or overlap of information between the moments with different orders and repetitions. This property enables the contribution of each moment to be unique and independent of the information in an image.

- Complex Zernike moments of order *n* with repetition *m* are finally defined as;

$$Znm = \frac{n+1}{\pi} \int_0^{2\pi} \int_0^1 f(\rho,\theta) V^*_{nm}(\rho,\theta)\rho d\rho d\theta \qquad (3)$$

Where, $f\,(x, y)$ is the image function and * denotes the complex conjugate.

To compute Zernike moments from a digital image, the integrals in the equation are replaced by summations and the coordinates of the image must be normalized into [0, 1] by a mapping transform.

2) Speeded up Robust Features

Speeded up Robust Features (SURF) is a fast and robust algorithm for local image representation. It selects interest points of an image from the prominent features of its linear space, and then builds local features based on the image intensity distribution.

In SURF, the Hessian matrix approximation is used in the interest point detection.

Given a point X = (x, y) in an image U, the Hessian matrix H(X, σ) in X at scale σ is defined as follows:

$$H(X,\sigma) = \begin{bmatrix} L_{xx}(X,\sigma) & L_{xy}(X,\sigma) \\ L_{xy}(X,\sigma) & L_{yy}(X,\sigma) \end{bmatrix}$$

Where Lxx (X, σ), is the convolution of the Gaussian second order derivative $g(\sigma)\frac{\partial^2}{\partial x^2}$ with the image U in point x, and similarly for Lxy (X, σ) and Lyy (X, σ).

In interest point description, a distribution of intensity content within the interest point neighborhood is extracted. It is built using the distribution of first order Haar wavelet response in x and y direction.

B. Related Work

1) Local Descriptors

Many different local extraction approaches for object recognition have been introduced in the literature. Hu [3] and Vogel [4] use regular grid of the local regions to extract patches from an image.

Other like Maree [5] uses random sampling and segmentation methods; then images are classified using randomly extracted sub-windows. Recent popular approaches use local features to detect key points in an image. In 1999, Lowe and David G. [6] introduced the well-known scale invariant feature transform (SIFT) descriptor for object recognition. In SIFT; a 128 dimensional feature vector is obtained from a grid of histograms of oriented gradients with automatic scale selection.

Later in 2006, Bay, Herbert, Tuytelaars, and Van Gool [7] proposed the SURF, which is a much faster scale and rotation invariant interest point descriptor. Like SIFT, both approaches produce hundreds of interest points per image. Each of these points is characterized by a dictionary of visual words. An image is then represented by a bag of words (BoW). This representation is then quantized and represented in a feature vector that contains the presence or absence of information of each visual word for the image.

2) Features Quantization

The most popular approach today is the (BoW) [8]. Its goal is to transform the image key point's features into a fixed vector of weights; each weight resembles the importance of the visual word in the image.

In BoW, the key point's descriptors are clustered using k-means clustering algorithm which encodes each key point with the index of its cluster by mapping it to the nearest centroid.

Each cluster is considered to represent a visual word; thus clustering process generates a codebook of visual words. Each image is then represented by a bag of words; which is quantized as a histogram of the frequency of occurrence of visual words.

Several approaches as presented by Mu [9], Jégou [10], and Kesorn [11] worked on enhancing the BoW model. An extension to BoW called Spatial Pyramid Matching (SPM) was proposed by Lazebnik [12]; it creates geometrical relationships between features. It partitions the image into increasingly finer spatial sub-regions and computes histograms of local features from each sub-region. Yang [13] introduced an extension to SPM; sparse coding is used followed by multi-scale spatial max pooling, and propose a linear SPM kernel based on SIFT sparse codes.

3) Global Descriptors

Other studies rely on global feature extractors. Due to moments ability to represent global shape features they have been used extensively in image processing as by Belkasim in [14]. Hu and Ming-Kuei [15] introduces one of the early studies illustrating the potential of image moments invariants that enable successful recognition against scaling, translation, and rotation.

ZMs are one of the most popular global descriptors. It is a complex orthogonal rotation invariant moment composed of a set of circle polynomials in two polar coordinates. Since, the moments are sensitive to rotation; the magnitudes of the moments were used as image features.

In many studies like Khotanzad [16], Kadir [17], Fleyeh [18], and Hwang [19], ZM was used in different image processing applications. In addition, Ono [20], Kim [21], Vretos [22] and Singh [23] used ZM in face recognition. As well, Wang [24] reported promising results using ZM in Chinese character recognition.

III. Proposed Algorithm

Global Zernike moments are mostly applied to images that have explicit shapes so they can be described well. On the other hand, local SURF is concerned more with the object details. In this work, a combined local/ object-based feature extractor is presented. In this method, the SURF features are calculated for each explicit object in the image and its moments to get the object shape characteristics of the texture around it. The proposed algorithm works as follows:

1) For an input image, detect the objects in it.
2) Extract object-based features for each object in an image.
3) Detect the local key points in an object and extract the local features for patches around each of the key points detected.
4) Construct visual codebooks for each extractor; randomly selecting training patches and cluster them using k-means.
5) Encode the image; create a feature vector of a histogram of visual word occurrences in it.
6) Concatenate the resulted visual vectors for each image.
7) Train a support vector machine classifier and predict the image category.

Figure 1 illustrates the main steps in the proposed method. First, images are divided equally into training and testing sets. For each input image, a preprocessing is done; all objects are annotated with a bounding box around it and then cropped, so that single objects are detected from an image instead of dealing with the entire image. ZM features are extracted for each object designated from the image. As well, the object local key points are detected using the SURF technique which results in hundreds of key points for a single image. After balancing all the strongest features among all images, a feature vector of SURF is calculated for a variable size block width around each key point; which is determined from the scale of each key point.

An image ends up with multiple feature vectors with the same dimensions describing each object in it. For each of the ZM and SURF extraction methods, the descriptor is then represented using BoW. The descriptors extracted from the images are grouped into clusters using k-means; resulting in a visual word dictionary. The number of clusters determines the size of dictionary. Each key point descriptor is encoded by mapping it to the index of the nearest cluster centroid; in which it belongs to. Only training images are used in building the dictionary.

Each image is then represented by a histogram of the visual words same length as the dictionary size, which is then converted to a single visual word vector with the frequency of the occurrence of each visual word in it accordingly. This histogram is normalized using L2 norm to make it invariant to the number of descriptors used.

Finally, the visual word vectors resulting from each extractor on a single image is concatenated to form the final feature vector for an image. Figure 2 depicts the method steps applied on a sample image from COREL data set.

Since there are more than two groups in the data sets, a multiclass Support Vector Machine (SVM) classifier is used to train the data set and predict a new image category.

A one-vs-all classification schema is used which constructs M binary SVM classifiers, each of which separates one class from all the remaining classes. The ith SVM is trained with all the training examples in class i with positive labels and all other classes with negative ones. When the M classifiers are combined to make the final decision, the classifier which generates the highest value from its decision function is selected as the winner. Accordingly, the corresponding class label is assigned without considering the competence of the classifiers.

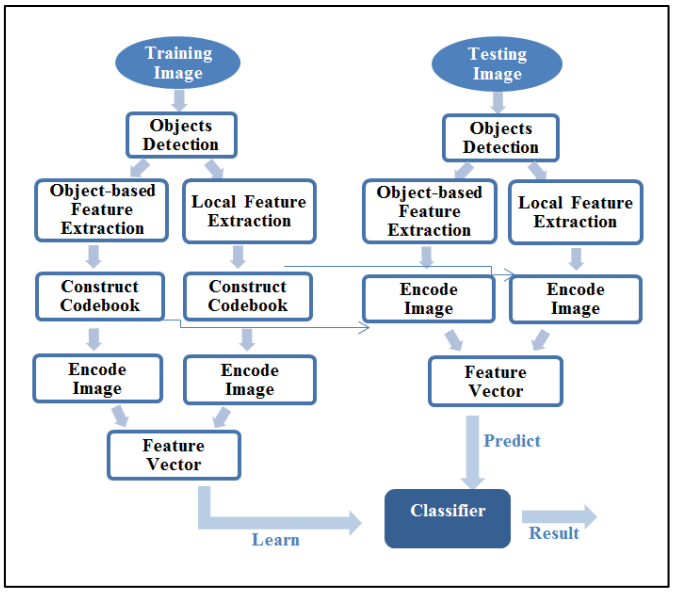

Fig.1. Block diagram for the proposed model

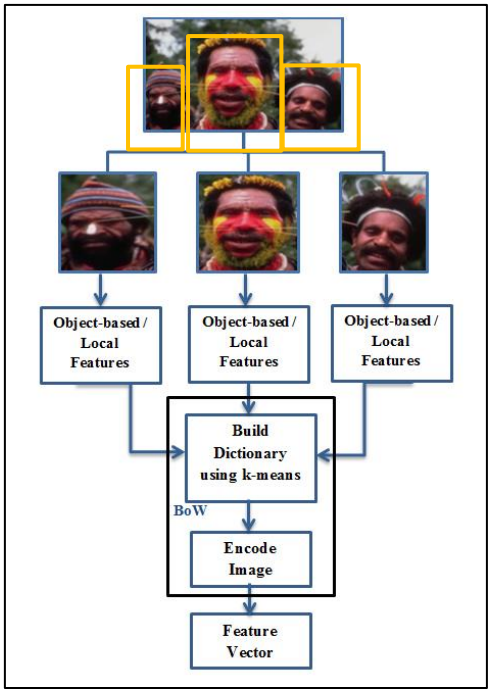

Fig.2. The proposed method applied on a sample image from COREL 1000 data set.

IV. Experimental Results and Discussion

In this section, the results on benchmark data sets are reported; Caltech-101 and Corel 1000. As well, a discussion of the results is offered. First, pre-processing is done on images by annotating objects and cropping them. For each object, ZM features with order 20 are calculated. Then using the SURF-64 descriptor, identify the region of interest points and pick the 80% strongest descriptors to calculate the features around them. The images were all pre-processed into gray scale. All dictionaries are trained for the tests with 2000 basis and random patches from the training image sets.

All results report the average of 10 runs with random selection of images for training and testing.

A. *Caltech-101*

The Caltech-101 data set contains 9144 images in 101 classes including vehicles, animals, objects, flowers, etc. The number of images per category varies from 31 to 800 images. Figure 3 shows sample of images from the data set.

The annotations given with the data set were used that outlines the objects in the image and crop the object accordingly. Examples of annotated objects are shown in Fig. 4.

Table I presents the results compared with other recently proposed ones. To make fare comparison, as recommended by the original data set [25] and suggested by other authors Griffin [26] and Zhang [27], the whole data set was divided into 5, 10, 15, 20, 25, and 30 training images per class and a maximum of 50 images for testing.

In the evaluation, 81.6% accuracy was obtained compared with several existing approaches as in [11], [12] and [26-31] as shown in Table I.

TABLE I. Image Classification Results on Caltech-101

Authors	**Training Images**					
	5	***10***	***15***	***20***	***25***	***30***
Wang [31]	51.1	59.77	65.43	67.74	70.16	73.44
Zhang [27]	46.6	55.8	59.1	62.0	-	66.20
Lazebnik [11]	-	-	56.40	-	-	64.60
Griffin [26]	44.2	54.5	59.00	63.3	65.8	67.60
Boiman [30]	-	-	65.00	-	-	70.40
Jain [29]	-	-	61.00	-	-	69.10
Gemert [28]	-	-	-	-	-	64.16
Yang [12]	-	-	67.00	-	-	73.20
Proposed Method	**54.5**	**61.3**	65.0	**68.3**	**75.6**	**81. 6**

Fig.3. Sample images from Caltech-101 data set

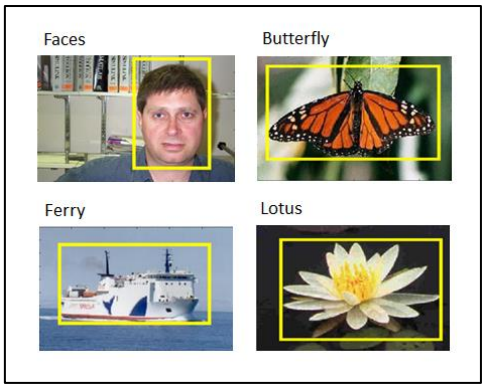

Fig.4. Object annotation in Caltech-101 data set

B. *COREL 1000*

The COREL 1000 dataset consists of 1000 images divided among 10 classes. Each category has 100 images. The categories are Africans, Architecture, Beach, Buses, Dinosaurs, Elephants, Flowers, Food, Horses, and Mountains. Figure 5 shows a sample from each category in the data set.

Objects annotations and cropping are done manually on the whole data set. Samples of the manually cropped objects from images are shown in Fig. 6. The same parameter setup was prepared as mentioned before for Caltech-101. Experiments were performed using 50 images for training and 50 for testing as recommended by Oliveira [32].

In the evaluation, 90.8% recognition performance was reached that is higher than SMK [34], ScSPM [12], LScSPM [33], and SSC [32] methods. Table II summarizes the results on Corel 1000 data set.

TABLE II. IMAGE CLASSIFICATION RESULTS ON COREL 1000

Authors	50 Training Images
Lu [34]	77.9
Yang [12]	86.20
Gao [33]	88.40
Oliveira [32]	88.40
Proposed Method	**90.80**

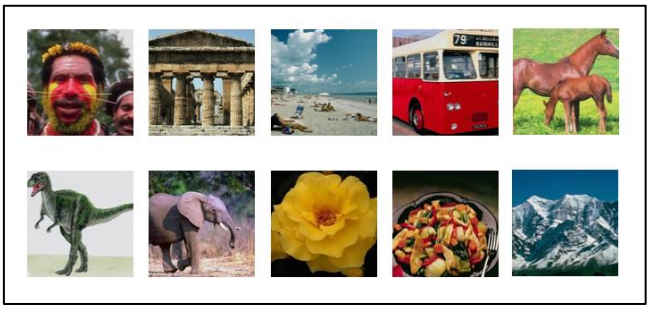

Fig.5. Sample image from each category in COREL 1000 data set

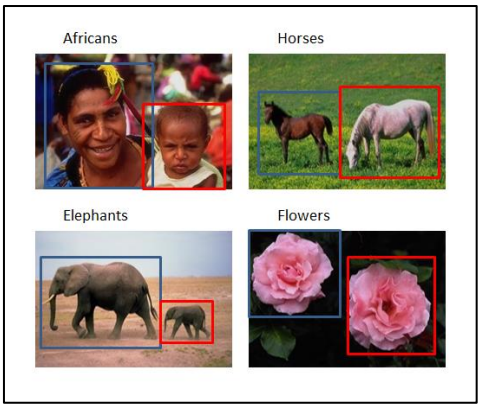

Fig.6. Objects annotation in COREL 1000 data set

The results confirm that object-based shape features and local features are both capable together to recognize objects in scene images with high accuracy. That is, neither applying ZM global feature descriptor nor SURF local descriptor alone was as effective as combining them for enhanced object recognition rates.

The main finding is that working on the detected objects in the image is way better than dealing with the whole image at once. Particularly, in ZM shape-based extractor, working on explicit objects helps to define clear boundary for the object rather than multiple objects in the same image.

Certainly, differences in dealing with the image objects contribute to the correctness variances between this study and the other previously stated ones. However, gray scale images were used in features extraction. Future research should include color features to enhance the recognition accuracy.

V. CONCUSION AND FUTURE WORK

The paper presented an approach for object recognition, which uses a combined feature extractor method. Objects in an image are detected first; some object annotations were already offered with the Caltech-101 data set and others for COREL 1000 were done manually.

The method then extracts the object-based features using ZMs and the local feature SURF descriptors for the captured objects in each image. As well, the technique uses bag of words for compact representation of images so that the image can be represented in a single feature vector.

Based on COREL 1000 and Caltech-101 data sets, the experimental results show that, the proposed method achieves better accuracy according to the published ones on the same databases. Future work will focus on automating the object annotations in images and enhancing the recognition performance by including the color parameter in the SURF extractor instead of dealing with gray images.

REFERENCES

[1] Schmitt, Drew, and Nicholas McCoy. "Object classification and localization using SURF descriptors." CS 229 Final Projects, 2011.

[2] Teague, Michael Reed. "Image analysis via the general theory of moments*."*JOSA* 70.8 (1980): 920-930.

[3] Hu, Ming-Kuei. "Visual pattern recognition by moment invariants." *Information Theory, IRE Transactions on* 8.2 (1962): 179-187.

[4] Vogel, Julia, and Bernt Schiele. "On performance characterization and optimization for image retrieval." *Computer Vision—ECCV 2002*. Springer Berlin Heidelberg, 2002. 49-63.

[5] Maree, Raphael, et al. "Random subwindows for robust image classification."*Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. Vol. 1. IEEE, 2005.

[6] Lowe, David G. "Object recognition from local scale-invariant features."*Computer vision, 1999. The proceedings of the seventh IEEE international conference on*. Vol. 2. Ieee, 1999.

[7] Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool. "Surf: Speeded up robust features." *Computer vision–ECCV 2006*. Springer Berlin Heidelberg, 2006. 404-417.

[8] Lavoué, Guillaume. "Combination of bag-of-words descriptors for robust partial shape retrieval." *The Visual Computer* 28.9 (2012): 931-942.

[9] Mu, Yadong, et al. "Randomized locality sensitive vocabularies for bag-of-features model." *Computer Vision–ECCV 2010*. Springer Berlin Heidelberg, 2010. 748-761.

[10] Jégou, Hervé, et al. "Aggregating local descriptors into a compact image representation." *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010.

[11] Kesorn, Kraisak, and Stefan Poslad. "An enhanced bag-of-visual word vector space model to represent visual content in athletics images." *Multimedia, IEEE Transactions on* 14.1 (2012): 211-222.

[12] Lazebnik, Svetlana, Cordelia Schmid, and Jean Ponce. "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories."*Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Vol. 2. IEEE, 2006.

[13] Yang, Jianchao, et al. "Linear spatial pyramid matching using sparse coding for image classification." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.

[14] Belkasim, Saeid O., Malayappan Shridhar, and Majid Ahmadi. "Pattern recognition with moment invariants: a comparative study and new results."*Pattern recognition* 24.12 (1991): 1117-1138.

[15] Hu, Ming-Kuei. "Visual pattern recognition by moment invariants." *Information Theory, IRE Transactions on* 8.2 (1962): 179-187.

[16] Khotanzad, Alireza, and Yaw Hua Hong. "Invariant image recognition by Zernike moments." *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 12.5 (1990): 489-497.

[17] Kadir, Abdul, et al. "A comparative experiment of several shape methods in recognizing plants." *arXiv preprint arXiv:1110.1509* (2011).

[18] Fleyeh, Hasan, et al. "Invariant road sign recognition with fuzzy artmap and zernike moments." *Intelligent Vehicles Symposium, 2007 IEEE*. IEEE, 2007.

[19] Hwang, Sun-Kyoo, and Whoi-Yul Kim. "A novel approach to the fast computation of Zernike moments." *Pattern Recognition* 39.11 (2006): 2065-2076.

[20] Ono, Atsushi. "Face recognition with Zernike moments." *Systems and Computers in Japan* 34.10 (2003): 26-35.

[21] Kim, Hyoung-Joon, and Whoi-Yul Kim. "Eye detection in facial images using Zernike moments with SVM." *ETRI journal* 30.2 (2008): 335-337.

[22] Vretos, Nicholas, Nikos Nikolaidis, and Ioannis Pitas. "3D facial expression recognition using Zernike moments on depth images." *Image Processing (ICIP), 2011 18th IEEE International Conference on*. IEEE, 2011.

[23] Singh, Chandan, Neerja Mittal, and Ekta Walia. "Face recognition using Zernike and complex Zernike moment features." *Pattern Recognition and Image Analysis* 21.1 (2011): 71-81.

[24] Wang, Tiansheng, and Shengcai Liao. "Chinese character recognition by Zernike moments." *Audio, Language and Image Processing (ICALIP), 2014 International Conference on*. IEEE, 2014.

[25] Fei-Fei, Li, Rob Fergus, and Pietro Perona. "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories." *Computer Vision and Image Understanding* 106.1 (2007): 59-70.

[26] Griffin, Gregory, Alex Holub, and Pietro Perona. "Caltech-256 object category dataset." (2007).

[27] Zhang, Hao, et al. "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition." *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*. Vol. 2. IEEE, 2006.

[28] Van Gemert, Jan C., et al. "Kernel codebooks for scene categorization."*Computer Vision–ECCV 2008*. Springer Berlin Heidelberg, 2008. 696-709.

[29] Jain, Prateek, Brian Kulis, and Kristen Grauman. "Fast image search for learned metrics." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[30] Boiman, Oren, Eli Shechtman, and Michal Irani. "In defense of nearest-neighbor based image classification." *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008.

[31] Wang, Jinjun, Fengjun Lv, and Kai Yu. "Locality-constrained linear coding systems and methods for image classification." U.S. Patent No. 8,233,711. 31 Jul. 2012.

[32] Oliveira, Gabriel L., et al. "Sparse spatial coding: A novel approach for efficient and accurate object recognition." *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012.

[33] Gao, Shenghua, et al. "Local features are not lonely–Laplacian sparse coding for image classification." *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010.

[34] Lu, Zhiwu, and Horace HS Ip. "Image categorization by learning with context and consistency." *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009.