GEO-SPATIAL OBJECT DETECTION USING LOCAL DESCRIPTORS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÇAĞLAR AYTEKİN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
ELECTRICAL AND ELECTRONICS ENGINEERING

JULY 2011

Approval of the Thesis

“GEO-SPATIAL OBJECT DETECTION USING LOCAL DESCRIPTORS”

Submitted by **ÇAĞLAR AYTEKİN** in partial fulfillment of the requirements for the degree of **Master of Science in Electrical and Electronics Engineering, Middle East Technical University,** by,

Prof. Dr. Canan Özgen

Dean, Graduate School of **Natural and Applied Sciences** _______________

Prof. Dr. İsmet Erkmen

Head of Department, **Electrical and Electronics Engineering** _____________

Prof. Dr. A. Aydın Alatan

Supervisor, **Electrical and Electronics Engineering, METU** ______________

Examining Committee Members

Prof. Dr. Uğur Halıcı
Electrical and Electronics Engineering, METU ______________

Prof. Dr. A. Aydın Alatan
Electrical and Electronics Engineering, METU ______________

Prof. Dr. Gözde Bozdağı Akar
Electrical and Electronics Engineering, METU ______________

Dr. Kubilay Pakin
ASELSAN ______________

Dr. Emre Başeski
HAVELSAN ______________

Date: 26.07.2011

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Lastname : Çağlar Aytekin

Signature :

ABSTRACT

GEO-SPATIAL OBJECT DETECTION USING LOCAL DESCRIPTORS

Aytekin, Çağlar
M.S., Department of Electrical and Electronics Engineering
Supervisor: Prof. Dr. A. Aydın Alatan

26.07.2011, 72 pages

There is an increasing trend towards object detection from aerial and satellite images. Most of the widely used object detection algorithms are based on local features. In such an approach, first, the local features are detected and described in an image, then a representation of the images are formed using these local features for supervised learning and these representations are used during classification . In this thesis, Harris and SIFT algorithms are used as local feature detector and SIFT approach is used as a local feature descriptor. Using these tools, Bag of Visual Words algorithm is examined in order to represent an image by the help of histograms of visual words. Finally, SVM classifier is trained by using positive and negative samples from a training set. In addition to the classical bag of visual words approach, two novel extensions are also proposed. As the first case, the visual words are weighted proportional to their importance of belonging to positive samples. The important features are basically the features occurring more in the object and less in the background. Secondly, a principal component analysis after forming the histograms is processed in order to remove the undesired redundancy and noise in the data, reduce the dimension of the data to yield better classifying performance. Based on the test results, it could be argued that the proposed approach is capable to detecting a number of geo-spatial objects, such as airplane or ships, for a reasonable performance.

Keywords: Harris, SIFT, object detection, bag of visual words, weighting words, scale information.

ÖZ

YEREL TANIMLAYICILAR KULLANARAK YER UZAMSAL NESNE TESPİTİ

Aytekin, Çağlar
Yüksek Lisans, Elektrik-Elektronik Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. A. Aydın Alatan

26.07.2011, 72 sayfa

Uydu görüntülerinden nesne tanıma problemi üzerine çokça eğilinen bir problemdir. Nesne tanıma probleminde geniş olarak kullanılan algoritmaların çoğu yerel öznitelik tabanlı algoritmalardır. Böyle bir yaklaşımda, yerel öznitelikler çıkarılır ve tanımlanır, daha sonra bu öznitelikler kullanılarak görüntünün sayısal bir gösterimi çıkartılır ve daha sonra bu gösterimler sınıflandırma için kullanılır. Bu tezde yerel öznitelikler, SIFT ve Harris yerel öznitelik çıkarıcısı ile çıkarılmış ve tanımlayıcı olarak SIFT tanımlayıcısı kullanılmıştır. Görsel kelime çantası modeli görüntüyü görsel kelimeler histogramları halinde ifade etmek için kullanılmıştır. Son olarak eğitim setinden artı ve eksi örnekler ile SVM sınıflandırıcısı eğitilmiş ve bu eğitilmiş SVM parametreleri kullanılarak test görüntüleri sınıflandırılmıştır. Geleneksel görsel kelime çantası modeline ek olarak iki temel yenilik sunulmuştur. İlk olarak görsel kelimeleri, önem ölçümleriyle doğru orantılı olarak ağırlandırma önerilmiştir. Önemli kelimeler basit olarak açıklanacak olursa nesnelerde daha fazla, arkaplanda daha az çıkan kelimelerdir. İkinci olarak gürültü azaltılması, gereksiz fazlalıklar atılarak boyut azaltılması ve performans arttırılması amacıyla histogramlar çıkartıldıktan sonra bir temel bileşen analizi yapılmıştır. Performans simulasyonlarına bakıldığında önerilen yöntemin mantıklı bir performans aralığında gemi ve uçak gibi nesneleri bulma yeteneğine sahip olduğu görülmektedir.

Anahtar Kelimeler: Harris, SIFT, nesne tanıma, görsel kelime çantası, kelimeleri ağırlıklandırma, ölçek bilgisi.

To My Parents

ACKNOWLEDGEMENTS

I would like to express my gratitude and deep appreciation to my supervisor Prof. Dr. A. Aydın Alatan for his guidance, positive suggestions and also for the great research environment he had provided.

I would like to thank to Neslihan Bayramoğlu, Osman Serdar Gedik, Engin Tola, Bertan Günyel, Cevahir Çığla, Yoldaş Ataseven, Engin Türetken, Eren Halıcı, Elif Vural, Osman Topçu, Duygu Arslan, Erhan Gündoğdu, Mustafa Ergül and Ozen Şener for their assistance and friendship.

I would like to thank and express my sincere love to my parents. Their endless love is one of very few things that make me truly happy.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

During the last decades, there has been growing interest in usage of local features in object recognition from visual data. The advantage of using local features lies beneath their robustness to occlusion and clutter; and most importantly, no prior segmentation is required for local feature extraction. The availability of many different feature extraction and description techniques also makes local feature analysis highly extensible. Moreover, the abundant number of features that can be generated from objects is another important advantage of local features. Although, the advantages of local features are promising, a local feature must satisfy certain specifications; such as invariance to illumination, rotation, scaling, minor changes in viewing direction, noise and cluttering. Moreover, for object recognition task, local features must be repetitive, descriptive and distinct.

1.1 Overview of the Thesis

This thesis is devoted to the problem of geospatial object recognition from satellite images and the proposed solution exploits local features extracted from visual data. First, extraction of local features, i.e. key points, is achieved by using a popular blob detector, namely scale invariant feature transform (SIFT), and Harris corner detector; then, these key points are described by SIFT descriptor vector. These feature descriptors are then clustered into manually defined number of clusters using K-means clustering to form a visual word dictionary. In other words, all these vectors are represented by K difference codewords. In this manner, the images are considered as sentences formed by combinations of visual words from the dictionary. The visual words forming the image are obtained by assigning local features in an image to the nearest visual codeword. Since all the images are represented as histograms, which are combinations of different number of visual words, a training set of images for each object containing positive and negative samples are used to train a support vector machine. Next, recognition of objects from images are achieved by first forming a histogram of visual words and then classifying these histograms as object or non-object using learned SVM parameters.

1.2 Fundamental Approaches to Object Detection

Object detection from visual data can be classified into three main tracks:

- Local feature based techniques [1,2]
- Appearance-based approaches [3,4]
- Shape-based methods [5,6]

There are also other approaches, such as part-based or model-based techniques that are preferred in object detection problem. However, this thesis is devoted to aerial and satellite images and the aforementioned three main tracks are more suitable to this kind visual data. Hence, these approaches are examined in more detail.

1.2.1 Local Feature Based Techniques

In computer vision and image processing, the local features are defined as the certain local interesting regions or patches that contains a piece of information, which is relevant for solving the computational task related to a certain application; in our case object detection problem.

Sun et. al [1] proposed a method to solve the problem of detecting geospatial objects presented in high-resolution remote sensing images, automatically. Each image is represented as a segmentation tree by applying a multi-scale segmentation algorithm at first, and all of the tree nodes are described as coherent groups, instead of binary classified values. All of the segments are described as histogram of visual words by implementing Bag of Visual Words model. The trees are matched to select the maximally matched sub-trees, denoted as common subcategories. Then, these subcategories are organized to learn the embedded taxonomic semantics of objects categories, which allow categories to be defined recursively, and express both explicit and implicit spatial configuration of categories.

Tao et. al [2] presents a method for airport detection from large high-spatial-resolution IKONOS images. To this end, airport is described by a set of scale-invariant feature transform (SIFT) keypoints and detect it using an improved SIFT matching strategy. After obtaining SIFT matched keypoints, to both discard the redundant matched points and locate the possible regions of candidates that contain the target, a novel region-location algorithm is proposed, which exploits the clustering information from matched SIFT keypoints, as well as the region information extracted through the image segmentation. Finally, airport recognition is achieved by applying the prior knowledge to the candidate regions.

In addition, Mikolajczyk et. al [39] compared the performance of a large number of local detectors and descriptor in the context of object class recognition and this work provides an extensive evaluation of local detectors and descriptors.

1.2.2 Appearance Based Approaches

Appearance-based methods try to exploit the visual outlook of an object as a whole in order to detect this entity. Template matching of different appearances of the object to be detected can be given as a simple example for this type of algorithms.

Perrotton et. al [3] proposes an algorithm in order to detect and localize objects, for example airplanes on highly cluttered background on remote sensing imagery. First, the discriminative keypoints are obtained and a robust feature description to variations in background is proposed. A number of local descriptors are studied and compared with the new descriptor Histogram Distance on Haar Regions (HDHR). The flaw of this approach is that it assumes that the object desired to be found and the background possesses different texture structures.

Cai et. al. [4] presented an approach to detect airplanes in panchromatic remote-sensing images. The filter first extracts candidate points of airplane centers. Then through a simple clustering method, airplane centers can be located.

Appearance-based approaches are not robust to illumination changes, furthermore contrast between the object and the background is an important factor in learning and detection procedure, for example a bright-colored plane and a dark-colored plane cannot be learned and detected as the same object and needs utilization of different features. Moreover, these kinds of techniques also tend to memorize the object class and cannot yield a good generalization in detection process.

1.2.3 Shape Based Methods

Shape-based object detection algorithms require segmentation in order to extract object regions, then these regions are described by shape descriptors and objects are detected by matching the training descriptor set with descriptors of the tested regions.

Hsieh et. al [5] introduces a hierarchical classification approach in order to recognize aircrafts for remote sensing. In order to have rotation invariance, a new algorithm using symmetry is proposed to guess the orientation of an aircraft. In addition, several image preprocessing techniques, such as noise removal, binarization, and geometrical adjustments are also applied to removing the above variations. After these steps, discriminative keypoints are obtained from each airplane for airplane recognition. In order to combine features, a new boosting approach is introduced to learn weights from training samples.

Iisaka et. al [6] proposed a robust approach for shape description, in aerial or satellite images. This algorithm represents the object in pattern series and structural elements with varying size. The initial few coefficients nearly approximates the shape of an object in aerial or satellite images in terms of shape variences.

The weakness of shape based methods is the segmentation procedure in the beginning of the algorithm. Usually the segmentation procedure is not robust to illumination changes, foreground-background contrast, shadows and other objects in contact with the object desired to be segmented. Hence, a robust segmentation algorithm is needed in shape based methods, which is generally very difficult to be utilized.

Considering the disadvantages of appearance and shape based approaches, local features based algorithm is selected for the purpose of this thesis. The local features based detection algorithm and the reason of this selection will be given in detail in Chapter 3.

1.3 Outline of the Thesis

In Chapter 2, the leading algorithms in the literature exploiting local features are provided.

In Chapter 3, a detailed analysis of local feature extraction and description is given. Furthermore, Harris corner detection and scale invariant feature transform (SIFT) is also explained in detail and Bag of Visual Words (BoVW) algorithm and extensions to BoVW algorithm is discussed. Finally, as a classification method support vector machines (SVM) algorithm is explained.

In Chapter 4, contribution of this thesis to the problem is presented. A novel bag of visual words algorithm for geospatial object recognition is proposed by weighting visual words. The novel proposed method is then compared to the original bag of visual words algorithm and evaluated extensively by experiments by using images containing different objects.

Finally, the summary, as well as conclusions from the thesis, is given in Chapter 5 with some suggested future directions.

CHAPTER 2

RELATED WORK ON LOCAL DESCRIPTOR-BASED OBJECT DETECTION

In a satellite image, in order to detect and recognize an object; first, interesting points or regions are determined and a description is provided for those regions. For the object detection task, these descriptions can be used to locate an object in an image. In order to achieve high detection performance, a local feature should satisfy some certain specifications, which are robustness to noise, occlusion and change in illumination, scale, orientation and viewing angle. Furthermore, in order to achieve a high recognition performance, a local feature needs to be repetitive in objects for different images and in order to achieve low false alarm rate, it must be distinct enough to occur only on the objects of interest, but not at the background.

2.1 Feature Detection

Feature detection methods detect informative small regions, for example blobs, edges and corners. The detected features are guessed to be more interesting than the others, for example the features attractive to humans, and this is thought to be useful for recognizing. Some famous detectors are Harris affine region detector [7], SIFT detector [8] and maximally stable regions (MSER) [9]. Among these keypoint detectors, SIFT is one of the most widely used technique [10, 11, 12, 13]. In some works, simpler approaches are also presented, such as using regular grid method for feature detection [14, 15]. This approach segments the image in same distances with lines and local features are obtained by these horizontal and vertical lines. In the following section, two popular feature extraction techniques, namely SIFT and Harris corner detector, are examined.

2.1.1 Scale Invariant Feature Transform

Scale Invariant Feature Transform [8] is proposed by Lowe in 2004. As mentioned before, scale invariance is one critical characteristic that a good feature descriptor should satisfy. In SIFT, a novel method is proposed to obtain scale invariance and also presented descriptor is rotation and translation invariant.

Briefly, the algorithm convolves the image by 2D Gaussians at different scales and then calculates difference between these convolved images. The local features are then obtained by a scale space extrema search on these difference images.

Before going into further detail in SIFT algorithm, a brief summary of the scale space theory [17] could be useful.

2.1.1.1 Scale Space Representation

The scale space representation L: $\mathbb{R}^N x \mathbb{R}^+ \longrightarrow \mathbb{R}$ of a continuous signal f: $\mathbb{R}^N \longrightarrow \mathbb{R}$ is defined as the solution to the heat diffusion equation [17]:

$$\partial_t L = \frac{1}{2}\nabla^2 L \tag{2.1}$$

The solution to (2.1) is a family of convolutions with Gaussians of different variances.

$$L(.\,;t) = g(.\,;t) * f(.) \tag{2.2}$$

where g: $\mathbb{R}^N x\mathbb{R}^+ \longrightarrow \mathbb{R}$ is

$$g(\boldsymbol{x};t) = \frac{1}{(2\pi t)^{\frac{N}{2}}} e^{-\frac{(x_1^2+\cdots+x_N^2)}{2t}} \tag{2.3}$$

In [17] it is proven that Gaussian function is the unique function for generating a scale-space.

Scale invariance can be achieved by exploiting this scale-space representation. Due to Gaussian function, as the scale grows larger the value of the Gaussian and so the scale space representation of a function decreases in value. In [17], it is proposed to add a normalization term to compensate this decrease. The scale-normalized Laplacian-of-Gaussian function $\sigma^2\nabla^2 G$ is studied in [18, 8].

2.1.1.2 Difference of Gaussian as Approximation of Laplacian

It has been shown that the extrema of $\sigma^2\nabla^2 G$ produces stable features [19, 8]. The difference-of-Gaussian function (DoG) successfully approximates the scale normalized Laplacian-of-Gaussian function. The 2D Gaussian function is defined as:

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x^2+y^2)}{2\sigma^2}} \tag{2.4}$$

The derivative of (2.4) with respect to σ can be approximated as the difference of two consecutive scaled Gaussians, as below:

$$\frac{\partial G}{\partial \sigma} = \sigma\nabla^2 G \approx \frac{G(x,y,k\sigma)-G(x,y,\sigma)}{k\sigma-\sigma} \tag{2.5}$$

The equation can be re-written as:

$$G(x, y, k\sigma) - G(x, y, \sigma) \approx (k-1)\sigma^2\nabla^2\mathrm{G} \qquad (2.6)$$

It can be interpreted from the above equation that if a constant factor of k is used between the scales of two consecutive DoG functions, then DoG automatically includes the scale normalization term σ^2. It should be noted that constant (k-1) term has no effect on the extrema location.

2.1.1.3 Interest Point Detection in Scale-Space

In order to detect interest points, the images are first convolved with a Gaussian to obtain the following scale space of the original image, *I*(*x*,*y*);

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y) \qquad (2.7)$$

Then, differences of Gaussian convolved images are obtained to approximate scale normalized Laplacian of Gaussian:

$$D(x, y, \sigma) = L(x, y, k\sigma) - L(x, y, \sigma) \qquad (2.8)$$

The scale space images $L(x, y, \sigma)$ are grouped in octaves. Each octave contains a number of scale space images where the variance of the Gaussian to produce each scale space image is k times the former. The DoG convolved images $D(x, y, \sigma)$ are then obtained by subtracting the consecutive scale space images $L(x, y, \sigma)$.

After obtaining the DoG convolved images $D(x, y, \sigma)$, the next octave is processed. The first image of the next octave is the image obtained by down-sampling the last scale space of the former octave by two. The process of obtaining DoG convolved images $D(x, y, \sigma)$ is shown in Figure 2.1.

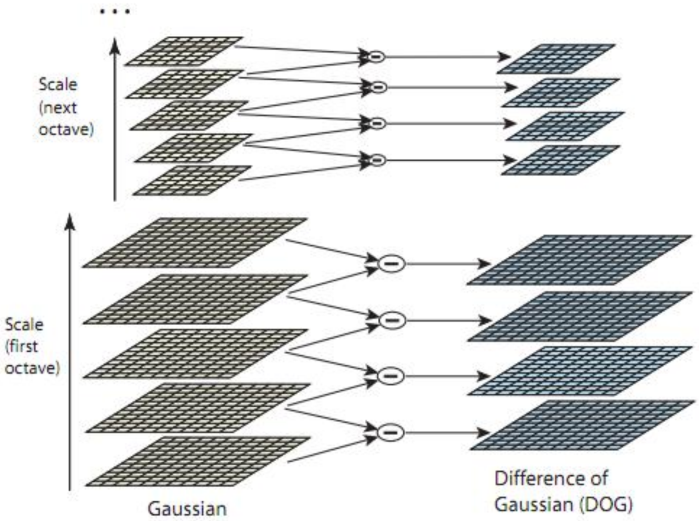


Figure 2.1: In each octave, differences of consecutive Gaussian convolved images are determined. The first image in the next octave is obtained by down-sampling the last image in the previous octave [8].

After this step, the detection of local extrema of $D(x, y, \sigma)$ is achieved by comparing each sample point by its eight nearest neighbors in the current image and nine nearest neighbors in the scales below and above. The pixel is selected, if its value is larger than all of the neighboring pixels. Although this might be assumed as a time consuming process, one should consider that most of the pixels will be eliminated in the first few checks. The selections of local extrema are illustrated in Figure 2.2.

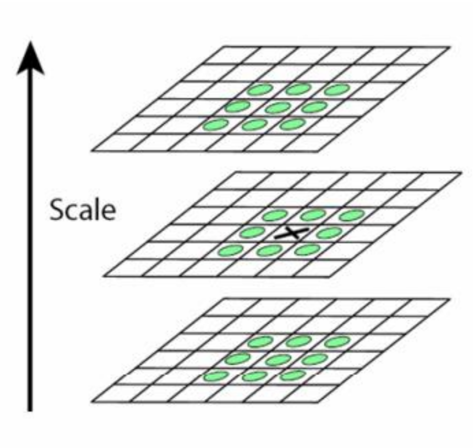


Figure 2.2: Local extrema selection process

When carefully analyzed; SIFT local extrema detection favors blob-like regions due to the shape of difference-of-Gaussian function. This shape is illustrated in Figure 2.3.

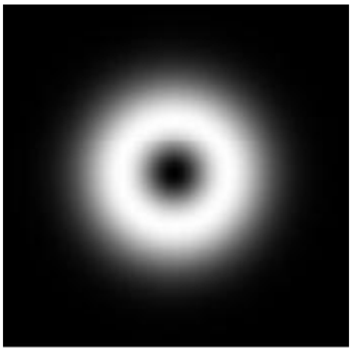


Figure 2.3: The shape of difference-of-Gaussian function

The detected local extrema is in pixel resolution. For more accurate localization of local extremas, a 3D quadratic function is fit to the image intensity function around the keypoint and local extrema of modeling function is detected. The fitting function is obtained using Taylor series expansion of the intensity function around the origin which is the center of the pixel of the original keypoint is.

After finding the exact locations of local extrema, unstable regions are eliminated. Unstable regions are defined as the regions having low contrast and those belonging to edges. Elimination of unstable regions due to low contrast is easily achieved by thresholding the contrast of the resulting region.

The response of an edge to difference-of-Gaussian will have a high value in the direction of edge normal and a low value in the direction perpendicular to the edge normal. The principal curvatures can be computed from the following 2x2 Hessian matrix:

$$H = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{yx} & D_{yy} \end{bmatrix} \tag{2.9}$$

Derivatives are simply estimated as the differences between neighboring pixels. Since the eigenvalues of the Hessian matrix is proportional with the principal curvatures of D, the ratio of the eigenvalues can be used to detect edges. Let the largest eigenvalue be α, the lowest be β and r be the ratio between them.

$$\begin{gathered} Trace(\boldsymbol{H}) = D_{xx} + D_{yy} = \alpha + \beta \\ Det(\boldsymbol{H}) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta \\ \frac{Trace(\boldsymbol{H})}{Det(\boldsymbol{H})} = \frac{(r+1)^2}{r} \end{gathered} \tag{2.10}$$

In order to decrease the computational cost, instead of evaluating the eigenvalue ratio, $\frac{(r+1)^2}{r}$ can be used, which is automatically obtained from trivial calculation of $\frac{Tr(\boldsymbol{H})}{Det(\boldsymbol{H})}$. Hence, if this value is above some threshold, the region is assumed to be edge-like and eliminated.

2.1.2 Harris Corner Detection

A very popular feature detection algorithm is the corner detection algorithm proposed by Harris and Stephens [7]. It is argued that the corners are distinctive features in visual data and can be informative for many different purposes. The corner detection is based on the simple idea that if a window covering a corner region is moved by a small amount in any direction, the change in the average intensity of the pixels in the window must be non-zero. If the region is somewhat constant in terms of intensity, this change will be ignorable, or even if the region contains an edge, the change should be still small in the direction along the edge. The change in average intensity $E(x, y)$ with respect to the move of the window can be defined approximately as:

$$E(x, y) = [x\ y]M[x\ y]^T,$$
$$M = \begin{bmatrix} \sum_W I_x^2 & \sum_W I_x I_y \\ \sum_W I_x I_y & \sum_W I_y^2 \end{bmatrix}, \tag{2.11}$$

where W is the region of the image under the window, I_x and I_y are the partial derivatives of the image in x and y directions, respectively. If the region under the window contains a corner, then the two principal curvatures of the change in average intensity $E(x, y)$ must be both high.

Using the fact that the eigenvalues of M are proportional to two principal curvatures, a cornerness measure can be defined as [1],

$$R = Det(M) - kTr^2(M), \tag{2.12}$$

Where k is constant, and $Det(M)$ and $Tr^2(M)$ are the determinant and the trace of the matrix M.

When the two eigenvalues are similar to each other, the cornerness measure R takes high values; otherwise, it takes relatively low values. If the value of R is the maximum among its 8 closest neighbors and above a threshold, then the pixel is accepted as a corner.

2.2 Feature Description

After detection of features, these points should be described to be able to make matches between similar points in different views. Feature description is the process of mapping these points or regions into useful numbers. A good descriptor should satisfy rotation, scale, affine and translation invariance, so that matching of similar features could be achieved in different views.

One of the most widely used feature descriptors is scale invariant feature transform (SIFT) [8]. SIFT basically determines the distribution of edge orientations at the neighborhood of a point and converts each neighborhood patch to a 128-dimensional vector. After this description, each image becomes a collection of vectors of the same dimension.

Another widely used feature representation method is GLOH [16] which is an extension of the SIFT descriptor for a log-polar location grid with 3 bins in radial direction (the radius set to 6,11 and 15) and 8 in angular direction which results 17 location bins. The gradient orientations are clustered in 16 bins giving a 272 bin new histogram. Then a PCA is utilized. The largest 128 basis vectors are utilized for description.

Histograms of Oriented Gradients (HOG) [35] feature descriptors are also widely used in many areas in image processing in order to detect objects. The idea for this descriptor is that the description of an image can be obtained by the distribution of intensity gradients or edge directions. The image is divided into small patches and histograms of gradient directions are evaluated for each and combinations of these histograms are then used as descriptors.

2.2.1 SIFT Descriptor

SIFT description starts with an orientation assignment. This orientation is assigned due to the magnitudes and directions of the gradients. It should be noted that the orientations are evaluated from the scale space images; thus, satisfying scale invariance. The magnitude and the direction of the gradient of the scale-space image are computed as:

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$$
$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y))) \tag{2.13}$$

Using these gradient computations, histograms of gradients are formed in a region around the keypoint. The histogram bins are quantized as to include a 10 degrees; thus, resulting histogram has 36 bins each consisting of 10 degrees of intervals.

A Gaussian weighting function with σ equal to one half the width of the descriptor window is used to assign a weight to the magnitude of each sample point as illustrated into following Figure 2.4 by a circle. The reason behind this Gaussian smoothing is to be able to reduce sudden changes in the descriptor with the small changes in the window location. Moreover, the gradients far from the center are weighted less, since these gradients are mostly affected from registration errors.

In order to find the orientation, a parabola is fit to the peaks of the histogram and its neighboring bins and the peak position is interpolated to locate the exact orientation. For the cases where the orientation histogram has more than one peak, if these bins have at least 80 percent of the value of the largest peak, several descriptors are evaluated for the keypoint with different orientations.

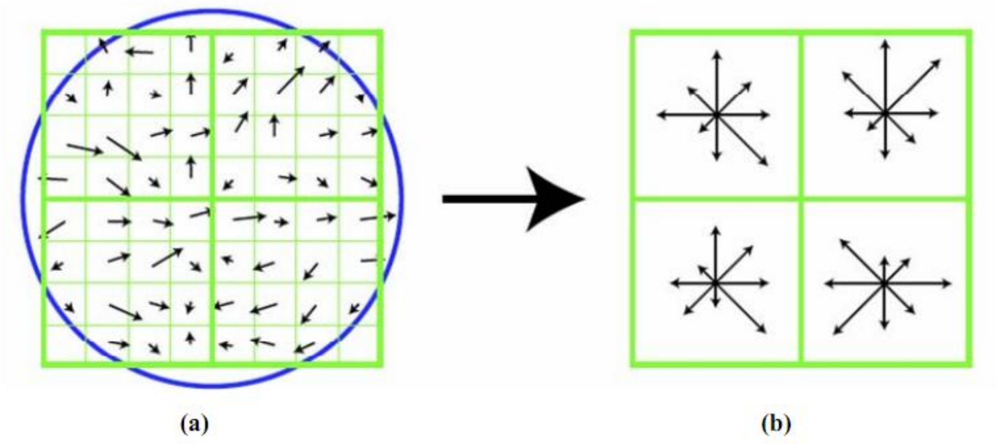


Figure 2.4: Gradient evaluation in a local feature

In order to assure rotation invariance, the descriptor is formed so that the orientation histogram is considered relative to the highest peak in the histogram. As shown in the Figure 2.4, the gradient measurements are achieved in 8x8 grids which are grouped into 4x4 grids. Each of the orientation bins has 8 orientations for gradient histograms; hence, the resulting descriptor is a histogram containing 4x4x8=128 orientation bins.

The described SIFT descriptor is utilized in this thesis and exploited for the bag of visual words model described in detail in the next section.

2.3 Bag of Visual Words

Among many different automatic object detection algorithms, Bag of Visual Words is a very widely used method exploiting local features for object detection. Recently, in many scientific contests for object detection, such as TRECVID [36] or PASCAL [37], Bag of Visual Words, or its derivates, has performed quite remarkably. This method basically treats images as sentences combined of different number of visual words without considering the position of the words in the sentence

Bag of Visual Words algorithm typically includes the following steps:

1. Feature Extraction
2. Feature Description
3. Dictionary Generation
4. Mapping images to histograms of visual words
5. Classification

In order to have a better understanding of the algorithm, Figure 2.5 can be useful.

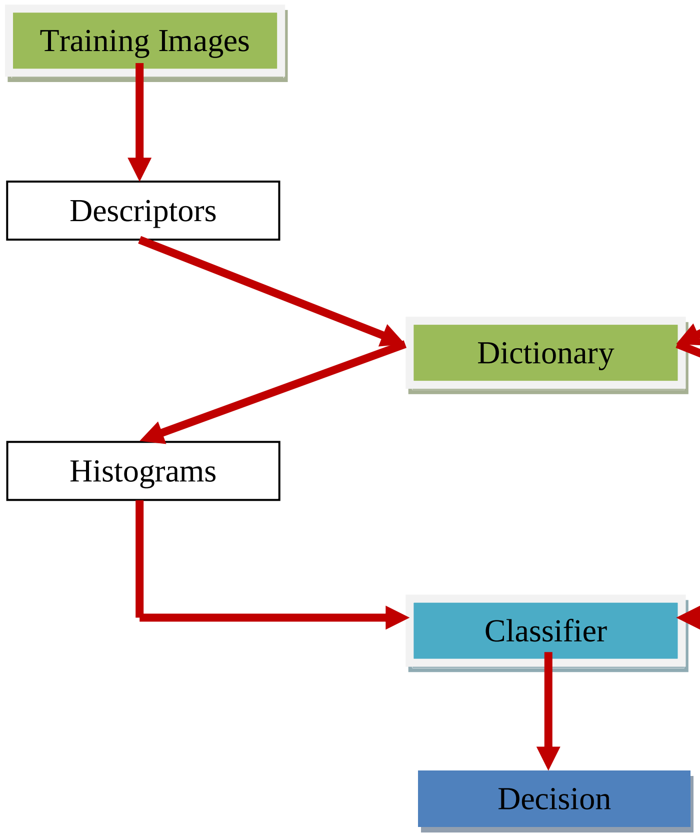


Figure 2.5: Bag of visual words algorithm block diagram

The next step after feature detection and description is to convert the descriptor vector to visual words in order to produce a visual word dictionary. A simple method to achieve this goal is to perform K-means clustering over all such descriptor vectors from visual data [20]. Visual words are then defined as the centroids of these clusters, while the number of the clusters becomes the visual word dictionary size. For more optimal clustering K-means++ algorithm [38] could also be used. This extension is an algorithm for choosing better initial values for the centroids of the classical K-means algorithm.

Thus, each patch in an image is mapped to a certain visual word through the clustering process and the image can be represented as the histogram of visual words from a fixed dictionary of K words. The shape of the histogram is assumed to be the most informative clue about the existence of an object in an image.

Category decision is then achieved by any classifier algorithm through utilization of the shape of the histogram.

Before going into further detail of the classifier used in this thesis (SVM), the extensions to the bag of visual words algorithm should be reviewed.

2.4 Extensions to Bag of Visual Words Algorithm

2.4.1 Soft Assignment

In classical Bag of Visual Words algorithm, histograms of visual words are formed by assigning each local feature to a visual word in the visual word dictionary. This assignment is achieved by increasing the bin corresponding to that visual word in the histogram by one and the feature distance of the visual word descriptor to the local feature descriptor is ignored. In [21], automatic image classifying by modeling soft-assignment in popular codebook model is studied.

In classical Bag of Visual Words Algorithm, histogram generation is obtained by the following relation:

$$CB(w) = \frac{1}{n}\sum_{i=1}^{n}\begin{cases} 1 & if\ w = arg\min_{v \in V}(D(v, r_i)) \\ 0 & otherwise \end{cases} \tag{2.14}$$

where n is the number of regions in an image, r_i is image region i, v is any visual word from the visual word dictionary V and $D(w, r_i)$ is the distance between a codeword w and region r_i.

Alternative histogram assignment methods to classical approach are examined in [21]. One of these approaches is the codeword plausibility:

$$PLA(w) = \frac{1}{n}\sum_{i=1}^{n} \begin{cases} K\big(D(w, r_i)\big) & if\ w = arg\min_{v \in V}(D(v, r_i)) \\ 0 & otherwise \end{cases} \tag{2.15}$$

where K is a kernel inversely proportional with the distance $D(w, r_i)$. In this method, histograms are formed by assigning local features to the visual word bins by increasing the bin value , not by 1, but by a value inversely proportional with the distance between the local feature descriptor and the nearest visual word descriptor. This is also the method examined in this thesis.

2.4.2 Spatial Information

Cao et. Al [22] proposed a novel bag of features algorithm in order to solve the problem of image retrieval in a large scale. This new developed class encodes geometric information. In order to exploit spatial information of words in sentences, projection of these local features to a different space and obtaining ordered bag of features can be a solution; these features are based on which different sets of spatial bag of features are designed to provide invariance of object translation, rotation and scaling. Then, the most representative features are selected based on a boosting-like method to generate a new bag of features like vector representation of an image.

Viitaniemi et al. [23] described spatial extensions to classical BoV and experimentally compared them. In particular, they compare two ways for tiling images geometrically: soft tiling approach and the traditional hard tiling technique. Based on the experimental results, soft tiling is proven to achieve better performance.

Lazebnik et. al [24] presented an approach to recognize scene classes based on near global geometric match. This method works by dividing the image into increasingly fine sub-regions and computing the histograms of local features existing inside each sub-region. According to the test results, this approach provides a boost in overall performance in some extent by exploiting spatial information.

Kobayashi et. al. [25] proposed a bag of hierarchical co-occurrence features method incorporating hierarchical structures for image classification. Local co-occurrences of visual words effectively characterize the spatial alignment of objects' components. The visual words are hierarchically constructed in the feature space, which helps to extract higher-level words and to avoid quantization error in assigning the words to descriptors.

Zhang et. al. [26] proposed high order features to incorporate geometrical information into the bag of feature representation. The authors have used Hough transform method to identify translation and scale invariant high order features co-occurring in two images. The co-occurrence is used to calculate a kernel for a SVM. Then, an efficient algorithm for localization with high order features is also proposed.

2.4.3 Representation Choices

Although feature detection, description, classification method selection is crucial for the performance of bag of visual words algorithm, some selection parameters related to the representation of the features are also very important. These can be listed as:

- Vocabulary size
- Stop-word removal
- Weighting schemes

These three techniques are examined in the next sections.

2.4.3.1 Vocabulary Size:

The vocabulary length of a visual word dictionary is different from the vocabulary length in text dictionary and it depends on clustering numbers. One should try to obtain an optimal cluster size in order to keep the balance in discriminativity and generalizability. If you use a small dictionary the discriminativity is not preserved since very unsimilar keypoints are assumed as the same words, on the other hand if you use a large vocabulary dictionary obviously generalizability will not be preserved since very similar words can be clustered into different words and also the assignment process will not be robust to noise in some degree. Also a large vocabulary use increases the computational cost.

There is no way to find an optimal size of a visual word vocabulary. The vocabulary size used in existing works varies from several hundred [24, 27], to thousands or even ten thousands [10, 28]. The performances of these different approaches cannot be compared due to difference in corpus and classification methods.

2.4.3.2 Stop-word Removal:

Stop-word removal is a standard technique in text categorization. In other words, many words, such as "a", "the", "is", could exist in all documents in English language and they should not be considered during document categorization. Sivic and Zisserman [10] also claimed that the highly occurring visual words in images are “stop-words” and they should be removed from the feature space. However, this idea is not proved with experiments yet.

2.4.3.3 Weighting Schemes:

Since weighting of terms is an important approach in information retrieval (IR) [29, 30], one should explore the use of it in image retrieval. Two dominant approaches in term weighting are term frequency (*tf*) and inverse document frequency (*idf*). Normalization can be thought to be a third factor, converting the feature into a probability distribution. The authors in [24, 27] have used *tf* weighting for image classification, where in another approach [10, 31] *tf-idf* has preferred.

Some weighting schemes are as follows:

- **Binary:** 1 if t_i is present in an image, 0 if not.
- **Term frequency:** tf_i (number of occurrences of term "i" in a document)
- **Term frequency, normalization**: $\frac{tf_i}{\sum_i tf_i}$
- **Term frequency, inverse document frequency:**$tf_i \log\left(\frac{N}{n_i}\right)$, where N is total images in a corpus, *n* is the number of images containing word $t_{i.}$

Most widely used weighting in bag of visual words are term frequency, normalization, which is also used throughout this thesis.

Although there are many works in expanding and improving the classical bag of visual words algorithm in order to improve the performance of object detection, there is still lack of study in many topics. In the Chapter 3, these problems are discussed and the contribution of this thesis is discussed thoroughly.

Before going into further detail about the contribution of this thesis and the novel bag of visual words approach proposed, the fundamental classifier used in this thesis is discussed in the next section.

2.5 Support Vector Machines

The main problem in a linear classifier is that the data is not generally linearly separable. As a classifier, Support Vector Machines rely on preprocessing the data, i.e. mapping it to a higher dimension to make it more separable in the new feature space [32]. With an appropriate nonlinear mapping function $\varphi()$ to a sufficiently high dimension, data from two categories can always become seperable by hyperplanes. It is assumed that each sample x_k from the data undergoes the transformation $y_k = \varphi(x_k)$. For each of the n samples, $k = 1,2,\dots,n,$ we let $z_k = \mp 1$ due to the class that the sample belongs to. A linear discriminant in the mapped $\boldsymbol{y}$ space is:

$$g(\boldsymbol{y}) = \boldsymbol{a}^t\boldsymbol{y} \tag{2.16}$$

where both the weight vector and the sample undergoing mapping are augmented (by $\boldsymbol{a}_0 = \boldsymbol{w}_0$ and $y_0 = 1$). Thus, a separating hyperplane should satisfy:

$$z_k g(\boldsymbol{y}_k) \geq 1 \qquad k = 1, \dots, n; \tag{2.17}$$

The margin is any positive distance from the decision hyperplane. Since a larger margin for separating positive and negative data is better for generalization of the classifier, the goal in the training of the SVM is to make the margin as large as possible. The distance from any hyperplane to a transformed sample $\boldsymbol{y}$ could be measured as: $|g(\boldsymbol{y})|/\|\boldsymbol{a}\|$ and with a positive margin b;

$$\frac{z_k g(y_k)}{\|\boldsymbol{a}\|} \geq b \qquad k = 1, \dots, n; \tag{2.18}$$

The goal is to find weights $\boldsymbol{a}$ which will maximize the margin b. In order to assure unique solution $b\|\boldsymbol{a}\| = 1$ constraint is imposed. Hence, $\|\boldsymbol{a}\|$ is tried to be minimized.

The support vectors are the training samples that satisfy $z_k g(\boldsymbol{y}_k) = 1$. This condition implies that the support vectors are equally close to the hyperplane and are the closest training samples. Furthermore, they are the training samples defining the hyperplane and obviously the hardest samples to classify. Thus, the goal is to find a transformation $\varphi()$ that will separate the data so that minimum number of support vectors satisfying maximum margin could be obtained which reduces the probability of misclassification.

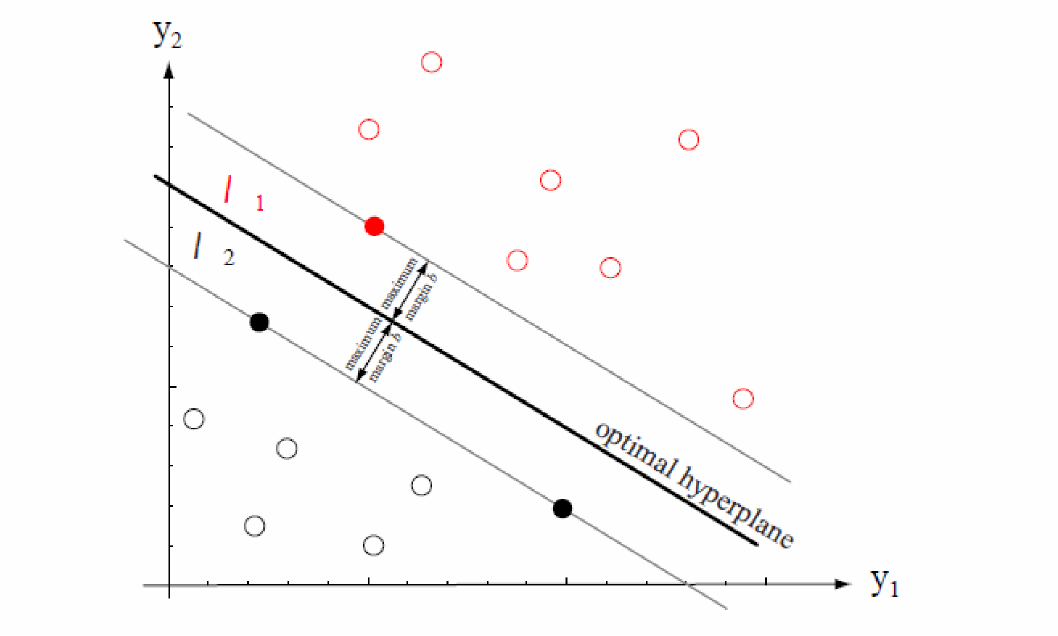


Figure 2.6: Illustration for finding the optimal hyperplane

The first step in training SVM is to choose an appropriate nonlinear transformation function $\varphi()$ that maps the input data to higher dimension. The choice of this mapping function can be related to the characteristics of the input data. If the information about characteristics of the data is absent one can use polynomials, Gaussians or other basis functions.

Since we are trying to minimize weights $||\boldsymbol{a}||$, we use Lagrange multipliers to recast the problem into an unconstrained problem. Hence the following function is constructed:

$$L(\boldsymbol{a}, \alpha) = \frac{1}{2}||\boldsymbol{a}||^2 - \sum_{k=1}^{n} \alpha_k [z_k \boldsymbol{a}^t \boldsymbol{y_k} - 1]. \qquad (2.19)$$

This formulation can be re-represented as follows:

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{k,j}^{n} \alpha_k \alpha_j \, z_k z_j \boldsymbol{y}_j{}^t \boldsymbol{y}_k, \qquad (2.20)$$

subject to the constraints

$$\sum_{k=1}^{n} z_k \, \alpha_k = 0 \qquad a_k \geq 0, \quad k = 1, \ldots, n \qquad (2.21)$$

These equations can be solved using quadratic programming.

The equation (2.22) can be re-written as:

$$L(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{k,j}^{n} \alpha_k \alpha_j \, z_k z_j k(\boldsymbol{y}_j{}^t \boldsymbol{y}_k), \qquad (2.22)$$

where $k(\boldsymbol{y_{j,}} \boldsymbol{y}_k) = \ \boldsymbol{y}_j{}^t \boldsymbol{y}_k.$

Vladimir Vapnik proposed the hyperplane algorithm [33] which was a linear classifier. Later on in [34] a novel method for creating nonlinear classifiers was proposed by applying the kernel trick. The resulting algorithm is very similar to the old one; the only difference is the replacement of each dot product by a nonlinear kernel function. This kind of selection allows the algorithm to fit the maximum-margin hyperplane in a transformed feature space. The transformation may be nonlinear and the transformed space high dimensional; thus, although the classifier is a hyperplane in the high-dimensional feature space, it may be nonlinear in the original input space.

Some common kernels can be listed as follows:

-Polynomial kernel (homogenous): $k(\boldsymbol{y_{j,}}\boldsymbol{y}_k) = \boldsymbol{y}_j{}^t\boldsymbol{y}_k{}^{\boldsymbol{d}}$

-Polynomial kernel (inhomogeneous): $k(\boldsymbol{y_{j,}}\boldsymbol{y}_k) = (\boldsymbol{y}_j{}^t\boldsymbol{y}_k + 1)^{\boldsymbol{d}}$

-Gaussian or Radial Basis Function: $k(\boldsymbol{y_{j,}}\boldsymbol{y}_k) = \exp\,(-\gamma\left\|\boldsymbol{y_j} - \boldsymbol{y_k}\right\|^2)$

-Hyperbolic tangent: $k(\boldsymbol{y_{j,}}\boldsymbol{y}_k) = tanh(\kappa\boldsymbol{y}_j{}^t\boldsymbol{y}_k + \boldsymbol{c})$

Based on experimental performance measures, Radial Basis Function is used in this thesis as a non-linear kernel, although the other selections could also be utilized.

CHAPTER 3

PROPOSED METHOD

For the proposed object detection method in this thesis, SIFT and Harris detectors are used as local feature detectors and SIFT is used as the only descriptor. These descriptors are then exploited during object detection problem by bag of visual words algorithm as explained in the previous section. Although many attempts are proposed in the literature to overcome some weaknesses of this algorithm as discussed in the previous chapter, there is still some chance to make further improvements in the performance of the algorithm.

The main contribution of this thesis in order to improve bag of visual words algorithm is weighting visual words and exploiting PCA to eliminate the undesired redundancy from the histogram of visual words and achieve better classifying performance.

3.1 Visual Word Weighting

In most of the studies that use bag of visual words algorithm, all of the visual words used in object detection have been treated by the same importance while forming the histograms. Hence, during detection of objects, the context information should be taken into account and more importance should be given to the object itself. Thus, while forming the histograms, the visual words belonging to the object itself should be favored. First of all, an importance measure has to be defined in order to implement this idea. A straightforward approach is to favor the features that are occurring more in the object and less in the context. The relation below defines the importance measure of a visual word:

$$imp(w) = \frac{\frac{CBO(w)}{CBNO(w)}}{\sum_{w=1}^{K} \frac{CBO(w)}{CBNO(w)}} \tag{3.1}$$

where $imp(w)$ is the importance of the word w in a dictionary having K visual words, $CBO(w)$ is the value of the bin corresponding to the visual word w in the histogram (codebook) occurring in the object and $CBNO(w)$ is the value of the bin corresponding to the same word in the codebook occurring in the context.

|Based on (2.17), the new histogram calculation taking these importance measures of words into account should be as follows:

$$PLA(w) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} imp(w)\, K\big(D(w, r_i)\big) & if\ w = arg\min_{v \in V}(D(v, r_i)) \\ 0 & otherwise \end{cases} \tag{3.2}$$

3.2 Principal Component Analysis

After adding scale information to descriptors and using an importance measure to weight words, while generating histograms of images, before the classification step, a PCA analysis should be performed. The motivation behind PCA analysis is to reduce noise and irrelevant information within histograms and more importantly dimension reduction of these histograms. This part is performed by following steps:

1. Evaluate scattering matrix S_C using following equation:

$$Sc = \sum_{j=1}^{l}(h_j - \bar{h})(h_j - \bar{h})^T \tag{3.3}$$

   where h_j is the j^{th} histogram in a dataset containing a total l number of histograms and $\bar{h}$ is the ensemble average of the histograms.

2. Find basis vectors of scattering matrix using Singular Value Decomposition.

3. Consider only the first *K* basis vectors.

4. Calculate projections of histograms onto these basis vectors.

5. Form *K*-length new histograms consisting of these projections.

As a final step an SVM is trained by histograms of images containing objects as positive samples and histogram of images that are not containing objects as negative samples. A test image is then classified as object or not-object by the trained SVM classifier.

The block diagram of the overall algorithm is shown in Figure 3.1 :

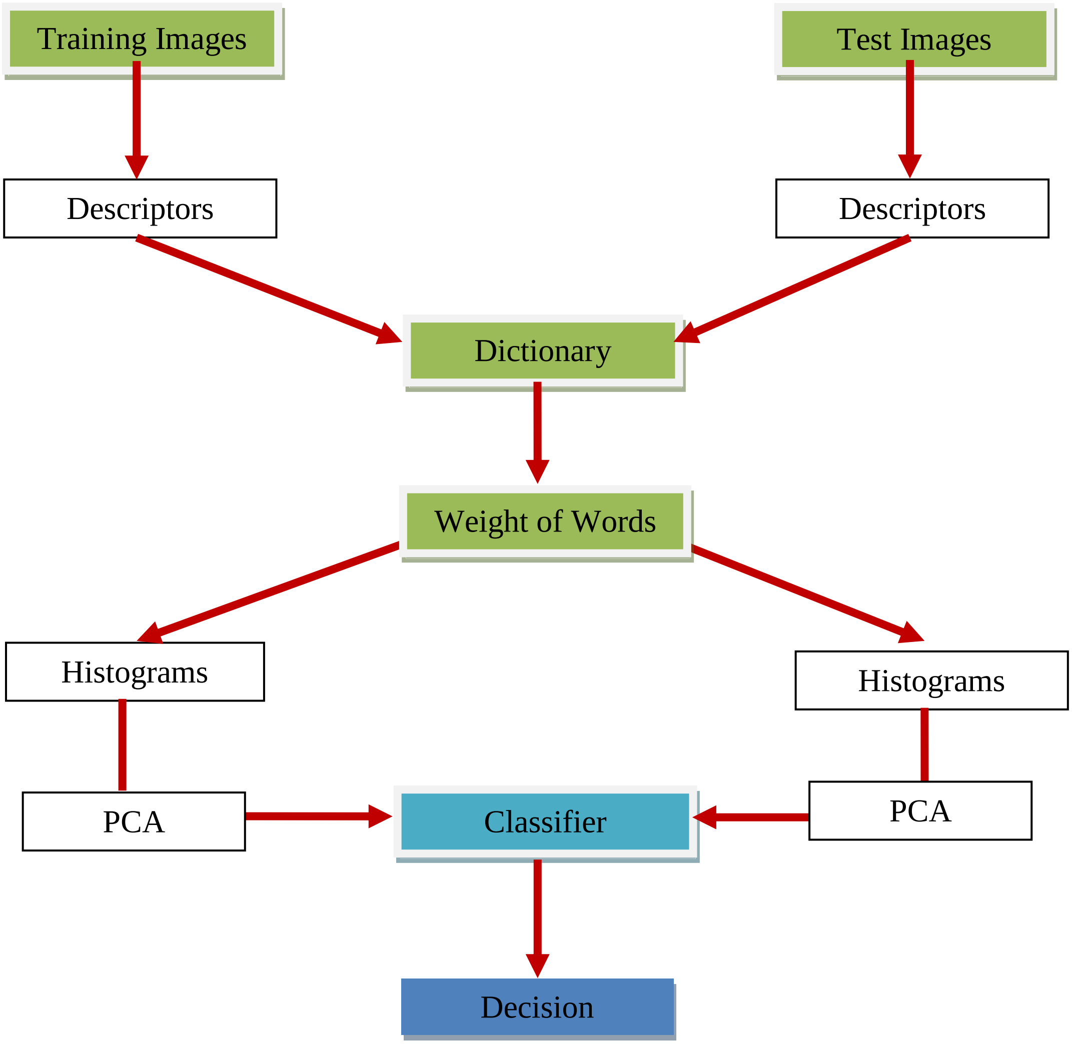


Figure 3.1: Block diagram of the overall algorithm

3.3 Tests on Proposed System

The performance of the proposed algorithm in Figure 3.1 is tested via tests. The tests are conducted first for the analysis of the repeatibility of features for SIFT points and Harris corners. Then, visually important words on test objects are examined in order to obtain a relation on different objects. Afterwards, the performance of the proposed system is tested against the conventional algorithm.

3.3.1 Repeatibility of Keypoints

As mentioned before, repeatability of a keypoint detector is very important, since it highly affects the performance of following steps after keypoint detection; i.e. classification. The test images are captured from Google Earth across the world in order to generalize the data. The tests are performed with two class of object namely ships and planes. The data set contains 18 different docks containing 149 ships and 9 different airports containing 119 airplanes. In the image gathering process, the eye altitude is set to the sum of the elevation terrain and a certain distance from earth which is selected as 700 meters in this thesis since it provides images in 0.5 m resolution.

In the following figures, the keypoint detection results for SIFT detector and Harris corner detector are presented. In Harris corner detection algorithm, the size of the window where a corner is searched is selected as 7x7, 15x15 and 21x21 and the results of each individual window size are merged.

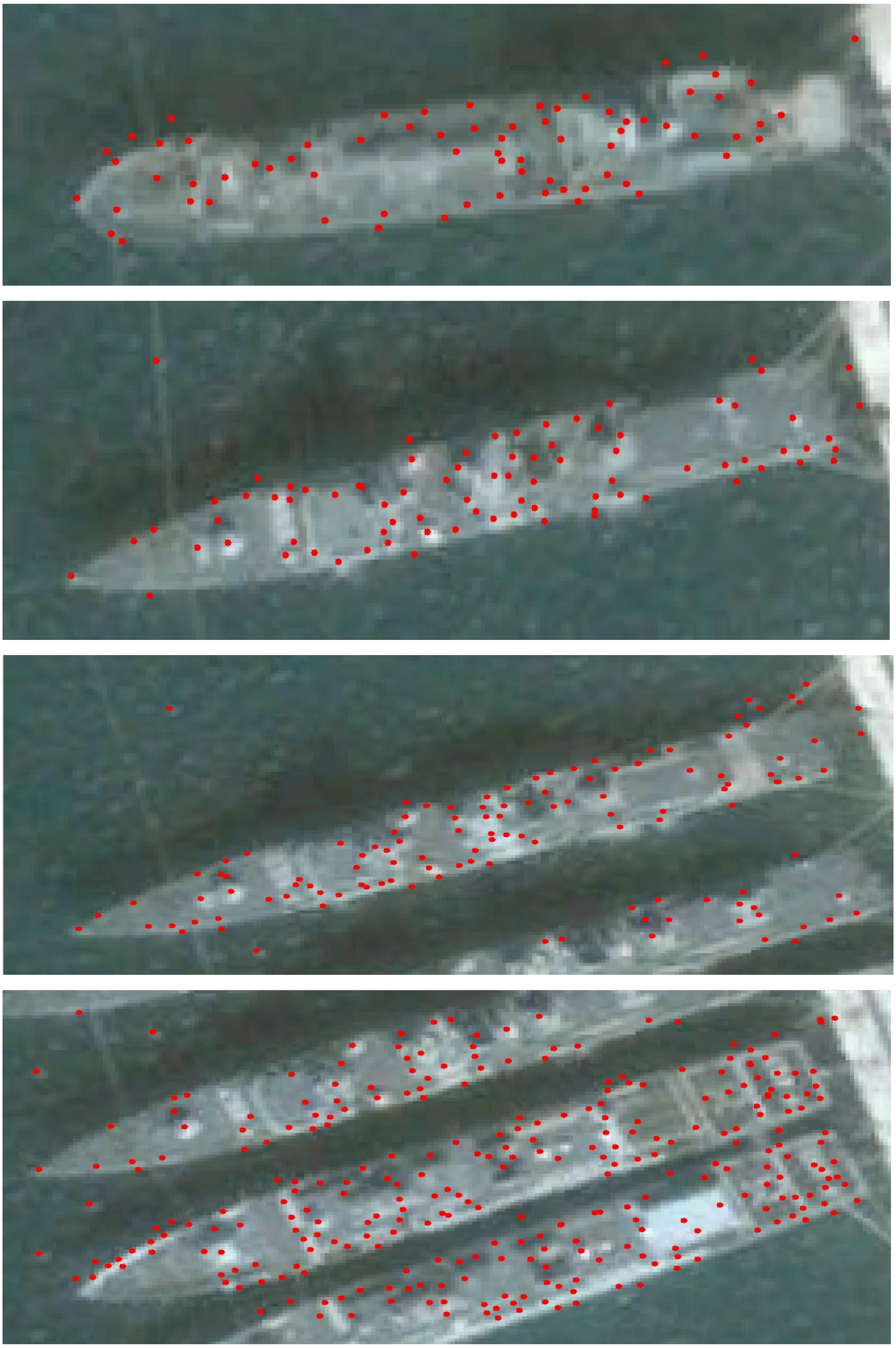


Figure 3.2 Keypoint detection results of SIFT keypoint detector in ship object class

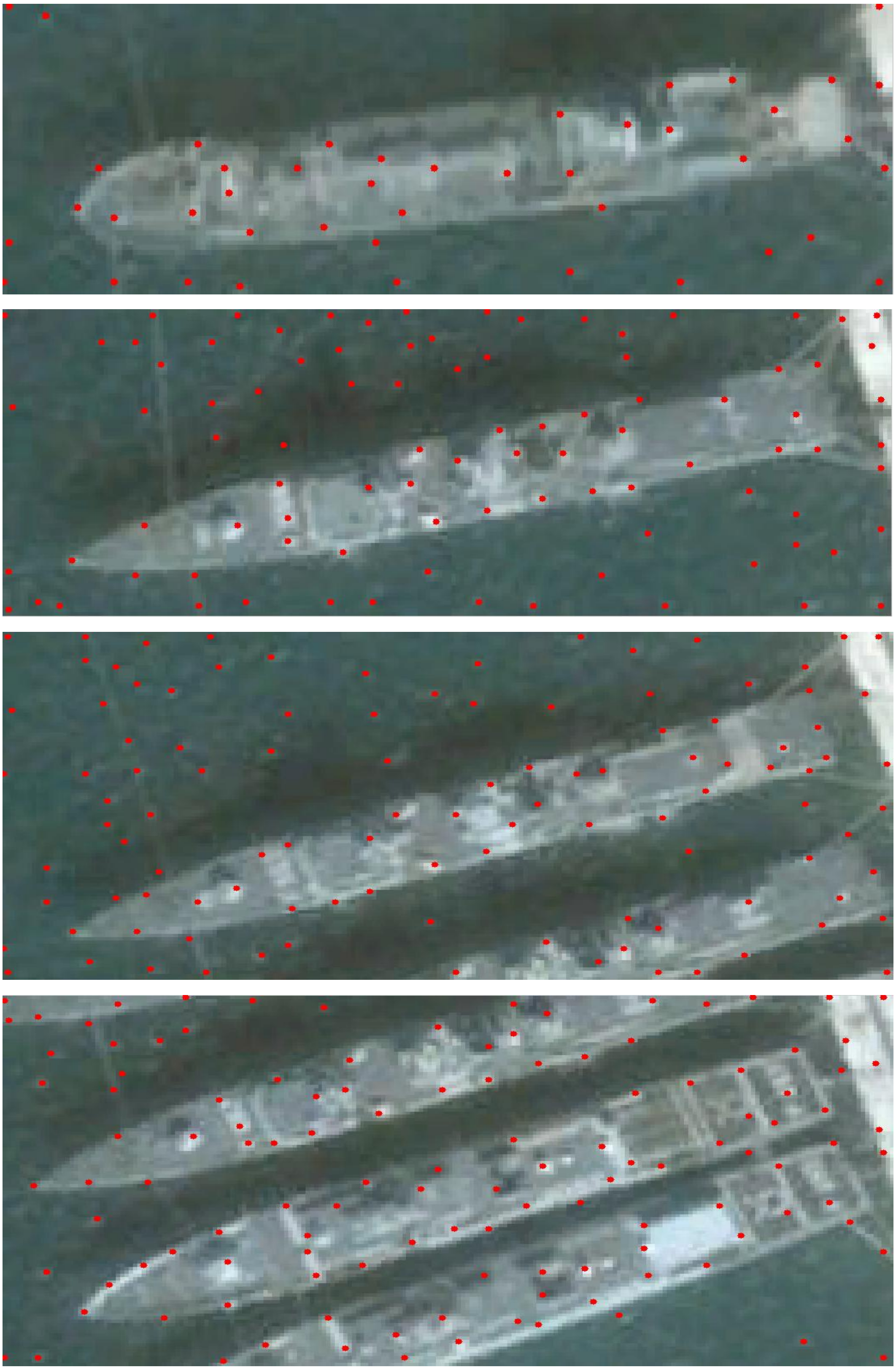


Figure 3.3 Keypoint detection results of Harris keypoint detector in ship object class

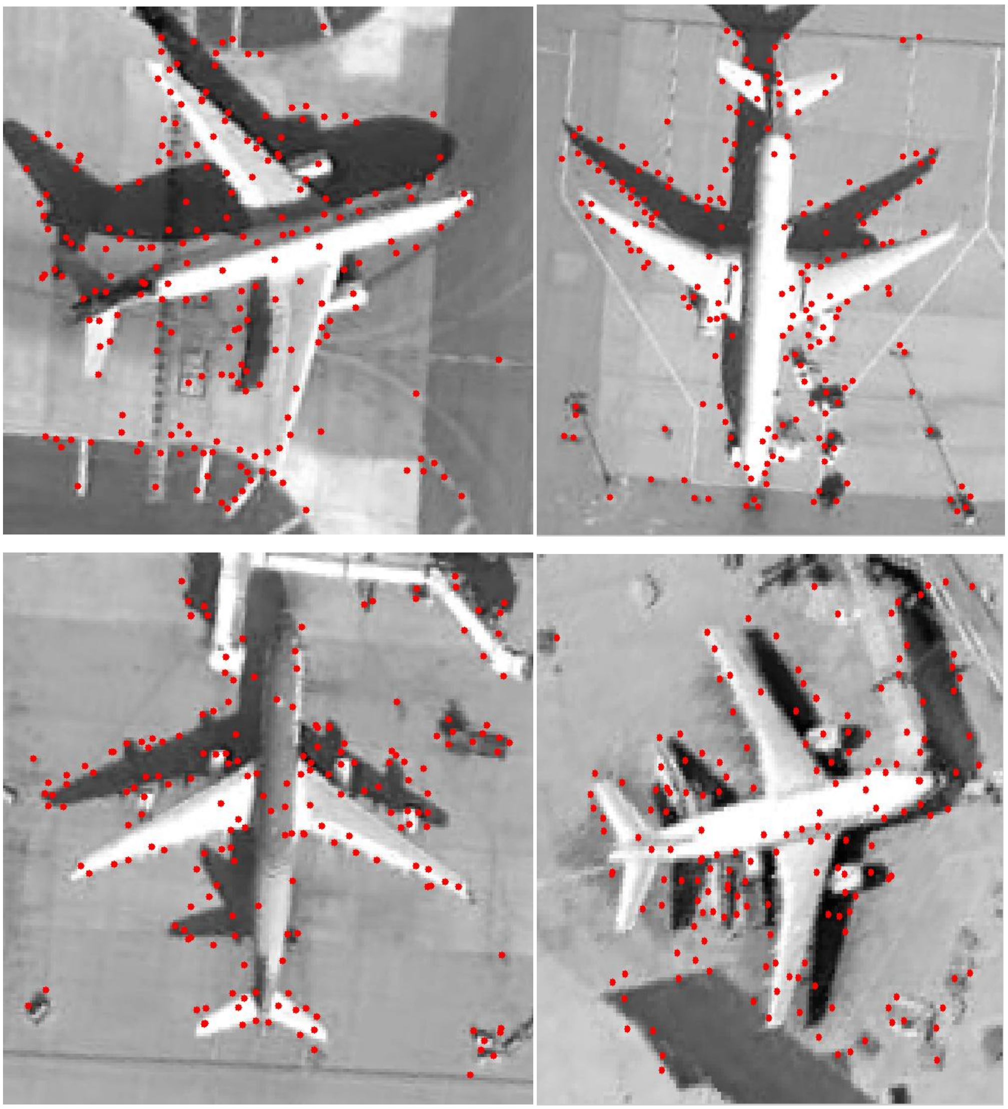


Figure 3.4 Keypoint detection results of SIFT keypoint detector in plane object class

An important observation from the above keypoint detection results for SIFT keypoint detector, is its unstable behavior compared to the corner-like keypoints around an important feature of the planes and ships. For example, the exact locations of the resulting keypoints around the wings and tails of the planes vary a lot in each picture. This observation is also valid for the ship object class. The important features characterizing ship object class, such as noses of the ships are sometimes missed by SIFT keypoint detector, even if the noses were detected, the exact locations of the keypoints differ around the nose. This result may yield a difference in the description of these keypoints. On the other hand, blob-like features, such as engines of the planes or square and circle objects in the ships are determined by a good precision in terms of localization. This result is due to the fact that SIFT is a blob-like region detector.

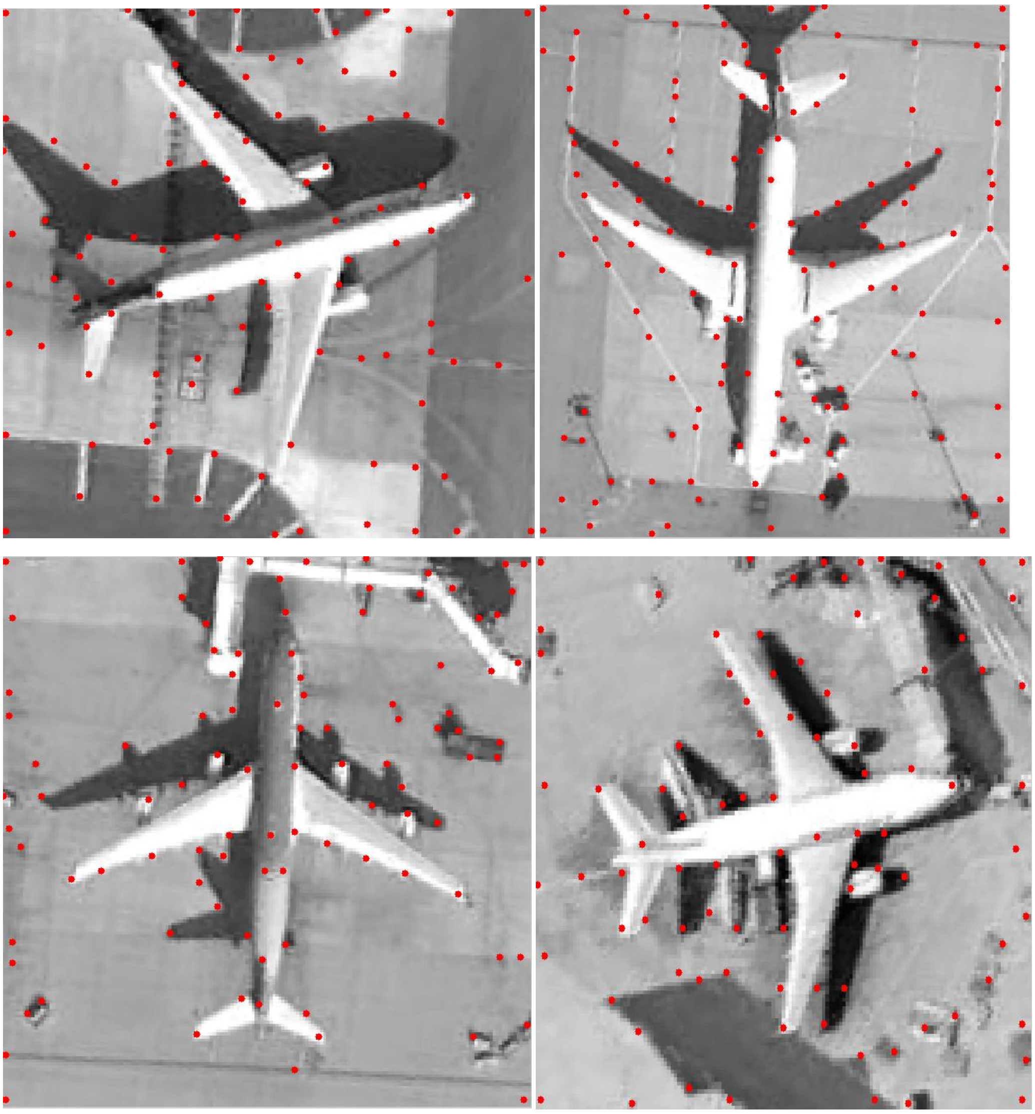


Figure 3.5 Keypoint detection results of Harris keypoint detector in plane object class

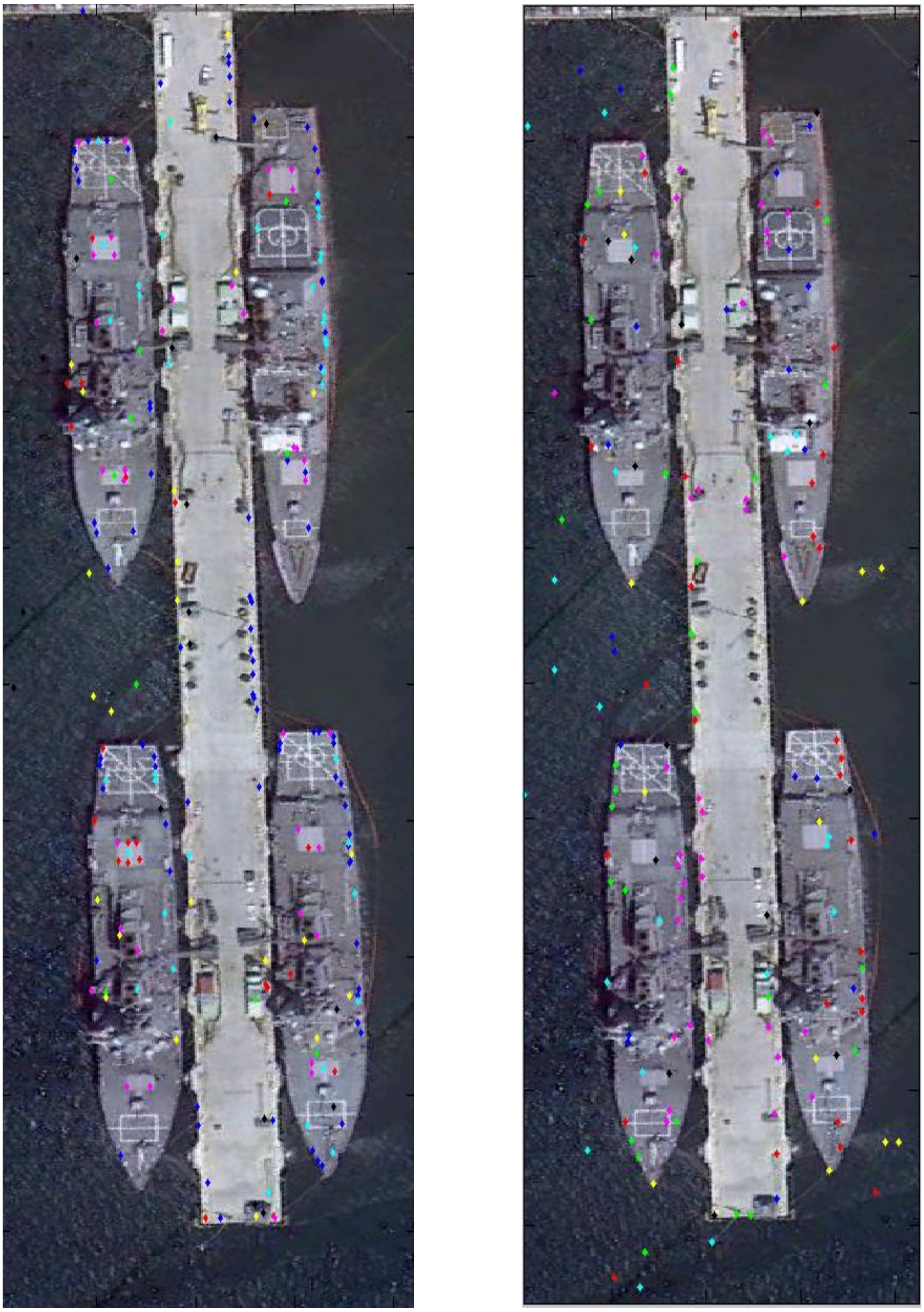
Corner-like keypoints, such as wings, tails and noses in the plane object class are obtained with near-perfect localization by Harris corner detector. Moreover, the intersection of the body of the plane and the wings or the intersection of the body and the tails are also determined repeatedly. It should be noted that these features are very important for characterizing the plane object class. In the ship object class noses and the corners on the stern of the ships are obtained repeatingly. Most particularly, the nose of the ships have a very unique formation and are very important features, thus repeating keypoints in this regions with fine precision in localization is required in order to improve performance of classification. On the other hand it should be noted that Harris corner-detection can be inefficient when trying to locate blob-like features like engines of the planes or circular or square objects on the ships.

For both of the keypoint detection algorithms, shadows distort the outputs of the detection; however, in case of SIFT, shadows might form blob-like regions with a very unique formation that is expected to occur only on planes. However, this kind of occurrence could be very dangerous, since it is highly unstable and very dependent to time, when the image was captured, the orientation of the plane and the sunlight direction.

As a conclusion, SIFT outperforms Harris corner detection output when blob-like features are in consideration; on the other hand, Harris keypoints perform better against SIFT, in case of corner-like features are tested in terms of repeatability. As an observation, it should be noted that in many man-made objects, corner-like regions occurs more and sometimes can be more informative compared to blob-like features.

3.3.2 Important Visual Words

The most important words extracted after the word weighting strategy discussed in the previous sections gives rich information about the important characteristics of the objects. It should also be very useful to check the most important words in SIFT and in Harris corner detector results, in order to evaluate the keypoint detection performance for object detection.



(a) (b)

Figure 3.6 Most important 7 words in ship object class with (a) SIFT (b) Harris (From most important to least important color codes: blue, green, red, cyan, magenta, yellow, and black.)

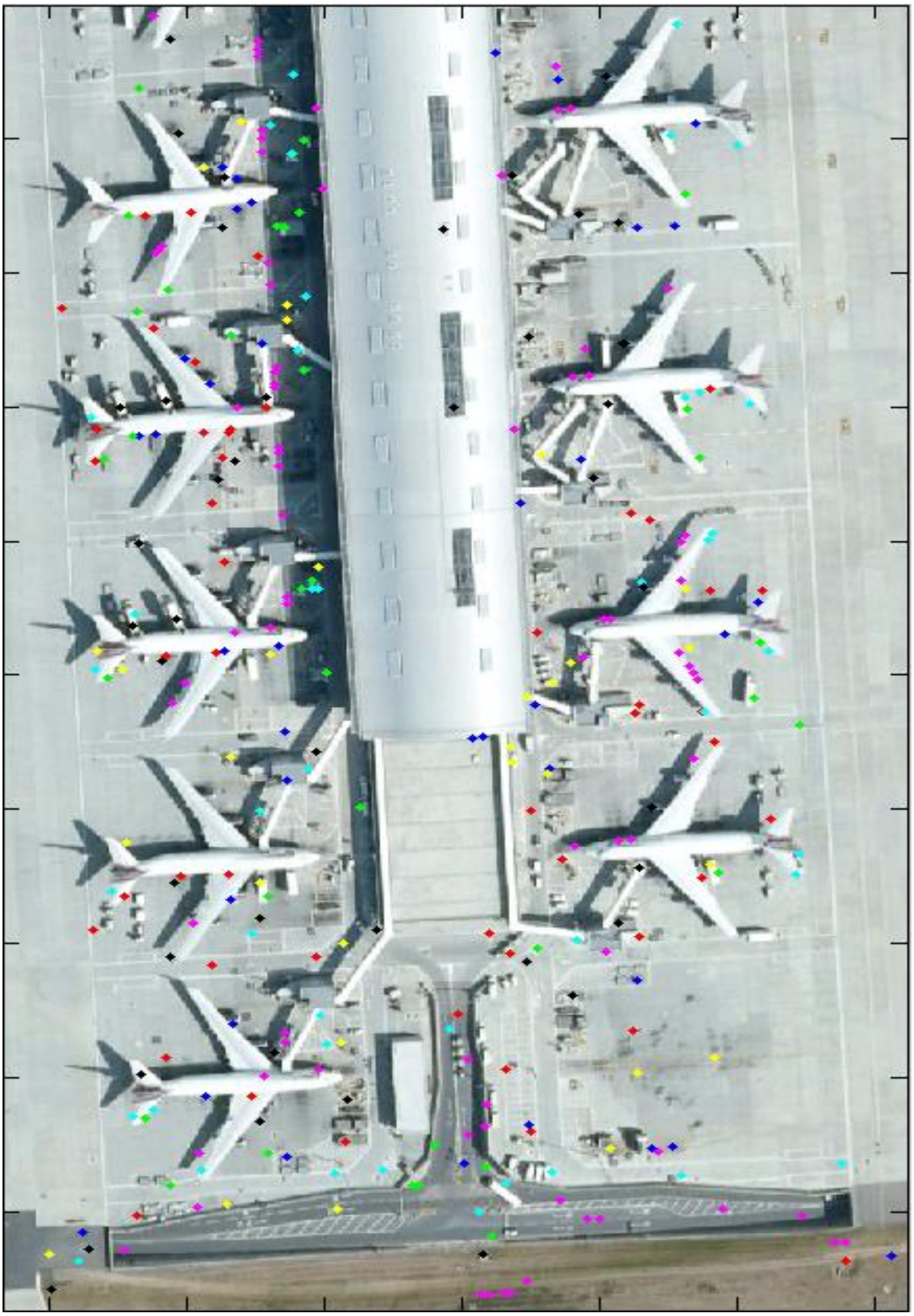


Figure 3.7 Most important 7 words in plane object class with SIFT detector (From most important to least important color codes: blue, green, red, cyan, magenta, yellow, and black.)

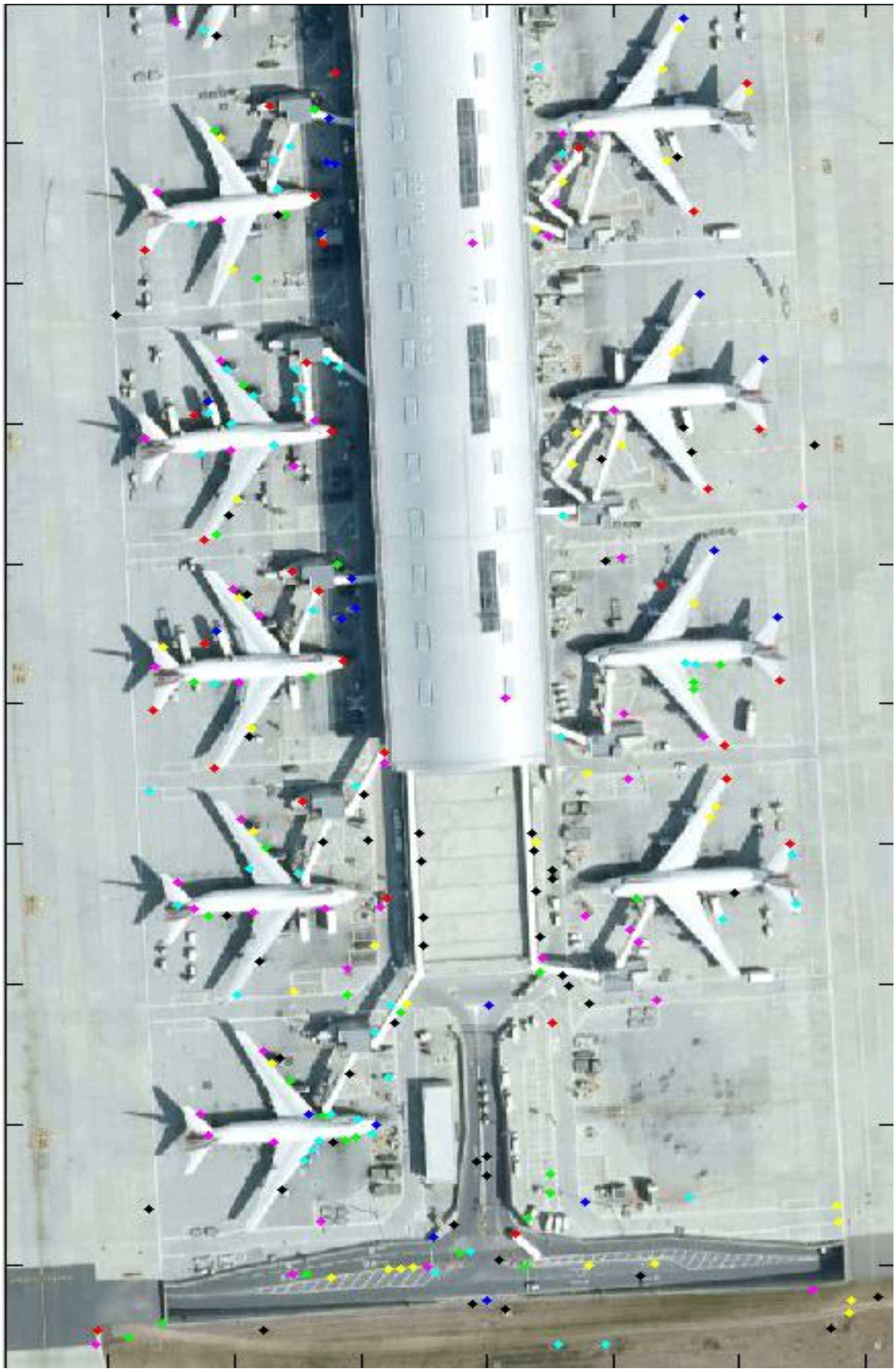


Figure 3.8 Most important 7 words in plane object class with Harris detector (From most important to least important color codes: blue, green, red, cyan, magenta, yellow, and black.)

As it can be interpreted from the figures above, visual word weighting favors the visual words occurring more on the object and less in the background. It can also be observed that the favored visual words are generally occurring in the regions having similar gray-level characteristics. An observation is that a visual word does not have to occur at a unique location of the object. For example, the same visual word occurs both in the wing and the tail of a plane, since these parts are very similar considering the descriptor of the points.

The most important three words in the ship object class for both of the keypoint detection algorithms (SIFT and Harris) occurs around the neighborhood of the painted regions on the ship or on some particular objects on the ship. Another critical observation is related to the case in which SIFT is selected as a keypoint detector: the two most important words turn out to be more repeating compared that of Harris, but it should be noted that they are highly unreliable in their locations. The visual words occurs less, when Harris is selected as a keypoint detector, but the locations of the visual words are quite stable, compared to SIFT; hence, these visual words tend to occur at numbers comparable to each other in every ship, which is not the case for SIFT keypoints.

Better localization of keypoints might result in better classifying performance, Harris keypoint detection result is much more repeatable with respect to SIFT detection results, hence gives a better localization performance which is expected to achieve a more robust detection result.

Furthermore, in the Harris test, a new visual word (in yellow) can be seen which is the visual words defining the nose of the ships. This basically justifies the comments made in the previous section that the SIFT descriptor is unsuccessful while trying to find the noses of the ships. Since the noses of the ships are very important features characterizing the ship class, Harris keypoint brings important descriptions that are invariant to objects in different images.

The weakness of the SIFT keypoints when trying to locate corners can be observed much severely during the plane test. In the Harris keypoint tests, the most important words occur dominantly on the wings, noses and tails of the planes and the transition between the parts of the planes, whereas in the SIFT test the most important word occurs in the engines of the planes, and, unfortunately, not very repeating. Thus, one can obviously state that Harris corner detector is highly reliable compared to SIFT for both of the object classes.

3.4 The Performance Tests

The tests are conducted by using cross-validation 9-fold for plane object class and 17-fold for ship object class. In other words, for *N*-fold test, *N*-1 subsets of the available data set are utilized during training and test by the remaining set. After repeating this process for *N* cases, the ensemble average of these performances are returned.

In order to evaluate the performance, one should define a performance criterion. The performance measure is given in terms of receiver operating characteristics curves (ROC curves). ROC curves can be very informative while trying to obtain the performance of an object detection algorithm. For this thesis, False Alarm Rate vs. Recall curves are used.

Before giving the definitions for False Alarm Rate and Recall, one should define the following related concepts.

- True Positives: The number of items correctly labeled as belonging to the positive class (i.e. the object being detected).
- True Negatives: The number of items correctly labeled as belonging to the negative class.

- <u>False Positives</u>: The number of items incorrectly labeled as belonging to the positive class.
- <u>False Negatives</u>: The number of items which are not labeled as belonging to the positive class, but should have been.

By the light of the definitions above, *Recall* is the ratio of correctly found objects (true positives) over the total object number in the image (true positive +false negative). False Alarm Rate is the ratio of the falsely found positives (false positives) over the total patches not containing objects in the image (false positives + true negatives). In the light of this explanation one can say that an ideal result should have value equal to 1 for Recall and 0 for False Alarm Rate, respectively. In summary,

Recall = *True Positive* / (*True Positive* + *False Negative*)
False Alarm Rate = *False Positive* / (*False Positive* + *True Negatives*)

The advantage of using this type of a ROC curve is due to the fact that it gives a solid indicator for the performance measure and the resulting curve must always be convex.

Extraction of ROC curve by using a SVM classifier is crucial. In classical SVM

$$\boldsymbol{y}_k = \begin{cases} 1 \quad if\ (\boldsymbol{w}\boldsymbol{x}_k + b) > T \\ -1\ if\ (\boldsymbol{w}\boldsymbol{x}_k + b) < T \end{cases}, \text{ k=1,2,…n} \tag{3.4}$$

where T =1 was the classification condition for the classical SVM. If T is chosen to be larger than the maximum of $(\boldsymbol{w}\boldsymbol{x}_k - b)$ in the dataset, obviously the classification algorithm will tend to classify all the data as negative class. Similarly, if T is chosen smaller than the minimum of $(\boldsymbol{w}\boldsymbol{x}_k - b)$ in the dataset, the classification algorithm will classify the data as being positive. By slowly varying T value between these two extremes, a ROC curve can be extracted by evaluating the False alarm rate and recall for each value of T in the interval.

In the detection procedure, the image is searched with overlapping NxN size windows, where N is manually selected for each object class. The search window is shifted $N/2$ in every direction so that the object is guaranteed to fall in at least one search window if $N/2$ is selected as the maximum length of an average object. The search procedure is visualized in Figure 3.9. In some cases the algorithm tends to find the same object more than once since an object can fall in more than one search window. In order to prevent multiple object detection, detected objects are represented as dots which are the centers of the most important word available in the window. Generally, if the same object is detected more than once with different search windows; the dots representing the objects in these search windows merge in one single dot which is the most important word available. Even though this process significantly reduces multiple detection, in some cases windows can contain different parts of the object which results in different important word representation. In performance evaluation, only one of these multiple detections are counted as a true positive and the others are counted as false positives. Furthermore, if the detected available most important word does not fall in the ground truth mask, even there is an object in the search window the detection result is counted as false positive and the object center is counted as false negative. In conclusion, the performance evaluation is achieved in a harsh manner in order to utilize an automatic performance evaluation.

The performance tests are performed in three main scenarios:

- Utilization of PCA during classification
- Comparison of Harris and SIFT detectors in terms of detection performance
- Comparison of word weighting against the baseline algorithm

Figures 3.10-3.13 illustrate the comparison of the cases for which PCA is utilized or not for plane and ship object classes, respectively for both Harris and SIFT keypoint detectors. The figures contain the average, minimum and maximum performance curves for a chosen basis vector K in (32 for planes, 64 for ships) and no PCA analysis.

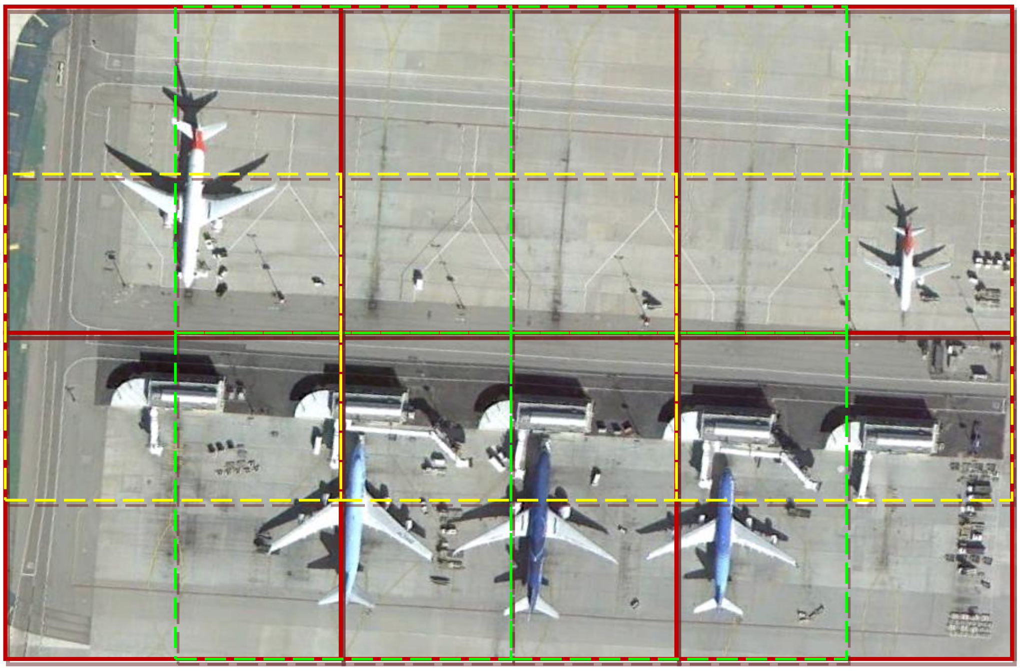


Figure 3.9: The search procedure

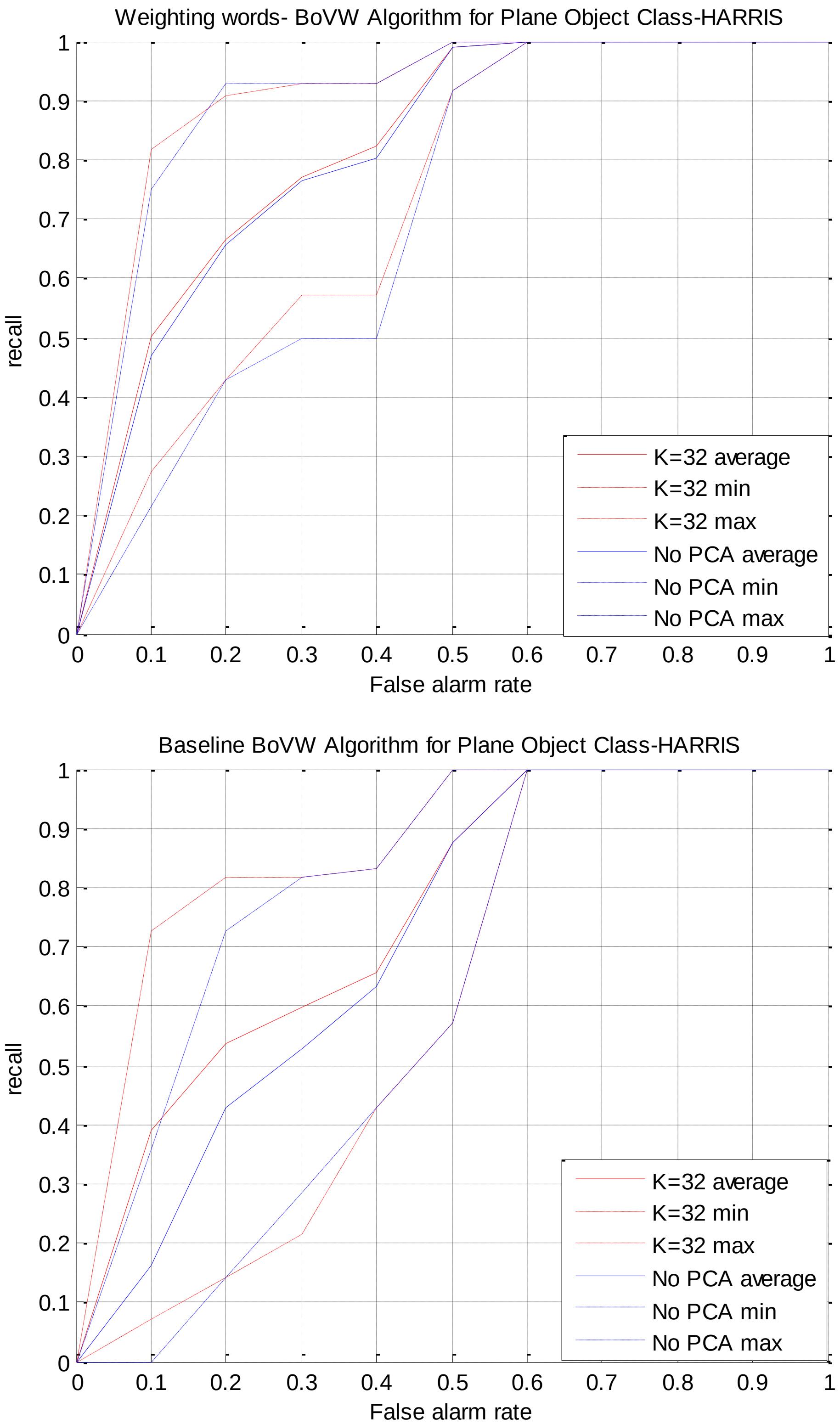
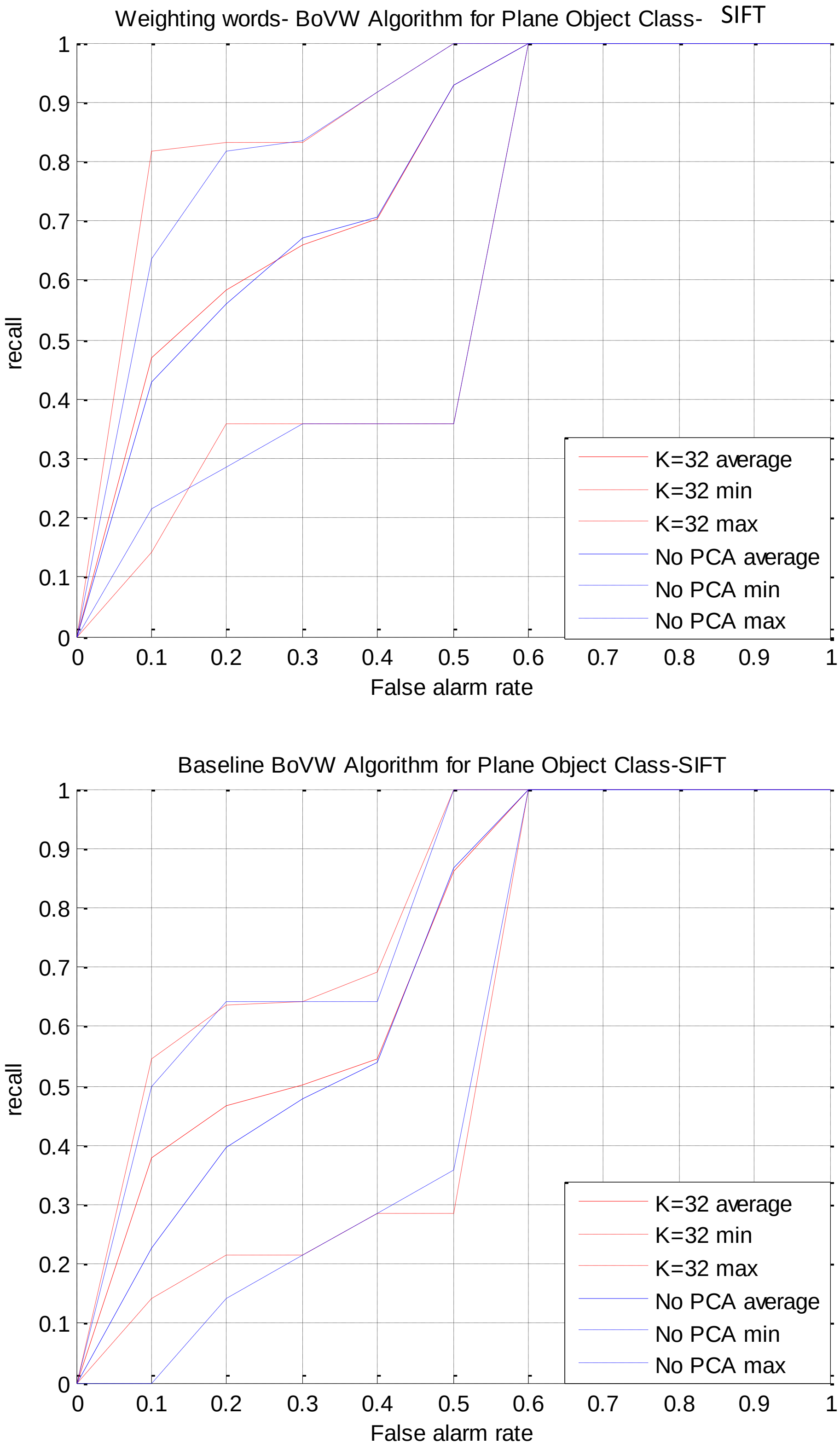


Figure 3.10: PCA comparison for Word Weighting and Baseline BoVW Algorithm using Harris keypoint detector for plane object class.


Weighting words- BoVW Algorithm for Plane Object Class- SIFT

Baseline BoVW Algorithm for Plane Object Class-SIFT

Figure 3.11: PCA comparison for Word Weighting and Baseline BoVW Algorithm using SIFT keypoint detector for plane object class.

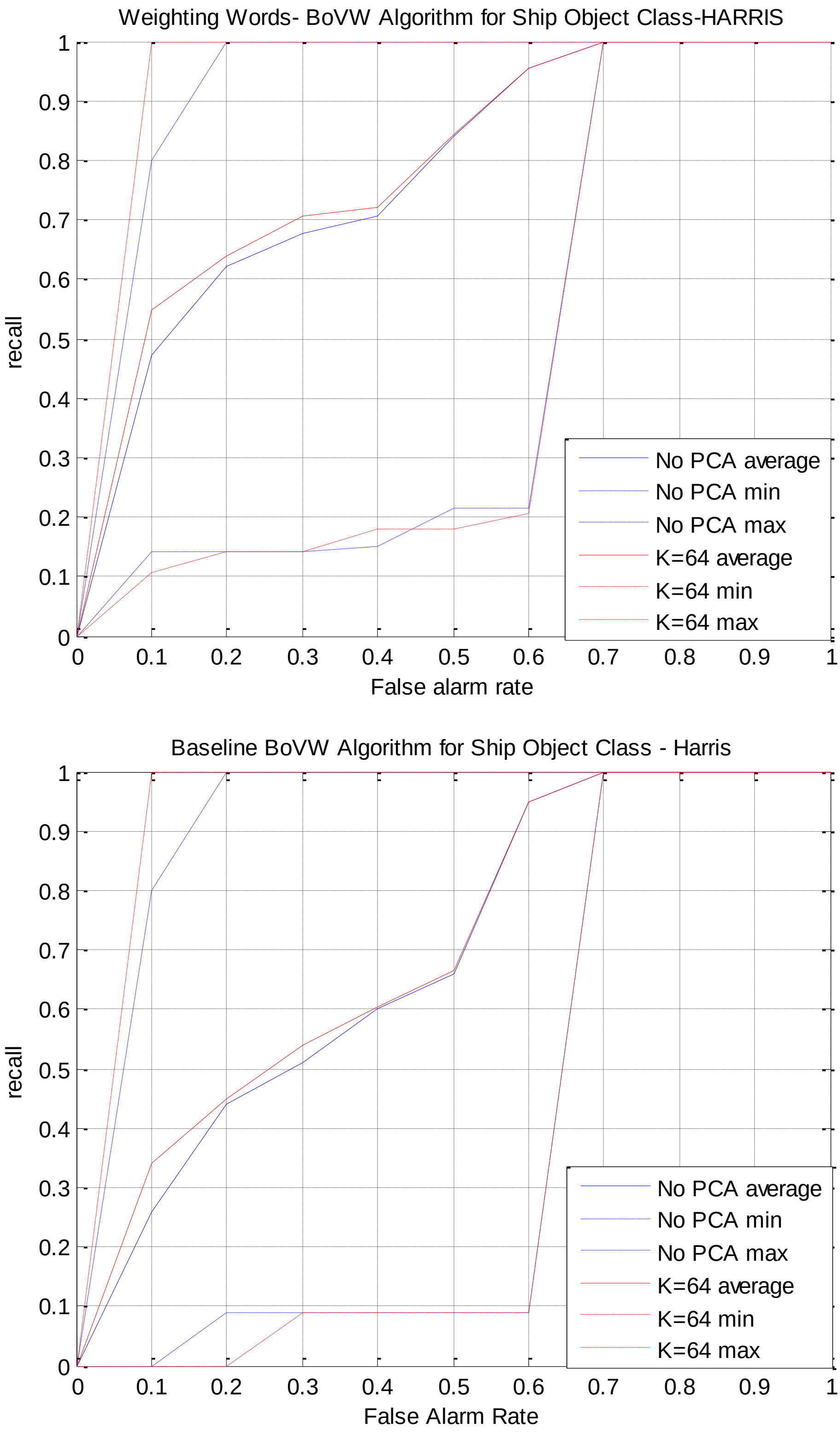


Figure 3.12: PCA comparison for Word Weighting and Baseline BoVW Algorithm using Harris keypoint detector for ship object class.

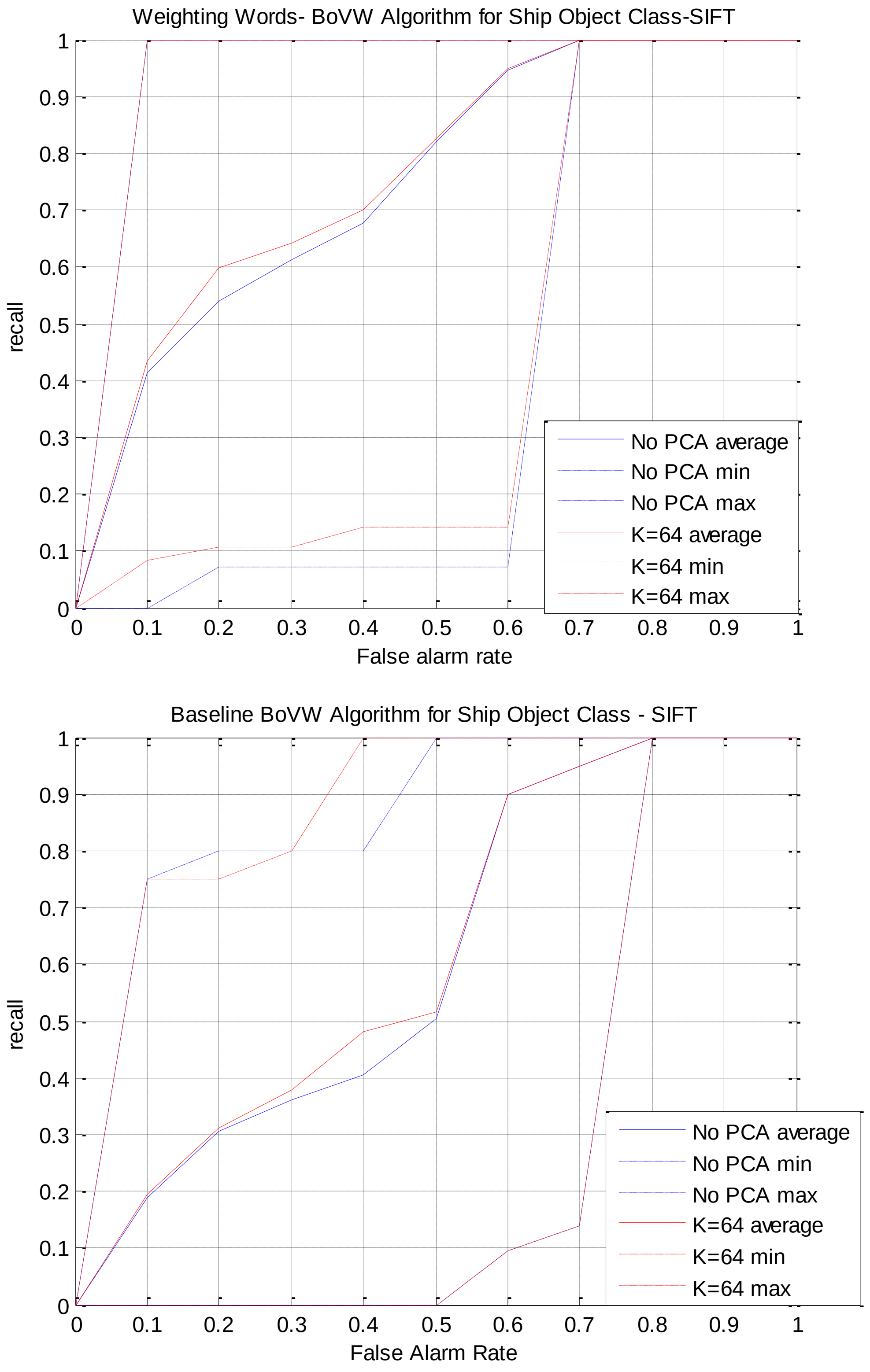


Figure 3.13: PCA comparison for Word Weighting and Baseline BoVW Algorithm using SIFT keypoint detector for ship object class.

As it can be interpreted from the above experiments that are presented by Figure 3.10 - 3.13, utilization of PCA increases the performance of the algorithm, since it eliminates the undesired small variation in the histogram or noise; i.e. background components hidden in the histograms. Although background information can be useful in some object detection algorithms, the bag of visual words algorithm is highly sensitive to this kind of redundancy, since it deals with histograms, and the histograms may change drastically with highly cluttered scenes. Thus, eliminating the noise is highly preferable via PCA.

Following figures focuses on the comparison of Harris and SIFT keypoint detectors for plane and ship object classes, respectively, using both weighting of words and baseline BOVW algorithm. The figures contain the average, minimum and maximum performance curves for a chosen basis vector K (32 for planes, 64 for ships) for PCA step.

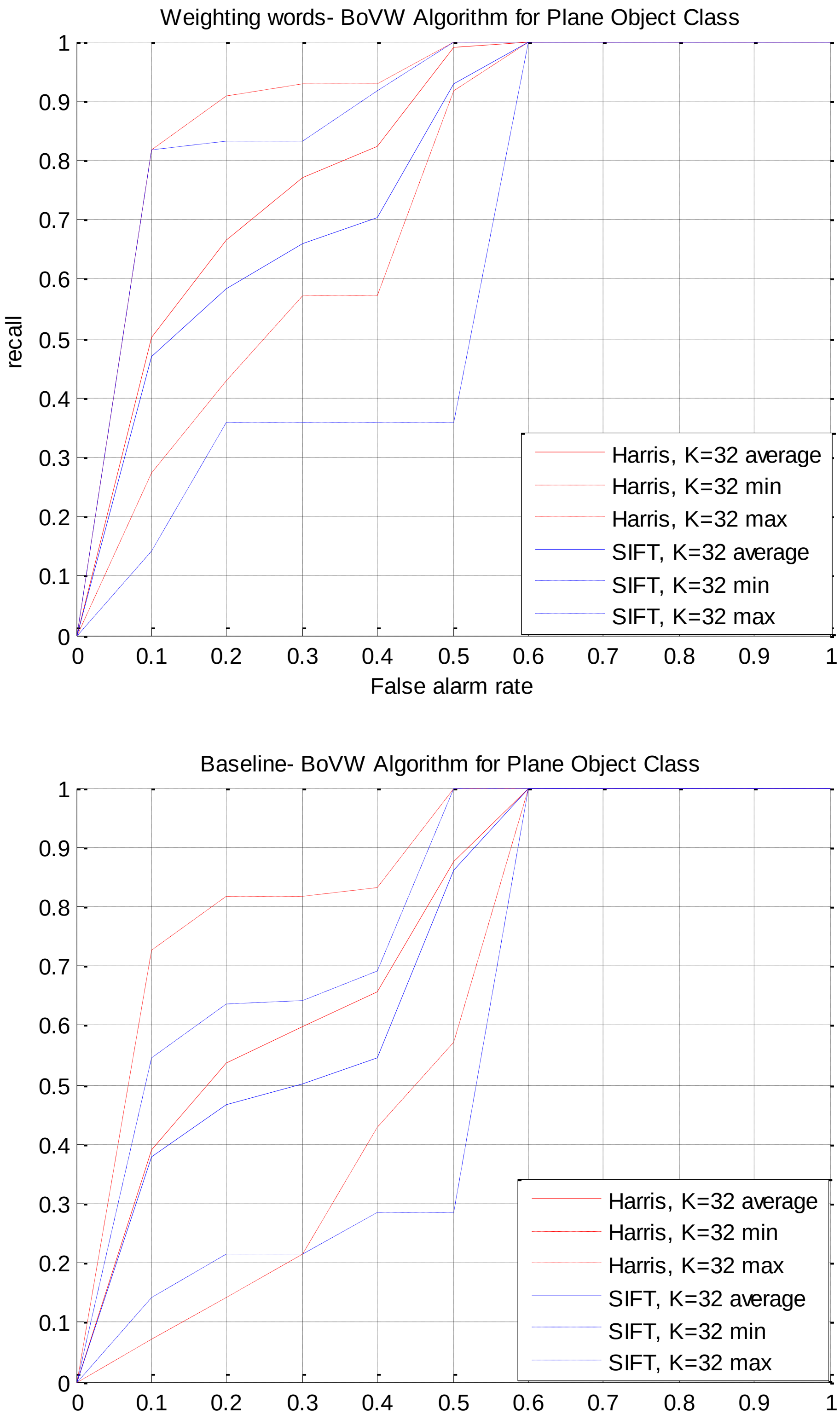


Figure 3.14: Keypoint detector comparison for Word Weighting and Baseline BoVW Algorithm using PCA (32 basis vectors) for plane object class.

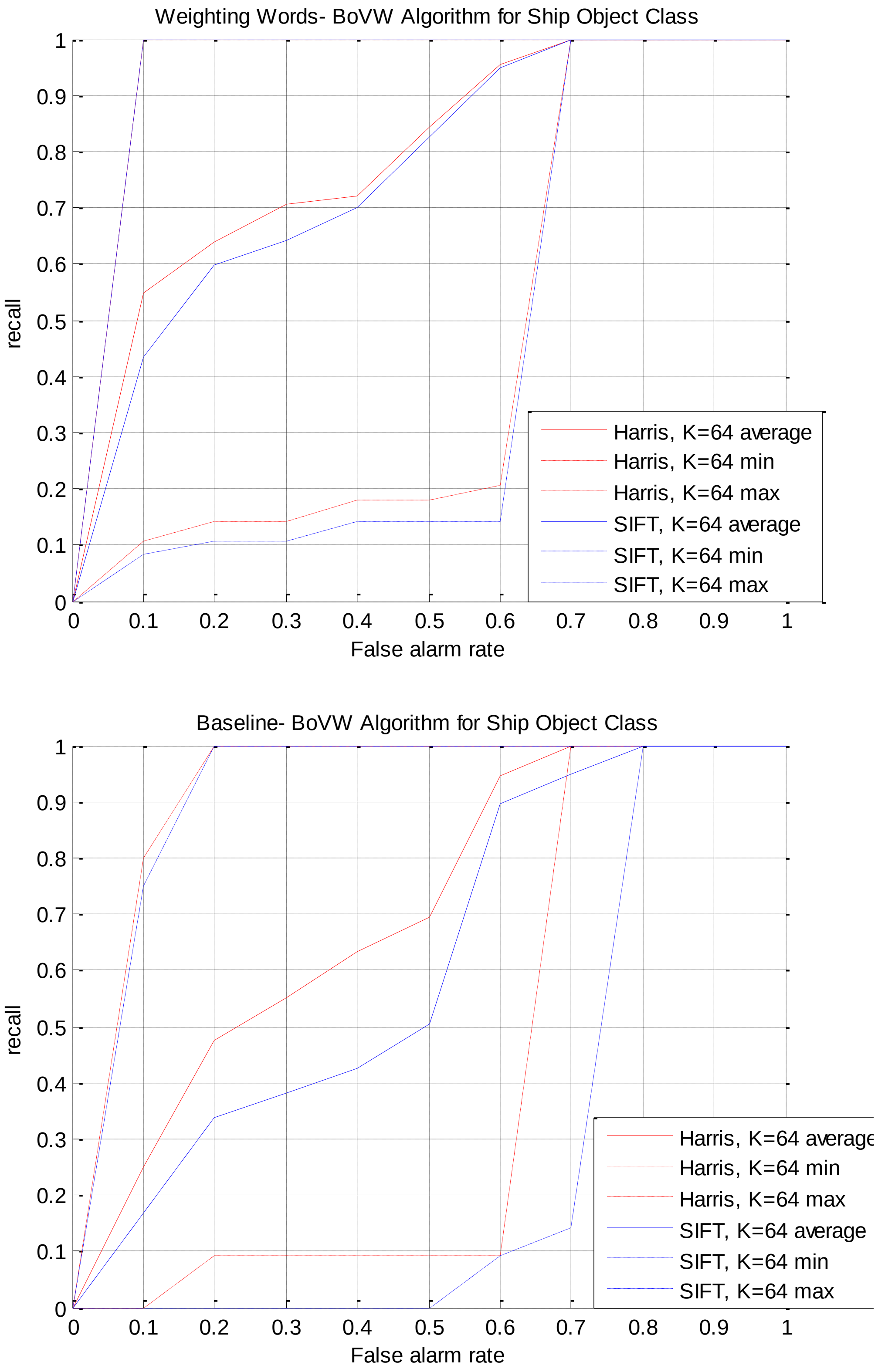
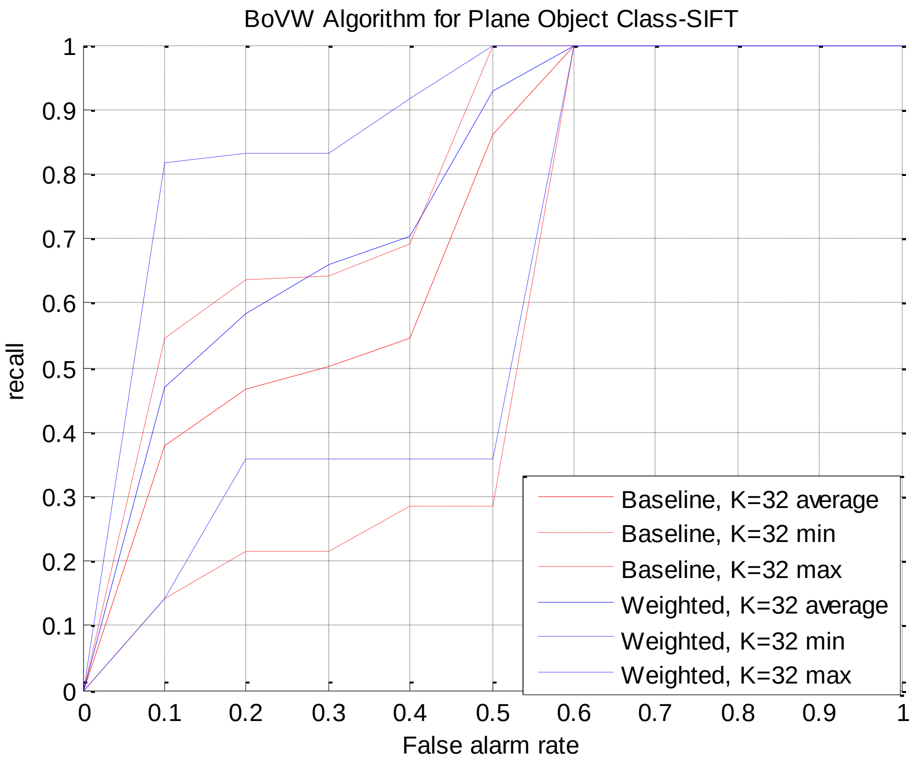


Figure 3.15: Keypoint detector comparison for Word Weighting and Baseline BoVW Algorithm using PCA (64 basis vectors) for ship object class.

As it can be examined and argued from the test results, Harris keypoint detector provides a better performance due to the reasons explained in detail in the beginning of this chapter.

Next figures summarize the comparison of baseline and weighted words BoVW algorithms for plane and ship object classes, respectively, by using both SIFT and Harris keypoint detectors. The figures contain the average, minimum and maximum performance curves for a chosen basis vector K (32 for planes, 64 for ships).



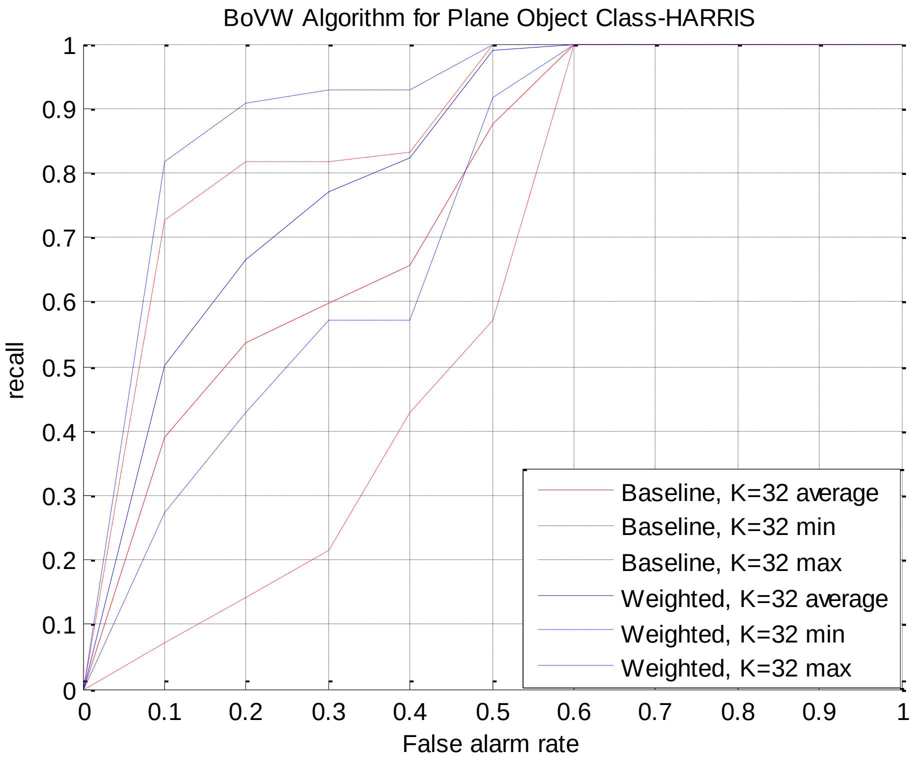
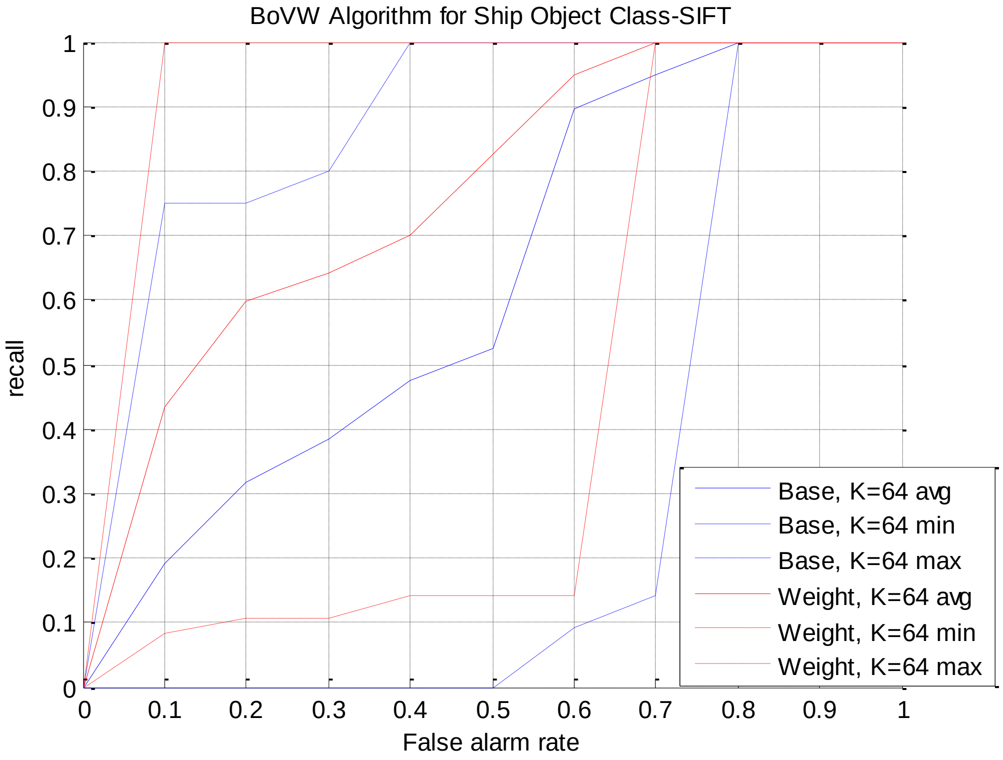


Figure 3.16: Algorithm type comparison for Word SIFT and Harris detector using PCA (32 basis vectors) for plane object class.



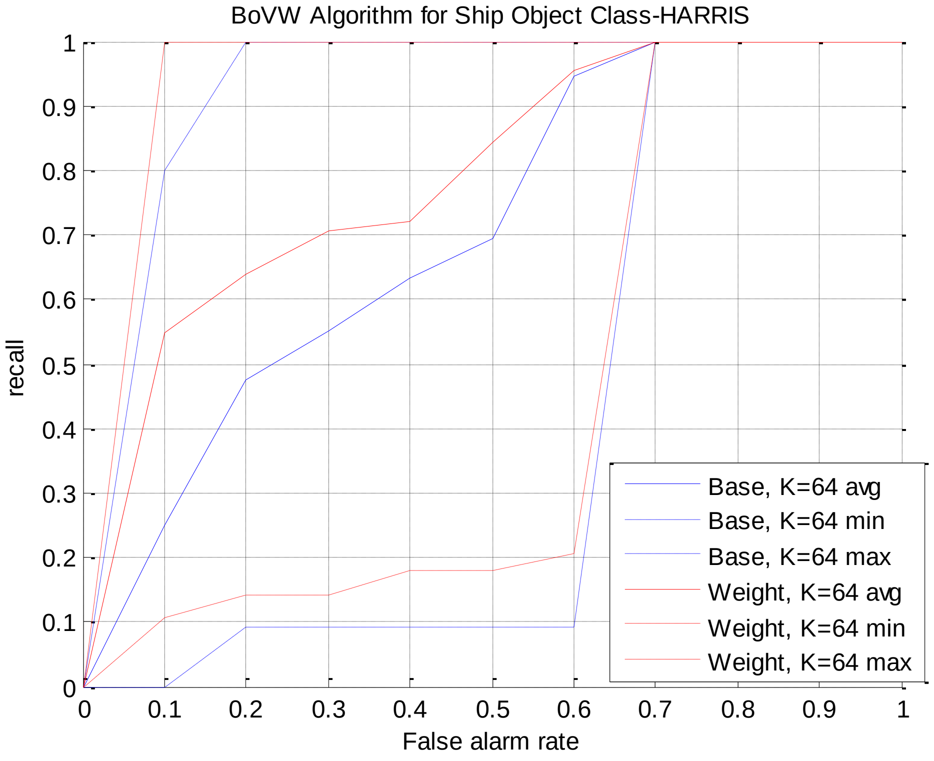


Figure 3.17: Algorithm type comparison for Word SIFT and Harris detector using PCA (64 basis vectors) for ship object class.

As it can be observed form the test results, weighting of words provides better performance, compared to the baseline BoVW algorithm as expected. Since weighting of words, thus, defining important words further reduces the effect of background on the histograms and provide better discriminability. Detailed analysis of the tests, the problems of BoVW algorithm and future work will be discussed in the next chapter.

CHAPTER 4

CONCLUSIONS

4.1 Summary of the Thesis

In this thesis, object detection problem is analyzed exploiting local features. A novel bag of visual words algorithm is proposed and different extensions of the algorithm are tested and compared for different object classes.

Classical Bag of words is a technique, whose main ideas are borrowed from text document analysis. Similar to the an analysis based on the occurrence of some text words in a document, the visual codewords that are obtained through quantization of the descriptions of keypoints on the image are examined in an image to state detection of a particular object.

In this thesis, two extensions to the classical bag of visual words algorithm are proposed. First of all, weighting visual words is presented and some experiments are performed in order to see the affects of weighting words before going into the classification step. Secondly, a Principal Component Analysis is proposed in order to remove the undesired redundancy and noise in the histogram shape

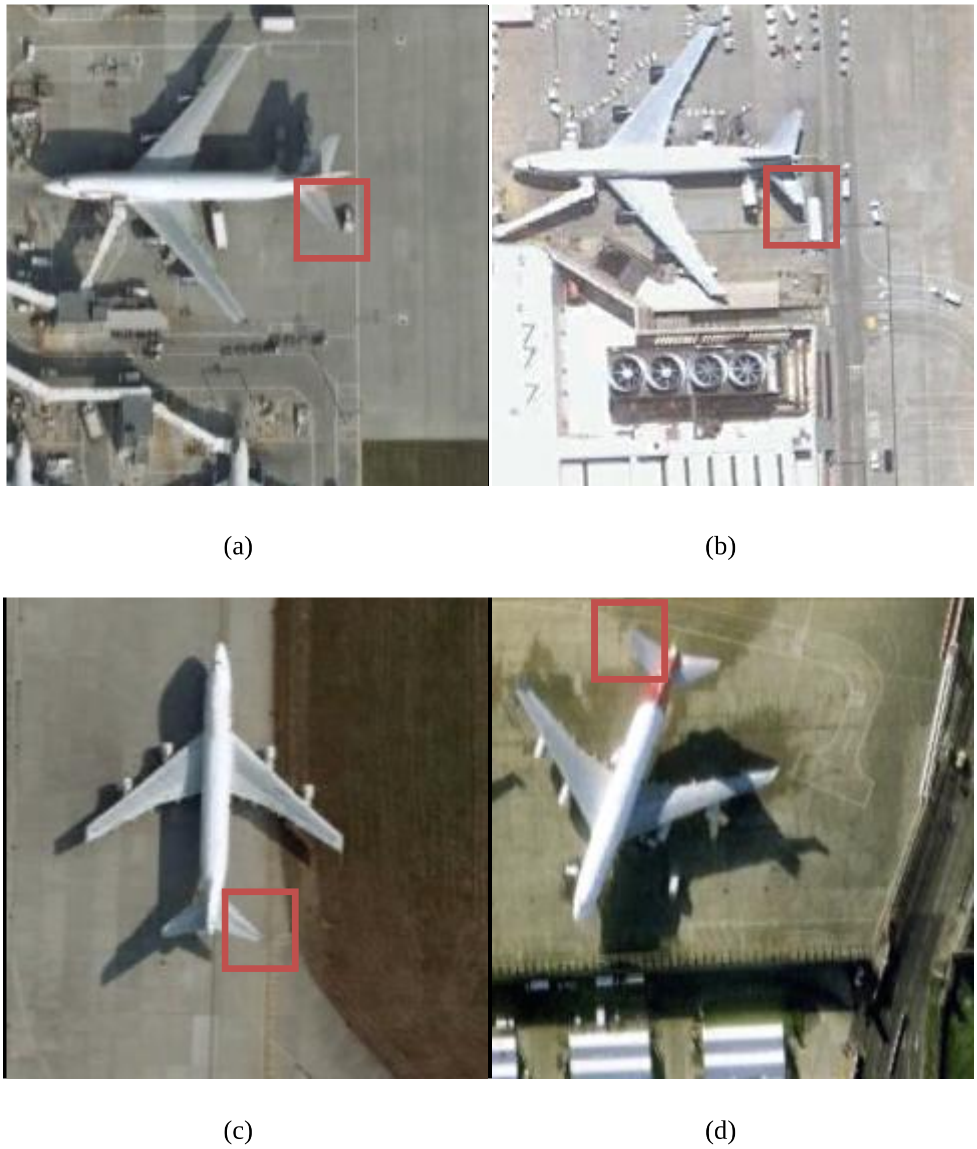
For the proposed algorithm, extensive experiments are performed; Harris and SIFT detector is compared in terms of repeatability, final performance in ROC curves with controlled experiments, principal component analysis is evaluated with comparing the ROC curves for this method and the ROC curves without PCA; finally, baseline algorithm and the algorithm with word weighting and soft histogram assignment is compared.

4.2 Discussions and Conclusions

The best performance results are achieved by Harris keypoint detector, weighting of words and principal component analysis. The reason behind this result for keypoint detector selection is due to the fact that Harris keypoint detector is more stable and reliable in terms of existence in consistent locations compared to SIFT keypoint. Moreover, corners are more distinctive features than blobs for many man-made object classes; and hence, an algorithm that is based on Harris keypoint, discriminates the object from the background better than it's SIFT counterpart.

The PCA and weighting of words both eliminates the noise in the histograms and provides histograms that are less affected by the background. Since BoVW algorithm deals with the distribution of the visual words in a patch of image, it is highly sensitive to noise, by proposed extensions BoVW algorithm is made more robust to background cluttering, thus providing a better performance.

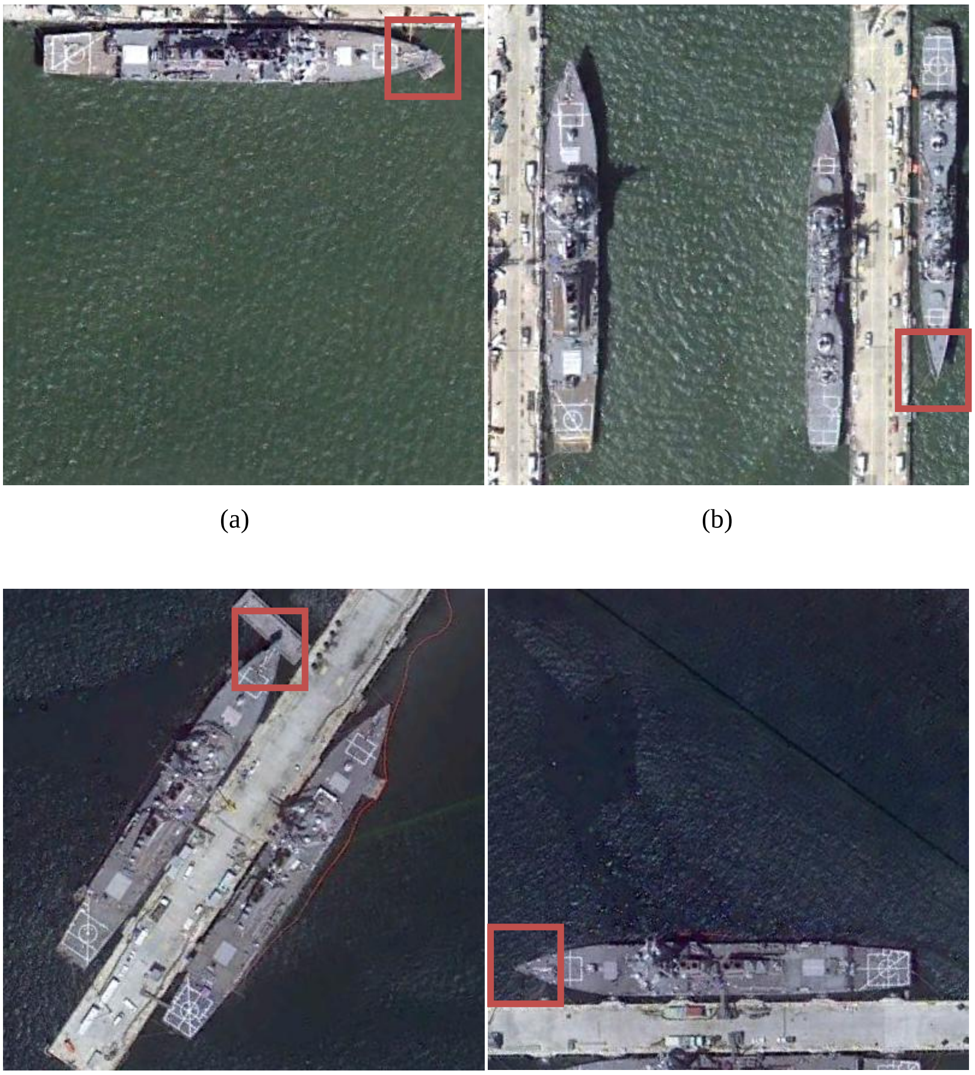
Although proposed extensions provide a great deal of performance increase, there are still some problems by some of the intermediate steps of the proposed algorithm. One of the main problems related with SIFT descriptor, is its shortcoming to describe the resulting features on different backgrounds. Since the descriptor is based on the gradient orientation histogram varying backgrounds can highly effect the orientation of the gradients. Some examples of varying backgrounds when trying to detect same regions of objects are given in Figures 4.1 and 4.2.



(a) (b)

(c) (d)

Figure 4.1: Illustration of the problem of background variation for describing the tail of airplanes. In (a) and (b) changing objects in the background and in (c) and (d) variations in the background color due to lines or stains are illustrated.



(a) (b)

(c) (d)

Figure 4.2: Illustration of the problem of background variation for describing the nose of ships. In (a) and (c) changing objects in the background and in (b) and (d) variation in the background color due to ropes or different sea colors are illustrated.

Another problem about the SIFT descriptor is its flaw while trying to describe a patch in case of a shadow. Shadow might completely change the gradient orientation histogram due to its very low intensity value. Some examples are presented for some parts of objects in Figure 4.3 and 4.4.

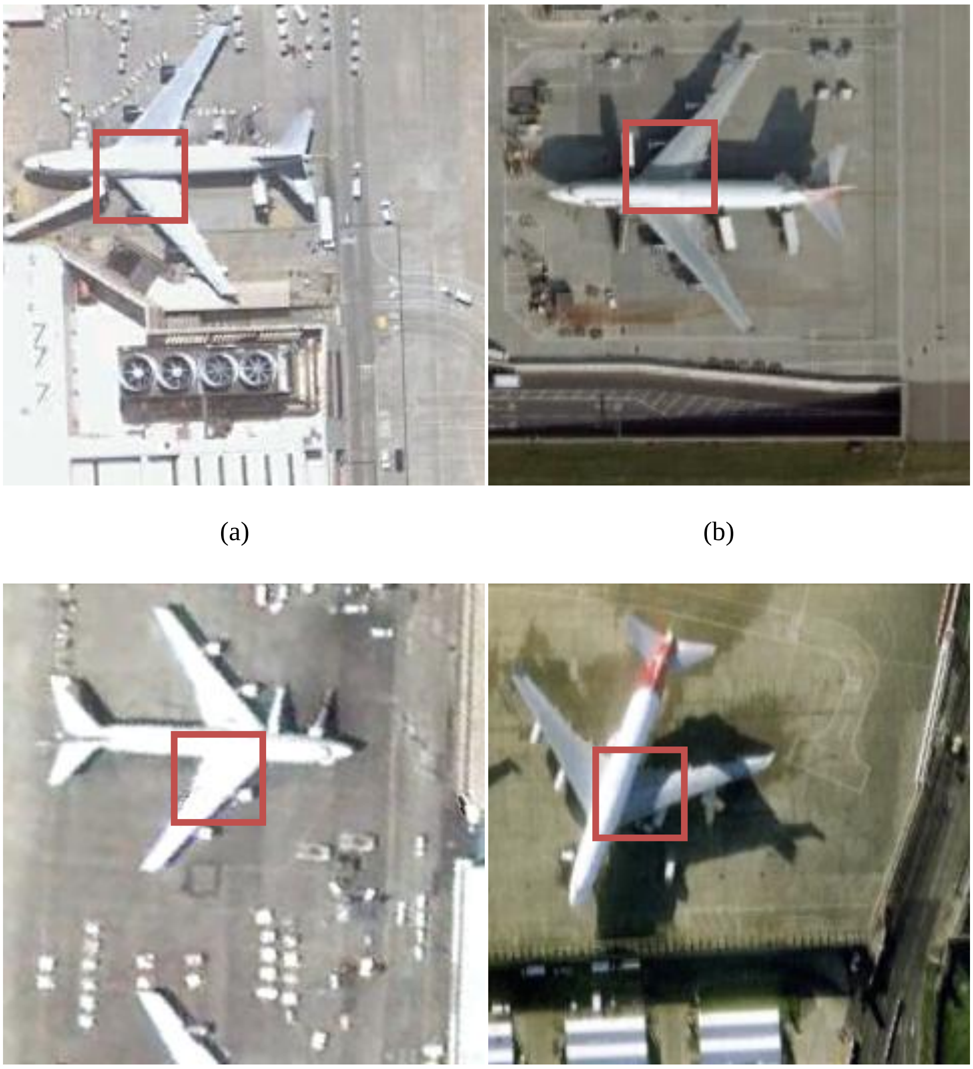
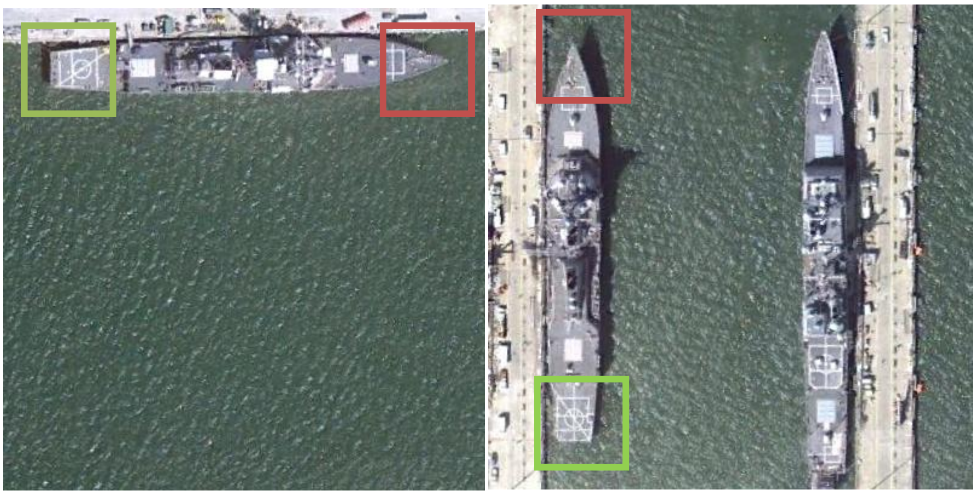


Figure 4.3: Illustration of the problem of self-shadow variation for describing the wing to body transition in airplanes.



(a) (b)

Figure 4.4: Illustration of the problem of cast shadow variation for describing the nose (red rectangle) and the stern of ships (green rectangle)

Another problem can be argued as the inter-class variations. An object class; for example, a ship can be quite different in various docks. Figure 4.5 and 4.6 illustrate this variance within the object classes. Since the proposed method suggests using of visual word histograms, obviously these histograms will change significantly, if the object type is somewhat different from the training data. In the light of this observation, one can say that BoVW algorithm is quite sensitive to intra-class variations.

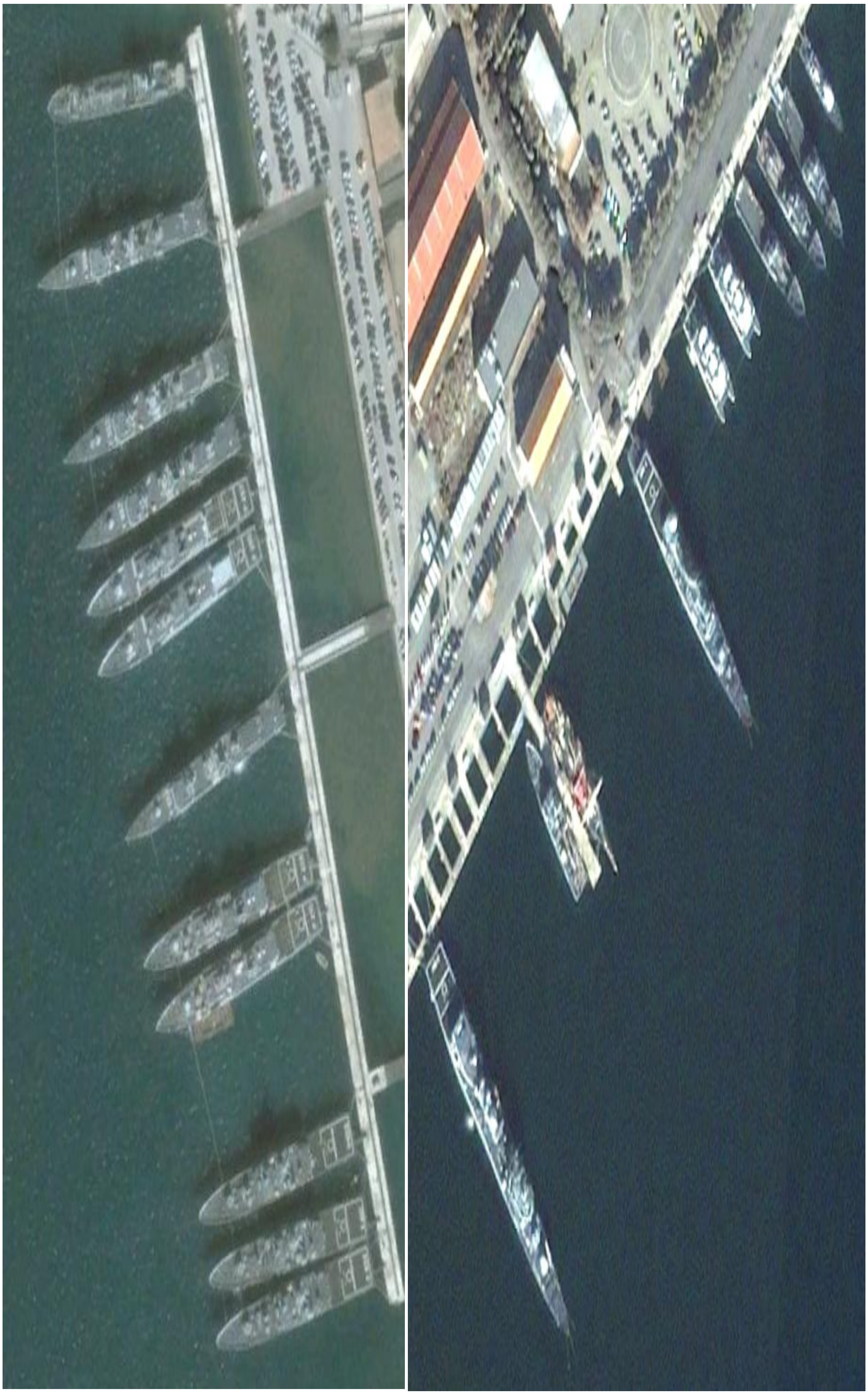


Figure 4.5: Illustration of inter-class variation within ships

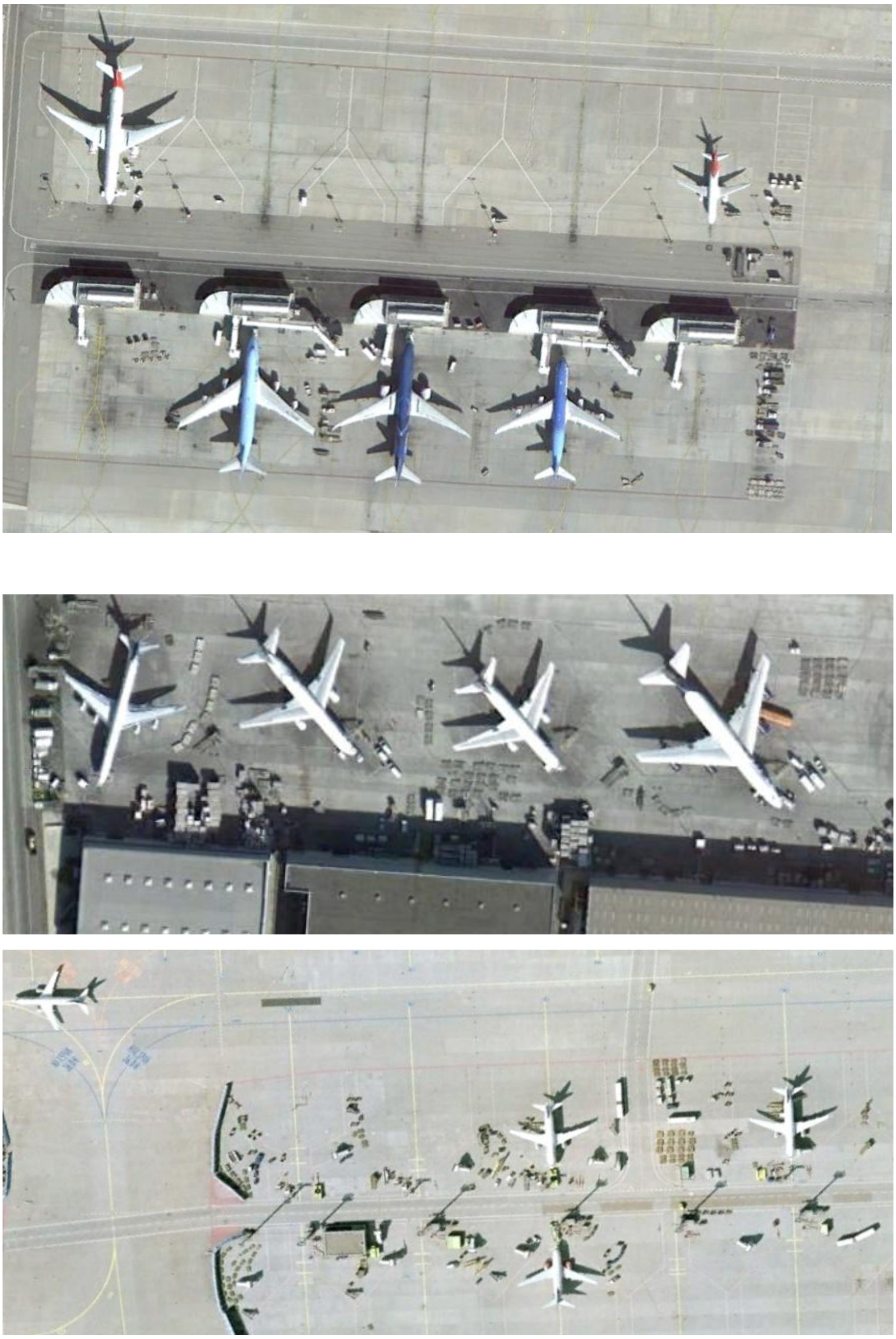


Figure 4.6: Illustration of inter-class variation within planes

4.3 Future Work

The problems about shadows can be handled in many ways. A shadow detection and restoration technique may be implemented as a preprocessing step before detection of SIFT or Harris features. This idea can be discussed in terms of computational cost; in addition, the performance of the shadow restoration algorithm highly depends on the shadow detection part, which is still an unsolved problem in the literature. Another solution to be proposed may be merging Harris, SIFT and many other keypoint detection results.

The inter-class variance problem can be tried to be solved by training different classifiers for each type within an object class; then, while trying to detect the object, all of these classifiers can be executed and the results can be merged. Since the false alarm rate is predicted to decrease, this may seem as an algorithm to reduce the precision, but a very high increase in recall can compensate the fall in performance.

Finally, adding spatial extensions could provide exploiting the structure of the object to be detected.

REFERENCES

[1] X. Sun, H. Wang and K. Fu, “Automatic Detection of Geospatial Objects Using Taxonomic Semantics”. *Geoscience and Remote Sensing Letters, IEEE.* Vol. 7 pp. 23-27, 2010.

[2] C. Tao, Y. Tan, H. Cai and J. Tian, “Airport Detection From Large IKONOS Images Using Clustered SIFT Keypoints and Region Information”. *Geoscience and Remote Sensing Letters, IEEE.* Vol. 8 pp. 23-27, 2011.

[3] X. Perrotton, M. Sturzel and M. Roux, “Automatic Object Detection on Aerial Images Using Local Descriptors and Image Synthesis”. *Proceedings of the 6th international conference on Computer Vision, ICVS,* 2008.

[4] H. Cai and Y. Su, “Airplane Detection in Remote Sensing Image with a Circle-frequency Filter”. *International Conference on Space Information Technology,* 2006.

[5] J.W. Hsieh, J.M. Chen, C.H. Chuang and K.C.Fan, “Aircraft Type in Satellite Images”, *In Vision, Image and Signal Processing, IEEE Proceedings-*, vol.152, pp. 307-315, 2005.

[6] J. Iisaka, T.S. Amano, “A Shape-based Object Recognition for Remote Sensing”, *Geoscience and Remote Sensing Symposium, IGARSS,* 1995.

[7] C. Harris and M. Stephens. "A combined corner and edge detector". *Proc. of the 4th Alvey Vision Conference*. pp. 147–151,1988

[8] D. G. Lowe. “Distinctive image features from scale-invariant keypoints”. *Int. J. Comput. Vision*, 60(2):91-110, 2004.

[9] J. Matas, O. Chum, M. Urba, and T. Pajdla. "Robust wide baseline stereo from maximally stable extremal regions." *Proc. of British Machine Vision Conference, pages 384-396,* 2002.

[10] J. Sivic and A. Zisserman. “Video google: A text retrieval approach to object matching in videos”. *In proc. Of the 9th IEEE Int’l Conf. on Computer Vision, Vol. 2,* 2003.

[11] K. Grauman and T.Darrell, “Approximate correspondances in high dimensions”, *NIPA, pp. 505-512,* 2006

[12] Y. Ke, R. Sukthankar, and L.Huston, “An efficient parts-based near-duplicate and sub-image retrieval system”, *ACM Multimedia, pp. 869-876,* 2004.

[13] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree”, *CVPR, 2006, vol2, pp 2161-2168.*

[14] F.-F.Li and P.Perona. “A Bayesian hierarchical model for learning natural scene categories”. In proc. Of the 2005 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition, pp 524-531, 2005.

[15] J. Vogel and B. Schiele . “On Performance Characterization and Optimization for Image Retrieval. *Proc. of European Conference on Computer Vision”*. pp. 51–55., 2002.

[16] K. Mikolajczyk and C. Schmid "A performance evaluation of local descriptors", *IEEE Transactions on Pattern Analysis and Machine Intelligence, 10, 27, pp 1615--1630*, 2005.

[17] T. Lindeberg, “Feature Detection with Automatic Scale Selection”, *International Journal of Computer Vision, vol. 30, No. 2*, 1998.

[18] T. Lindeberg, “Scale-space theory: A basic tool for analyzing structures at different scales”, Journal of Applied Statistics, vol. 21, No. 2, pp. 224-270, 1994.

[19] K. Mikolajczyk, “Detection of local features invariant to affine transformations”, Ph.D. thesis, Institut National Polytechnique de Grenoble, France, 2002.

[20] T. Leung and J. Malik. "Representing and recognizing the visual appearance of materials using three-dimensional textons". *International Journal of Computer Vision* 43 (1): 29–44 , 2001.

[21] J.C. van Gemert, C.J. Veenman, A.W.M. Smeulders, J.M. Geusebroek,” Visual Word Ambiguity”, *IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.32, no. 7,* 2010.

[22] Y.Cao, C.Wang, Z. Li, L.Zhang, L. Zhang, Spatial Bag of features, *CVPR '10. IEEE Conference on Computer Vision and Pattern Recognition,* 2010.

[23] V. Viitaniemi, J. Laaksonen, Spatial Extensions to Bag of Visual Words. *CVPR ’09 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[24] S. Lazebnik, C. Schmid, J. Ponce, Beyond Bag of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, *CVPR '06 Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, 2006

[25] T. Kobayashi, N. Otsu, Bag of Hierarchical Co-occurrence Features for Image Classification, *CVPR '10 Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010

[26] Y. Zhang, T. Chen, Weakly Supervised Object Recognition and Localization with Invariant High Order Features, *Proceedings of the British Machine Vision Conference*, 2010

[27] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid (2001). "Local Features and Kernels for Classification of Texture and Object Categories: a Comprehensive Study". *International Journal of Computer Vision* 73 (2): 213–238.

[28] W.Zhao, Y.-G. Jiang, and C.-W. Ngo. "Keyframe retrieval by keypoints: Can point-to-point matching help?" In *Proc. Of 5th Int'l Conf. on Image and Video Retrieval (CIVR),* pages 72-81, 2006.

[29] G. Salton and C. Buckley. "Term weighting approaches in automatic text retrieval". *Information Processing and Management: ant Int'l Journal*, 25(5):513-523, 1988.

[30] R.Baeza-Yates and B.Riberio-Neto. "Modern Information Retrieval". *ACM Press Series / Addison Wesley,* 1999.

[31] W.Zhao, Y.-G. Jiang, and C.-W. Ngo. "Keyframe retrieval by keypoints: Can point-to-point matching help?" In *Proc. Of 5th Int'l Conf. on Image and Video Retrieval (CIVR),* pages 72-81, 2006.

[32] R.o. Duda, P.E. Hart and D.G. Stork, "Pattern Classification" , 2nd Edition, pp. 567-568, 2000.

[33] V. Vapnik, and A. Lerner, "Pattern recognition using generalized portrait method". *Automation and Remote Control*. pp. 774–780, 1963

[34] B.E. Boser, I.M. Guyon, and V.N Vapnik, "A training algorithm for optimal margin classifiers". *COLT '92: Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. pp. 144–152, 1992.

[35] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection". *CVPR '05 Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol 1, 2005.

[36] A. F. Smeaton, P. Over and W. Kraaij. "Evaluation campaigns and TRECVid". *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR),* pp. 321-330, 2006.

[37] M. Everingham, L. Gool, C. K. Williams, J. Winn and A. Zisserman, "The Pascal Visual Object Classes (VOC) Challenge". *International Journal of Computer Vision*, vol. 88, Issue 2, 2010.

[38] D. Arthur and S. Vassilvitskiik, "k-means++: The Advantages of Careful Seeding". *Prooceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. pp. 1027–1035, 2007.

[39] K. Mikolajczyk, B. Liebe and B. Schiele, “Local Features for Object Class Recognition”. *Tenth IEEE International Conference on Computer Vision.* 2005.