2018

Image Classification Using Bag-of-Visual-Words Model

Kaiqiang Huang
Technological University Dublin

Image Classification using Bag-of-Visual-Words model

Kaiqiang Huang

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Data Analytics)

January 2018

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institutes guidelines for ethics in research.

Signed:

Date: 03/01/2018

Abstract

Recently, with the explosive growth of digital technologies, there has been a rapid proliferation of the size of image collection. The technique of supervised image classification has been widely applied in many domains in order to organize, search, and retrieve images. However, the traditional feature extraction approaches yield the poor classification accuracy. Therefore, the Bag-of-visual-words model, inspired by Bag-of-Words model in document classification, was used to present images with the local descriptors for image classification, and also it performs well in some fields.

This research provides the empirical evidence to prove that the BoVW model outperforms the traditional feature extraction approaches for both binary image classification and multi-class image classification. Furthermore, the research reveals that the size of the visual vocabulary during the process of building BoVW model impact on the accuracy results of image classification.

Keywords: Image processing, Bag-of-visual-words, Image classification, Supervised machine learning

Acknowledgments

I would first like to thank my thesis supervisor Prof. Sarah Jane Delany in the School of Computing at Dublin Institute of Technology. I cannot complete this thesis at Master's level without her powerful support, flexible arrangement, patient guidance, comprehensive suggestion and encouragement.

I would like to thank my parents for their support, care and love during my thesis, even during my Master's study.

Lastly, I would to thank my loving girlfriend Dan Xu, who always encourages me to pursue the Master degree, and supports me at all time.

Contents

List of Figures

List of Tables

List of Acronyms

BD(i)	The ith dataset for binary classification experiment
BoW	Bag-of-Words
BoVW	Bag-of-Visual-Words
DoG	Differences of Gaussian
RGB	Red, Green and Blue
FP	False Positive
FN	False Negative
HOG	Histogram of Oriented Gradients
HSV	Hue, Saturation and Value
HSR	High Spatial Resolution
LBP	Local Binary Patterns
LoG	Laplacian of Gaussian
MRI	Magnetic resonance imaging
MDA(i)	The ith dataset in Group A for Multi-class classification experiment
MDB(i)	The ith dataset in Group B for Multi-class classification experiment
SIFT	Scale-Invariant Feature Transform
SURF	Speed Up Robust Feature
TP	True Positive
TN	True Negative

Chapter 1

Introduction

1.1 Background

Over the last decades, with the development of the Internet and social media, the increasing number of images has been generated and studied using methods to acquire, process, analyze, and understand in the computer vision community. One of the most key subfields in computer vision is image classification, which copes with constructing systems that attempt to identify objects represented in images. However, the task of image classification is a complicated process, and it is difficult to gain the high accuracy by the supervised machine learning algorithms (Kurian & Karunakaran, 2012). For examples, the effect of illumination is sensitive to the pixel level that could cause the significant variations in the intensity of the pixels. And also, the visual objects often exhibit variation for their sizes in the real world, and the most of objects do not have the rigid feature that can be deformed in extreme ways. Therefore, the main challenge of image classification is to find out the feature representation of the images, which are the vectors of feature extracted by images.

In the earlier work(Torralba, Fergus, & Freeman, 2008), the feature of raw pixel was regarded as one of the most straightforward possible image representation. However, it could discard all of the high-frequency image features, resulting in the poor accuracy for image classification. Furthermore, the color histogram with RGB color space is one of the oldest known representation approaches for image classification (Swain &

Ballard, 1991). Similar to the feature extraction approach of raw pixel, it still does not provide a significant improvement for image classification.

The Bag-of-Words (BoW) model (Z. S. Harris, 1954) has been successfully applied in the field of document classification and text categorization where the occurrence of each is used as a feature for training a classifier. As the motivation, the state-of-the-art approach, called Bag-of-Visual-Words (BoVW) model, was proposed by Csurka, Dance, Fan, Willamowski, and Bray (2004) for image representation with the SIFT descriptors used in supervised image classification. Similar to the process of BoW, the local descriptors, which are extracted from the regions of interest, are clustered to a vector, which is called a visual word, and many visual words are combined as the visual vocabulary.

1.2 Research project

As introduced in the background, the performance made by traditional feature extraction approach and BoVW model is theoretically different. Therefore, as a motivation, the aim of the research is to compare traditional feature extraction techniques, namely raw pixel and color histogram, to the BoVW model with the SIFT descriptor and SURF descriptor for the supervised image classification. Then, the research question is stated as follows.

Can the feature extraction approaches of SIFT and SURF with the BoVW model outperform the feature extraction approaches of raw pixel and color histogram for image classification using the linear SVM algorithm ?

According to the defined research question, the objective of the research is to determine whether the BoVW model can produce the greater accuracy than traditional feature extraction approach for image classification. Some experiments will be performed in order to fulfill the aim of this research and obtain the high-quality results.

1.3 Research methodologies

This research focuses on the comparison of BoVW feature extraction for image classification to more traditional techniques with the existing data source, and therefore, it belongs to secondary research. According to the secondary research, the existing literature about supervised image classification is reviewed and studied.

The methodology of this research belongs to empirical and quantitative research. The designed experiments will be performed to yield the expected results in order to answer the proposed research question. Moreover, the chosen statistical test will be conducted to prove the defined hypotheses. In addition, this research will be made conclusion based on the results generated by the experiments, so it belongs to inductive.

1.4 Document outline

This research contains four more chapters and the relevant overview is outlined below for each chapter.

Chapter 2(Review of existing literature) provides the existing literature review about the contents of image processing with global and local feature, BoVW model introduction and its related work in the different fields, image classification introduction and evaluation measurement, and statistical test methods.

Chapter 3(Experiment design and methodology) provides the design of three experiments in details, which are the experiments of binary classification, multi-class classification, and selecting vocabulary size. The data that will be used in the experiments is provided and analyzed. Also, the methodologies of the approaches, evaluation, and statistical test are presented and discussed. Also, the experimental software design is provided at a high level for conducting each experiment and obtaining all results.

Chapter 4(Experimentation and results) provides the details about the implementation, results, and statistical test for each experiment. Furthermore, the deep discussion and key findings are presented and analyzed based on the given results.

Chapter 5(Conclusion) concludes the summary of results and findings in this research. And also, it provides the general description for each conducted experiment. Moreover, the limitation of this research and the future work are presented.

Chapter 2

Review of existing literature

This chapter presents the detailed literature review to introduce the image processing, and also describe the related work about the field of image classification, especially using BoVW pattern. Also, this project investigates how the BoVW approaches compare to the more traditional approaches of feature representation for image analysis in the area of image classification. Therefore, at first, section 2.1 presents the essential knowledge and literature of image processing at a high level, including global feature and local feature. Then, the brief histories, motivations, and developments of BoVW in many industries are presented in section 2.2. After that, section 2.3 presents the overview of image classification with supervised and unsupervised learning approaches and the evaluation methodologies. Lastly, the introduction of statistical testing is shown in details in section 2.4.

2.1 Image processing

All image analysis requires representing an image as a vector of features that represent some aspects of the image. There are a large variety of ways to extract and detect features from images used by the computer vision community. These vary from the most straightforward gray-scale representation and color histograms to more complex BoVW approaches. They are used in a variety of applications, such as image classification and retrieval system (Stottinger, Hanbury, Sebe, & Gevers, 2012; Liu & Bai,

2012), robot navigation and mapping system (Nicosevici & Garcia, 2012) and object recognition and matching system (Dollar, Wojek, Schiele, & Perona, 2012; Miksik & Mikolajczyk, 2012).

2.1.1 Global feature

The global feature is proposed to describe an image through the whole perspective, and it is interpreted as a distinctive feature of the image with each pixel. In the global feature representation, the image is represented by the multidimensional feature vectors where describe the whole image. In a nutshell, the approach of global feature generates a single vector with values, measured by different aspects of the images, such as color, texture, and shape. Furthermore, the advantages of global features are that they are much faster and easier to compute, and require small amounts of memory than local feature's requirement. Moreover, it also has some limitations as they are not invariant to significant transformations and sensitive to clutter and occlusion (Hassaballah, Abdelmgeid, & Alshazly, 2016).

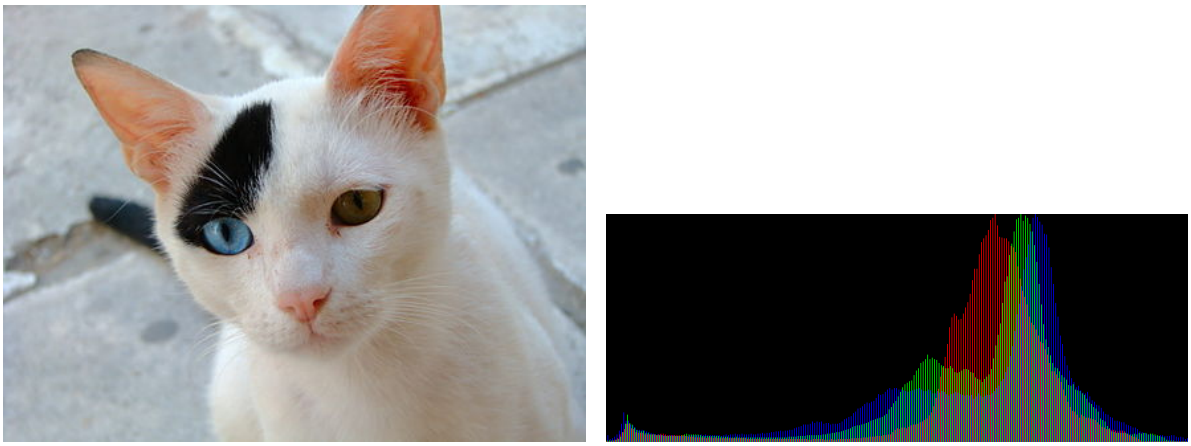
Raw pixel

The feature of the raw pixel regarding as the global feature is inspired by the research (Torralba et al., 2008). It is one of the most straightforward possible image representations based on the proposed tiny images. It works slightly better if the tiny image is made to have zero mean and unit length. This is not a particularly good representation because it discards all of the high-frequency image content and is not especially shift invariant. Torralba et al. (2008) proposed several alignment methods to alleviate the latter drawback. It demonstrates that the simple non-parametric methods, along with the tiny image dataset, can give reasonable performance on object classification.

Color histogram

Color histogram is one of the oldest known global features used in image processing. The early work proposed to use color histograms with RGB (Red, Green, and Blue)

color space in image retrieval (Swain & Ballard, 1991). However, RGB model doesn't correspond to the way humans perceive color (Chatzichristofis, Zagoris, Boutalis, & Papamarkos, 2010; Sural, Qian, & Pramanik, 2002). However, HSV color space is explicitly designed to model human color perception, and is therefore used in most papers on histograms as a global feature. Another problem is that the color histogram has high sensitivity to noise interference, such as illumination intensity change and quantization error, and also the high dimensional color histogram is also another problem (Wang, Wu, & Yang, 2010). Some color histogram feature spaces usually take up more than one hundred dimensions. The color space of HSV (Hue, Saturation, and Value), therefore, is widely used to apply on histograms as the global feature to match the human color perception (Stricker & Orengo, 1995).



(a) An odd-eyed cat

(b) RGB-Histogram of the Odd-eyed cat

Figure 2.1: An example of color histogram representation (Commons, 2016)

Furthermore, a color histogram only concentrates on the proportion of the number of different types of colors, regardless of the spatial location of the colors. The values of a color histogram are from statistics. They illustrate the statistical distribution of colors and the essential tone of an image. For the further study, the relationship between color histogram data and physical properties of objects in the image, showing that they cannot only represent the color and illumination of objects, but also relate to the surface roughness and image geometry, and provide an improved estimation of illumination and object color (Novak & Shafer, 1992). The figure 2.1 shows the

example of an odd-eyed cat image and its RGB-based color histogram.

As discussed above, the color histogram is generated by RGB color space, which has the drawbacks as well. Another study presented that using a uniform color space can deliver the better retrieval performance, such as CIE $L^*a^*b^*$, namely Lab (Konstantinidis, Gasteratos, & Andreadis, 2005). In Lab color space, the term of L stands for the lightness of the color as 0 producing black and 100 producing a diffuse white. The term of a means the comparison between redness and greenness, then the term of b means the comparison of yellowness and blueness. However, the conversion from RGB to Lab is computationally expensive due to the calculation of cubes root. In a word, the main disadvantage of the histogram for classification is to represent the color of the object studied and ignored its shape and texture. The color histogram could be in the situation that two same images have the different object contents just to share the color information. On the contrary, without space or shape information, similar objects of different colors based on the comparison of the color histogram may not be distinguished.

Texture

Texture, treated as useful features for images, is commonly used in human visual systems for recognition and interpretation (ping Tian et al., 2013). In literature, a large number of techniques have been proposed to extract texture features where the texture feature is extracted and classified into the feature extraction approaches of spatial texture and spectral texture (Zhang, Wong, Indrawan, & Lu, 2000; WANG & Shi, 2006). For the former approach, texture features are calculated by the pixel frequencies or finding the local pixel structures in the original image domain, while the latter transforms an image into the frequency domain, and then computes features from the transformed images. Furthermore, the most well-known approach for texture feature extraction, called Gabor filter has been widely used in image texture feature extraction (Manjunath & Ma, 1996). Moreover, the Gabor filter was proposed to sample the entire frequency domain of an image by characterizing the center frequency and orientation parameters.

2.1.2 Local feature

Local feature representation aims to particularly describe the images based on regions of interest while remaining invariant to viewpoint and illumination changes. The images, therefore, are represented according to the local property by the local feature descriptors. In comparison, the local features provide the even higher performance that global feature's (Jegou et al., 2012). The process of extracting local feature contains two primary stages that are feature detection and feature description as following.

Feature detection

Computing of Laplacian-of-Gaussian (LoG) that is a linear combination of second derivatives is a memory-dependent and time-consuming process. To speed up the process, Lowe (2004) proposed the state-of-the-art approach based on local 3D extrema in the scale-space pyramid, along with Difference of Gaussian (DoG) filters. The DoG is an analogy to LoG. Hence, the type of features extracted by DoG can be treated as the same type of features as LoG. However, they have the typical limitation that is the local maxima can be detected by the area of straight edges, leading to the issues of sensitivity on outliers or light changes (Mikolajczyk & Schmid, 2004).

Harris Corner Detector, was proposed by (C. Harris & Stephens, 1988), is a corner detection approach, which is commonly used in computer vision algorithms to extract corners and infer features of an image. It takes into account the difference between the corner point directly rather than using the displacement block at every 45-degree angle, and is proved to be able to distinguish the angle more accurately (Dey, Nandi, Barman, Das, & Chakraborty, 2012). Furthermore, Harris-Laplace detector was proposed as the scale invariant corner detector (Mikolajczyk & Schmid, 2004), and it is consist of the Harris corner detector and the Gaussian scale space representation. In spite of the invariance of rotation and illumination changes by Harris corner detector, the points are not invariant to the scale. The Harris-Laplace approach significantly reduces the number of redundant interest points compared to Multi-scale Harris. The points are invariant to scale changes, rotation, illumination, and the addition of noise. Moreover, the interest points are highly repeatable. However, the Harris-Laplace detector returns

the much smaller number of points compared to the LoG or DoG detectors.

The feature detectors, such as DoG and Harris-Laplace, present the invariance of rotation, orientation, and consistent scaling. However, the scale can be different in each direction rather than uniform scaling if the localization and scale are useless for the affine transformation so that it leads to the fail of the scale invariant detectors in affine transformations. With the development of image processing, some features detectors have been extended to extract features invariant to affine transformations. Schaffalitzky and Zisserman (2002) modified the Harris-Laplace detector by affine normalization as the extension. And also, Mikolajczyk and Schmid (2004) proposed the approach for scale and affine invariant interest point detection.

Feature description

Scale-Invariant Feature Transform (SIFT) is an algorithm in computer vision to detect and describe local features in images, proposed by Lowe (2004). The SIFT descriptor is invariant to consistent scaling, orientation, illumination changes, and partially invariant to affine transformation. There are four main steps in SIFT algorithm. The first step is scale-space extrema detection. As known, it is impossible to use the same window to detect keypoints with different scale. Therefore, SIFT makes use of DoG, which is obtained as the difference of Gaussian blurring of an image with two different values. It is processed for various octaves of the image in Gaussian Pyramid, shown in 2.2. After obtaining DoG, the images can be found for local extrema through scale space. After getting the potential locations for keypoints, SIFT is required to acquire more precise results as refinement because scale-space extrema detection generates few unstable keypoints. The aim of this step is to remove the low contrast keypoints. Besides, the DoG is sensitive to edges so that it is necessary to be removed according to the detector of Harris corner. After that, orientation is assigned to each keypoint to keep invariance to image rotation. A neighborhood is taken around the keypoint location depending on the scale, and the gradient magnitude and direction is calculated in that region for all pixels around the keypoint using equation 2.3. The most important gradient orientations are identified using the histogram. Lastly, the

keypoint descriptor is generated, and a 16*16 neighborhood around the keypoint is taken. It is divided into 16 sub-blocks of 4*4 size. For each sub-block, 8 bin orientation histogram is created. Therefore, a total of 128 bin values are generated. SIFT descriptor representation is designed to avoid the problems of boundary changes in location, orientation and scale do not cause radical changes in the feature vector.

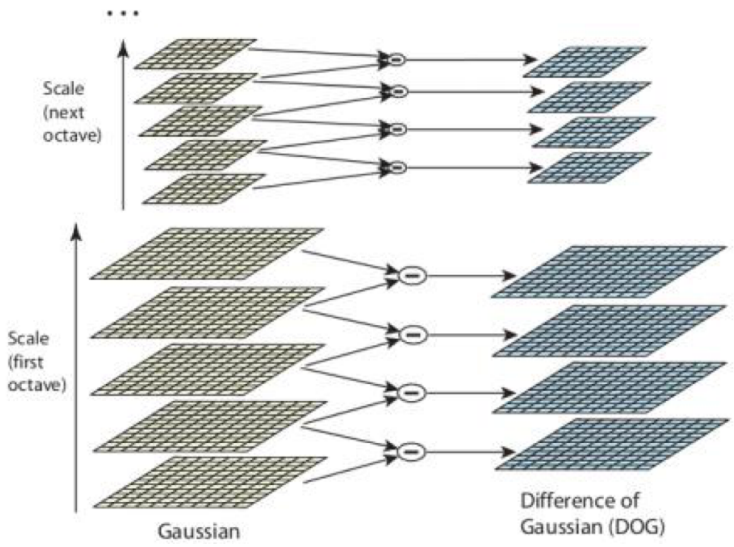


Figure 2.2: DoG representation (Sinha, 2017)

$$m(x,y) = \sqrt{(L(x+1,y) - L(x-1,y))^2 + (L(x,y+1) - L(x,y-1))^2}$$

$$\theta(x,y) = \tan^{-1}((L(x,y+1) - L(x,y-1))/(L(x+1,y) - L(x-1,y)))$$

Figure 2.3: Gradient magnitude and orientation (Sinha, 2017)

Speeded up robust feature (SURF), was proposed by Bay, Tuytelaars, and Van Gool (2006), is local feature descriptor inspired by SIFT descriptors. The SURF descriptor is based on the same principles and steps as SIFT. However, the details are different. The algorithm contains three critical steps, including interest point detection, local neighborhood description, and matching. The SURF was designed to the approximation to LoG with box filter, which is the better to calculate the convolution using box filter for integral images. Besides, the SURF depends on the determination of Hessian matrix for both scale and location. During the step of orientation assignment,

the SURF makes use of wavelet responses in horizontal and vertical direction for a neighborhood, and also, enough Gaussian weights are applied to it. The dominant orientation is estimated by calculating the sum of all responses within a sliding orientation window of angle 60 degrees. Then, a square region is extracted in order to describe the region around the points. The point of interest is divided into 4x4 square sub-regions, and the Haar wavelet responses are extracted at 5x5 regularly sample points. Compared to SIFT, the SURF can accelerate the calculation process since it employs 64-dimensional feature vector to describe the local feature as advantages rather than 128 dimensions in SIFT.

Furthermore, the Histogram of Oriented Gradient (HOG) was proposed to extract local features in images, which is the variant of SIFT (Dalal & Triggs, 2005). In this research, it indicated that the HOG provides the excellent performance relative to other existing feature sets including wavelets. Also, Ojala, Pietikainen, and Maenpaa (2002) proposed the approach of Local Binary Patterns (LBP) to extract the spatial information of the texture with the invariant to monotonic transformations of the gray levels. In a nutshell, the different approaches of feature extraction in image processing, global feature and local feature, could deliver the different performance because of the existence of various situations for images, such as scalability, illumination, and rotation. Hence, the performance of each approach should be multiple evaluated by different image datasets for image classification.

2.2 BoVW methodology

2.2.1 Introduction

Initially, the methodology of bag-of-words (BoW) is commonly used in the field of natural language processing and information retrieval, such as text categorization, and the term of BoW was early proposed by Z. S. Harris (1954) in a linguistic context. This model aims to represent texts with the number of times a term appears in the texts without the consideration of grammar and word order. After years, the methodology of BoVW was inspired by BoW model in the field of computer vision, proposed by

Csurka et al. (2004). In the process of image classification, a visual word is used in the BoVW model, generated by clustering low-level visual features of local regions points, such as color and texture along with the process of vector quantization. In other words, the BoVW is a sparse vector of occurrence counts of a vocabulary of local image features, which can be described as a histogram of visual words as well. It is possibly amazing that the BoVW schema could be effective and productive to match or surpass the other state-of-the-art performance in some developed applications because of the lack of spatial information and structure. However, the lack of spatial relationships between patches could lead to the issue of high misclassification rate in computer vision.

2.2.2 BoVW process

The process of creating BoVW model is shown in figure 2.4, which can be concluded to four key steps as follows. Firstly, it is to detect regions or points of interest. Then, computing local descriptors over those regions or points. After that, quantizing the descriptors into words to form the visual vocabulary. Lastly, finding the occurrences for each specific word in the vocabulary for constructing the BoVW model, namely the histogram of word frequencies (Tsai, 2012).

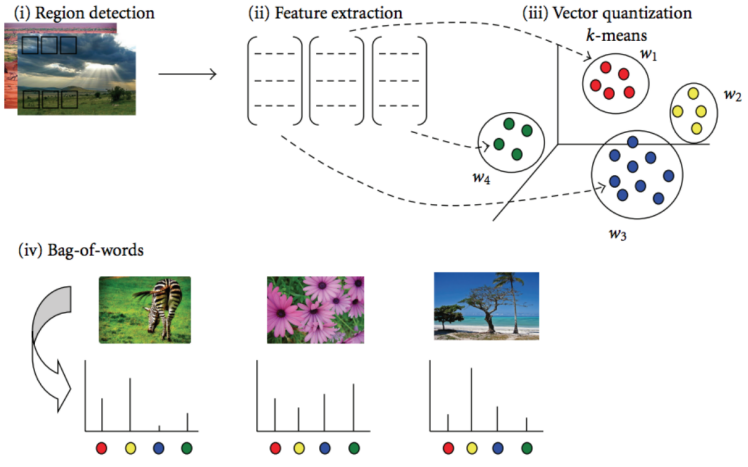


Figure 2.4: The process of generating BoVW model (Tsai, 2012)

The interest point detection detects keypoints with the scale space representations

of images. It is computed at predefined locations and scales, and also some popular detection methods were discussed by Mikolajczyk, Leibe, and Schiele (2005). In their research, they compared some well-known detectors based on affine normalization, and the conclusion is that the Hessian-Affine detector outperforms among others. Additionally, the interest points are detected by both sparse and dense approach (Horster & Lienhart, 2007). The interest points are detected at local extrema in the DoG pyramid for sparse features (Lowe, 2004). For dense features, the interest points are defined at sampled grid points.

Computing feature descriptor is an important step to decide how to represent the neighborhood of pixels near the localized region apart from making the decision where features exist in images. In BoVW literature, the SIFT descriptor (Lowe, 2004) is widely used as feature descriptors. In addition, SURF is the alternative to SIFT descriptor, and it has been widely used and applied as well (Bay et al., 2006). The process of SURF contains the procedures of feature detection and description. The purpose of SURF is to produce the similar features as produced by SIFT on Hessian-Laplace interest points, but more effective and accurate. In the study (Mikolajczyk et al., 2005), there has the comparison of some feature descriptors and concludes that the SIFT-based descriptors outperform the other descriptors in many areas. According to the study (Mikolajczyk & Schmid, 2005), the authors compared the performance of local descriptors, which are extracted by the Harris-Affine detector, and it indicated that SIFT-based descriptors deliver the best performance.

After detecting regions and extracting features for images, the final step of constructing the visual vocabulary for BoVW model is in accordance with vector quantization. Basically, the k-means clustering algorithm is used during this step, and the number of visual words generated is based on the number of clusters predefined. van de Sande, Gevers, and Snoek (2011) explained that the process of vector quantization during building BoVW model has the high computational cost using the k-means algorithm, which is to find the k number of neighbor clusters for each point. However, there have the limitations of creating the visual vocabulary in the traditional BoVW model, that is, it ignores the spatial information for images because of its orderless

collection. Therefore, Lazebnik, Schmid, and Ponce (2006) proposed the approach of spatial pyramid matching, treated as an alternative consideration of the orderless images. This method can allow the BoVW model to contain the spatial information during the process of generating visual vocabularies to improve the performance of image processing.

2.2.3 Related work

Medical science

Due to the rapid development of modern medical facilities, increasingly numerous medical images are captured and generated. For example, more than 640 million medical images have been stored over 100 National Health Service Trusts in UK in 2008 (Khaliq, Blakeley, Maheshwaran, Hashemi, & Redman, 2010). However, there have some special difficulties to classify images on the sizable medical database, such as imbalance number of training images among different classes, intra-class variability, and inter-class similarity. The research presented a BoVW-based approach to obtain high classification accuracy on ImageCLEF 2007 medical database, and the methodologies are based on BoVW for feature extraction with SIFT descriptors and the kernel of radial basis function of support vector machine classifier used in training phrase (Zare, Seng, & Mueen, 2013). Also, magnetic resonance imaging (MRI) is a powerful, non-invasive medical imaging technique widely used in neuroscience and brain disease research (Fatahi, Speck, et al., 2015), and in recent years BoVW has used to analyze MRI to complete the tasks of image classification. Daliri (2012) proposed the BoVW model with the feature extraction of SIFT descriptors from different slides in MR images and used SVM to classify them. Furthermore, Rueda, Arevalo, Cruz, Romero, and González (2012) proposed the model of BoVW model for brain MR images with the features of gray pixel intensities, based on SVM. As can be seen, the BoVW pattern will be further developed to the filed of medical science in the future.

Aerial imagery

The high spatial resolution (HSR) images can be captured and generated by devices, such as satellites and radars, in the domain of aerial imagery. The HSR aerial images can provide abundant spatial and textural information for classification (Xu, Fang, Li, & Wang, 2010). Therefore, the factors of feature detection and description are the crucial points in HSR image classification. Recently years, the BoVW model in image semantic analysis has been considered to improve image processing by many researchers. This state-of-the-art approach of image processing has been successfully applied to general visual categorization (Perronnin, 2008), texture categorization (Qin, Zheng, Jiang, Huang, & Gao, 2008) and object classification of aerial image (Xu et al., 2010). As concerned, the image classification based on BoVW model will be effectively and widely used in the filed of aerial imagery to improve military defense and civil applications.

Robotics

With the development of robotics over decades, the designed robots are purposed to assist human beings to complete tasks. Also, during the awareness process of robots, the image recognition is the necessary progress to allow robotic system to understand what images present. Recently years, the BoVW model has been developed to enhance the process of image classification in robotic system, such as robot navigation and mapping (Nicosevici & Garcia, 2012) and handicapped assistance (Ergene & Durdu, 2017). In the paperwork Nicosevici and Garcia (2012) explained that while discarding the geometric information in images, BoVW proved to be very robust methods to detect visual similarities between images, allowing efficient loop-closure detection even in the presence of illumination and camera perspective changes and partial occlusions. Besides, Ergene and Durdu (2017) proposed to make use of BoVW model to build the visual vocabulary to produce image classification on robotic hands with linear SVM. As the consideration of robotics development, the BoVW could have the potentially great effect on the image recognition in the domain of robotics in the future.

2.3 Image classification

2.3.1 Overview

Image classification is one of the fundamental problems in the domain of computer vision, which has attracted many attentions over the last decade. The goal of image classification is to predict the categories of the input images using its features. The image classification contains four main steps (Kamavisdar, Saluja, & Agrawal, 2013). First of all, the image pre-processing is important preparation before feature extraction to improve the quality of features, such as noise removal, image transformation, and principal component analysis. After that, the feature detection and extraction are conducted to generate the set of descriptors to describe images. Then, the training stage aims to train the selection of the particular features that describes the pattern at best with the machine learning algorithms. Lastly, the testing stage categorizes detected objects into predefined classes by using the suitable method that compares the image patterns with the target patterns.

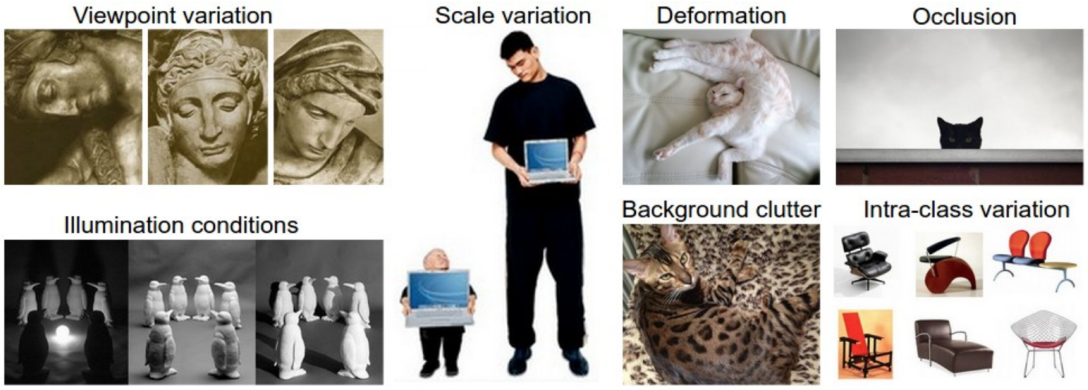


Figure 2.5: The Summary of image classification challenge (Johnson, 2017)

Furthermore, it still has many difficulties and challenges in image classification (Kurian & Karunakaran, 2012). The effect of illumination is sensitive to the pixel level that could cause the significant variations in the intensity of the pixels. A single object can be oriented in many ways concerning the camera by changing position while capturing that lead to the problem of viewpoint variation. Also, the visual objects

often exhibit variation for their sizes in the real world, and the most of objects do not have the rigid feature that can be deformed in extreme ways. Furthermore, the objects of interest could mix into their background, making them difficult to identify. And the objects can be occluded, only the small part of an object can be visible. In addition, the object classes can often be relatively broad. There could have many different types of these objects with the different appearance. The summary of challenges in image classification is shown in figure 2.5.

2.3.2 Machine learning approaches

Generally, there are two types of approaches in machine learning, which are the supervised learning for labeled data and unsupervised learning for unlabeled data. Supervised learning is to infer a function from labeled training data. It analyzes the training data and produces an inferred function, which can be mapped to the data to be assigned labels. Unsupervised machine learning is to infer a function to describe the hidden structure from unlabeled data, which means the information of categorization is not included in the observations. This approach is not widely used in the task of image classification, but it is used to generate the visual vocabulary in BoVW model, such as using k-means algorithm (Csurka et al., 2004).

Support Vector Machine (SVM) is the state-of-the-art supervised machine learning technique that is widely used in image classification. SVM builds the set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression, and outliers detection. The hyperplanes in SVM can be adjusted within the maximum margin, shown in figure 2.6. In many situations, it indicated that classification results in the issue of over-fitting in high dimensional feature spaces, however, in SVM over-fitting is controlled through the principle of structural risk minimization (Cortes & Vapnik, 1995). The problem of misclassification is minimized by maximizing the margin between the points and the boundary (Mashao, 2003).

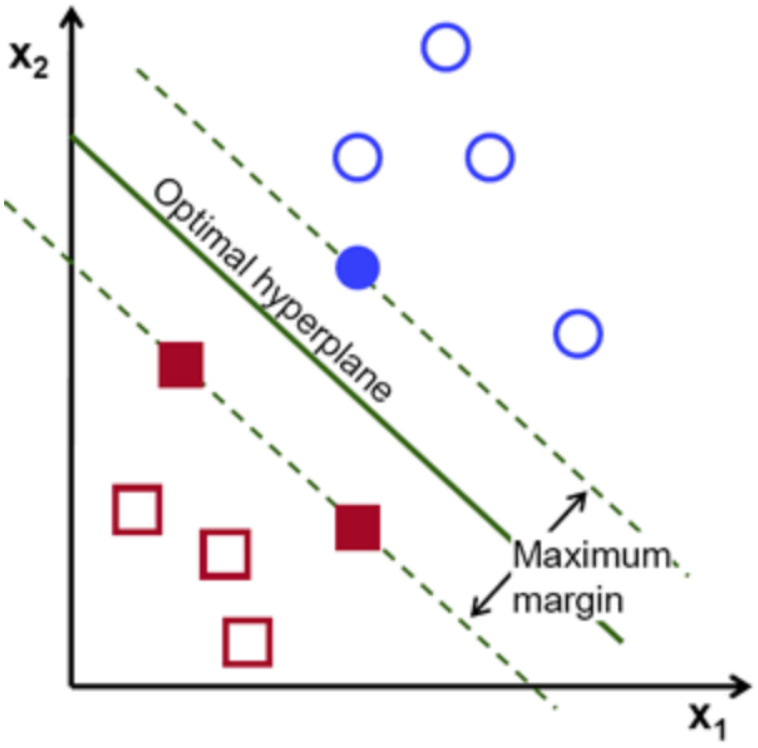


Figure 2.6: SVM representation

The SVM is naturally used for binary classification. However, it also can be extended to multi-class image classification, along with the strategies of one-against-one and one-against-all (Melgani & Bruzzone, 2004). The one-against-all inspires the most common SVM multi-class approach and involves the division of an N class dataset into N two-class cases. Also, the one-against-one approach consists in building a machine for each pair of classes resulting in N(N-1)/2 machines. Each classification gives one vote to the winning class, and the point is labeled with the class having most votes while applied to a test point. This approach can be further modified to weight in the voting process. In the research (Gualtieri & Cromp, 1999), it explained that the one-against-one approach outperforms one-against-all, because one-against-all can be compromised according to unbalanced training datasets. Additionally, the kernel method, called kernel trick as well, also plays a significant role in SVM-based classification. In fact, the state of linear is extraordinary, and the systems in the real world are not truly linear. Therefore, the non-linear model is suitable to solve the problem

of non-linearity rather than the linear model. There have some non-linear kernel functions with SVM, including polynomials and Radial Based Function. In a word, SVM gains flexibility in the choice of the form of the threshold and contains a nonlinear transformation. It also provides an excellent generalization capability, resulting in the reduction in computational complexity and simplicity of making decision rules.

Random Forest is another very successful classification algorithm, which was proposed as a combination of tree predictors so that each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest (Breiman, 2001). In another word, it is a pattern for constructing a classification ensemble with the set of decision trees, growing in the randomly selected subspace of data. It can be applied to object classification with the relatively small number of classes (Moosmann, Triggs, & Jurie, 2007). Also, the attractions of random forests have been widely developed in image classification (Bosch, Zisserman, & Munoz, 2007). Some approaches have been proposed to build random forest models from subspaces of data (Breiman, 2001; Ho, 1998). One of the most well-known forest structure, proposed by Breiman (2001), is to randomly select a subspace of features at each node to grow branches of decision trees, then to use bagging method to generate training data subsets for building individual trees, finally to combine all individual trees to form random forests model. In addition, owing to the image features of high dimensionality sparsity and multi-class labels, they could contain the uninformative feature, resulting in the problem of serious misclassification. Within the process of constructing forest, informative features could be possibly missed with the selection of small subspace from high dimensional data (Amaratunga, Cabrera, & Lee, 2008). In a nutshell, the over-fitting, mentioned before, is a serious problem, resulting in the unexpected data. However, the classifier does not tend to over-fit the model when enough trees are involved in the forest for the random forest algorithm. Then, the other advantage for random forest is that it can deal with missing values. Moreover, it is difficult to conclude that there was a significant difference performance between random forest and SVM used in image classification, and the different data distribution and various unexpected factors could significantly impact on image classification.

2.3.3 Evaluation measurement

The evaluation is the most significant stage after obtaining the classification results. Therefore, it can report how the performance of classifiers and even the performance of feature extraction approaches in image classification. With the respect of classification accuracy, it is commonly described as a metric computed from confusion matrix (Provost, Fawcett, Kohavi, et al., 1998) according to the testing sets, and also, it is estimated by different classifications, compared to indicate the significance of differences in the classification results (Foody & Mathur, 2004).

Typically, the confusion matrix contains the information about actual and predicted classifications done by the classification pattern. The following table 2.1 describes the confusion matrix for a two-class classifier where the term of true positives (TP) is the number of correct predictions that an instance is positive, the term of false positive (FP) is the number of incorrect predictions that an instance is positive, the term of false negatives (FN) is the number of incorrect of predictions that an instance negative, and term of true negatives (TN) is the number of correct predictions that an instance is negative.

	Predicted Positive	**Predicted Negative**
Actual Positive	True Positives (TP)	False Negatives (FN)
Actual Negative	False Positives (FP)	True Negatives (TN)

Table 2.1: Confusion Matrix

According to the representation of confusion matrix, the common and intuitive measure is calculated as the number of all correct predictions divided by the total number of the datasets, known as accuracy that is shown in equation 2.1.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.1}$$

Furthermore, the other measurement still plays a crucial role in the evaluation of classification that are sensitivity and specificity. Sensitivity, also called recall or true

positive rate, is calculated as the number of correct positive predictions divided by the total number of positives, shown in equation 2.2.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2.2}$$

Specificity, also called true negative rate, is calculated as the number of correct negative predictions divided by the total number of negatives, shown in equation 2.3.

$$Specificity = \frac{TN}{TN + FP} \tag{2.3}$$

Additionally, precision, called positive predictive value, is calculated as the number of correct positive predictions divided by the total number of positive predictions, shown in equation 2.4.

$$Precision = \frac{TP}{TP + FP} \tag{2.4}$$

2.4 Statistical test

Over the last decade, the field of machine learning has been increasingly aware of the need for statistical validation of comparisons (Demšar, 2006). Dietterich (1998) examines McNemars test on misclassification matrix as powerful as the 52 cv t-test in the case of the unreliability of running the algorithm morn than once, and using t-test is discouraged after cross validation. Nadeau and Bengio (2000) proposed the corrected re-sampled t-test that adjusts the variance over subsets of examples. However, none of the studies above found the approach to cope with evaluating the performance of multiple classifiers and the performance of classifiers, tested by multiple datasets. Therefore, the non-parametric testing is proposed to compare classifiers in information retrieval (Schütze, Hull, & Pedersen, 1995). And also, Vázquez, Escolano, Riaño, and Junquera (2001) studied ANOVA (Fisher, 1956) and Friedman's test (Friedman, 1940) for comparison of multiple models on single data.

As discussed above, the Friedman test is a non-parametric equivalent of the repeated-measures ANOVA, along with the ranks of the algorithms for each dataset. The

statistic of Friedman is distributed based on χ_F^2 with k-1 degrees of freedom, shown in equation 2.5. (Demšar, 2006). It can identify whether the significant difference appears over multiple datasets.

$$\chi_F^2 = \frac{12N}{k(k+1)}\left[\sum_j R_j^2 - \frac{(k(k+1)^2)}{4}\right] \tag{2.5}$$

The Nemenyi test (Nemenyi, 1962), treated as post-hoc test, is used to compare all classifiers to each one when null-hypothesis is rejected in Friedman test. It indicates which paired classifiers have the significant difference over multiple datasets if the corresponding average ranks differ by at least the critical difference, shown in equation 2.6.

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{2.6}$$

Additionally, the table 2.2 describes the critical values for the two-tailed Nemenyi test, as using after Friedman test.

#approaches	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$\mathbf{q_{0.05}}$	1.960	2.343	2.569	2.728	2.850	2.949	3.031	3.102	3.164
$\mathbf{q_{0.10}}$	1.645	2.052	2.291	2.459	2.589	2.693	2.780	2.855	2.920

Table 2.2: Critical values for the two-tailed Nemenyi test (Demšar, 2006)

2.5 Conclusion

This chapter reviewed regarding peer-reviewed papers in details, along with the field of image classification based on BoVW model in computer vision. First of all, in the section 2.1, the prior procedure of image classification, known as image processing, were studied, including both representations of global feature and local feature. Furthermore, the stages of feature detection and feature description while representing local feature were extensively discussed, and also, the state-of-the-art feature descriptors were introduced and compared, such as SIFT, SURF, and HOG. Next to the section 2.2, it reviewed the model of BoVW used in the task of image classification, along

with the relevant research work in different domains. Then, the process of generating BoVW was introduced and discussed in details. Also, the different selections of the approaches during the process of generating BoVW model were extensively discussed and compared, such as the selections of feature descriptors and the selections of vector quantization algorithms. In addition, the limitations of BoVW model were presented and explained as well.

After that, in the section 2.3 the overview of image classification were introduced at the high level, including the general process and its challenges. The supervised machine learning algorithms were considered to introduce in general, and the SVM and random forest were intensively discussed and compared in image classification. Additionally, the measurement of classification evaluation was presented and described with a little math knowledge. Furthermore, the section 2.4 was intensively discussed, and also the Friedman test and the relevant post-hoc test was deeply introduced and studied with a little math knowledge.

The next chapter, namely experiment design and methodology, will be presented to design the methodologies and experimental software for image classification based on the proposed research question in chapter 1 and the existing literature reviews in this chapter.

Chapter 3

Experiment design and methodology

3.1 Introduction

In this chapter, the experiments are designed and presented in details for this research, including the descriptions of image datasets, the methodologies of all experiments, the design of experimental software, and the conclusion of this chapter. The structure of this chapter is described as following. Firstly, in the section of data used, the image data sources will be described in general and illustrate its advantages. Secondly, in the section of evaluation methodology, the methodology will be presented in the aspects of approaches, performance measures, statistical tests. Thirdly, in the section of experimental software design, the detailed software structure will be designed and presented. Finally, the conclusion of this chapter will be presented, and describe what it will go through in the next chapter.

3.2 Data used

As the part of this research, the data source selection plays a significant role in the process of image classification, since it can impact on classification performance. The image data source, called Caltech-256, is used to the task of image classification for this

research. The original version, called Caltech-101 (Fei-Fei, Fergus, & Perona, 2007), was collected by selecting a set of object categories that downloaded instances from Google Images, and then manually screening out all images that did not fit the category. However, Caltech-256 was collected in similar methods with some improvements, such as more than a double number of categories, the increase of minimum number of images for any category from 31 to 80, and the avoiding of artifacts because of image rotation (Griffin et al., 2007). The summary between Caltech-101 and Caltech-256 is illustrated in table 3.1. As can be seen, there has a dramatic improvement from Caltech-101 to Caltech-256, including the number of categories, total images and minimal instances for each category. Therefore, the Caltech-256 is conducted to involved in image classification experiments to answer research question defined in the previous chapter.

Dataset	Released	Categories	Images total	Min	Med	Mean	Max
Caltech-101	2003	101	9144	31	59	90	800
Caltech-256	2006	256	30607	80	100	119	827

Table 3.1: Comparison between Caltech101 and Caltech256. The clutter categories are excluded.(Griffin et al., 2007)

The datasets will be determined to select from Caltech-256 and use to both experiments of binary classification and multi-class classification, which will be discussed and presented in details in the next chapter. Classification across all 256 images is a complex process so separate binary and multi-class datasets were extracted from the source data. It is time-consuming and dependent on the high-quality personal computer or laptop. Therefore, the dataset used to binary classification experiment contains 21 paired-category sub-datasets, assembled from 7 categories that are mountain bike, mushroom, mussels, necktie, octopus, ostrich and owl from Caltech-256, shown in table 3.2 and figure 3.1. Besides, BD(i) denotes that the i^{th} sub-dataset for binary classification experiment. In a word, each sub-dataset only contains two categories with the same number of instances for each category. For example, The categories of mountain bike and mushroom both have 82 instances in BD1 sub-dataset.

	Classes	#Instances
BD1	mountain bike, mushroom	82, 82
BD2	mountain bike, mussels	82, 82
BD3	mountain bike, necktie	82, 82
BD4	mountain bike, octopus	82, 82
BD5	mountain bike, ostrich	82, 82
BD6	mountain bike, owl	82, 82
BD7	mushroom, mussels	174, 174
BD8	mushroom, necktie	103, 103
BD9	mushroom, octopus	111, 111
BD10	mushroom, ostrich	109, 109
BD11	mushroom, owl	70, 70
BD12	mussels, necktie	103, 103
BD13	mussels, octopus	111, 111
BD14	mussels, ostrich	109, 109
BD15	mussels, owl	70, 70
BD16	necktie, octopus	111, 111
BD17	necktie, ostrich	103, 103
BD18	necktie, owl	70, 70
BD19	octopus, ostrich	111, 111
BD20	octopus, owl	70, 70
BD21	ostrich, owl	70, 70

Table 3.2: The summary of datasets for binary classification

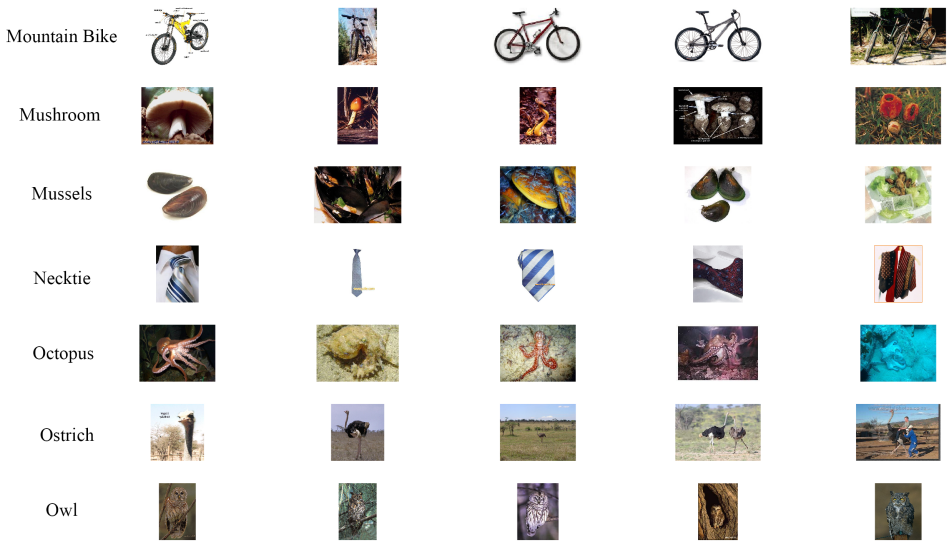


Figure 3.1: Examples for binary classification

The generated dataset for multi-class classification experiment is shown in table 3.4 and figure 3.2 for Group A, and table 3.5 and figure 3.3 for Group B. The MDA(i) and MDB(i) denotes that the i^{th} dataset in Group A and the i^{th} dataset in Group B for multi-class classification experiment, respectively. Each category is related to the class of animal in group A, and each category is randomly selected from the rest of categories, which do not belong to the class of animal in group B. Therefore, there has a difference that group A contains similar categories and images, and group B contains non-relevant categories and images. Each group contains eight sub-datasets that the number of classes is increasing from 3 classes to 10 classes. And also, the latter sub-dataset is generated by adding a new category, based on the former. For example, MDB1 has the categories that are baseball bat, bathtub and bulldozer, and MDB2 has the categories that are baseball bat, bathtub, bulldozer, and calculator. In addition, each sub-dataset contains the same number of instances for each category, based on the minimal number of instances among categories. It can keep balance for the proportion of each category in one sub-dataset to deliver the correct classification results.

In addition, the dataset, designed for the experiment of selecting vocabulary size , is displayed in table 3.3. It contains ten classes, which are randomly selected and obtained from both used datasets of binary classification experiment and multi-class classification experiment. As can be seen, the types of categories, such as animal-relevant and non-relevant, would not impact on the results in the process of vocabulary size selection. In another word, the results of this experiment are yielded, excluding the influence of samples selection.

	Classes	**#Classes**	**#Instances**
Dataset	bathtub, cormorant, deskglobe, horse mountain bike, mushroom, necktie penguin, T-shirt, windmill	10	82

Table 3.3: Dataset for vocabulary size selection

	Classes	#Classes	#Instances
MDA1	bear, cormorant, dolphin	3	102
MDA2	bear, cormorant dolphin, goat	4	102
MDA3	bear, cormorant, dolphin goat, goose	5	102
MDA4	bear, cormorant, dolphin goat, goose, horse	6	102
MDA5	bear, cormorant, dolphin goat, goose horse,octopus	7	102
MDA6	bear, cormorant, dolphin goat, goose, horse octopus, ostrich	8	102
MDA7	bear, cormorant, dolphin goat, goose, horse octopus, ostrich, penguin	9	102
MDA8	bear, cormorant, dolphin, goat goose, horse, octopus ostrich, penguin, swan	10	102

Table 3.4: Multi-class Classification: Group A (Animal-relevant Class)

Figure 3.2: Examples of Group A (Animal Class) for multi-class classification

	Classes	#Classes	#Instances
MDB1	baseballbat, bathtub, bulldozer	3	82
MDB2	baseballbat, bathtub bulldozer, calculator	4	82
MDB3	baseballbat,bathtub,bulldozer calculator,deskglobe	5	82
MDB4	baseballbat,bathtub,bulldozer calculator,deskglobe,golfball	6	82
MDB5	baseballbat,bathtub,bulldozer calculator, deskglobe golfball, laptop	7	82
MDB6	baseballbat,bathtub,bulldozer calculator,deskglobe,golfball laptop,segway	8	82
MDB7	baseballbat,bathtub,bulldozer calculator,deskglobe,golfball laptop,segway,T-shirt	9	82
MDB8	baseballbat,bathtub,bulldozer calculator,deskglobe,golfball,laptop segway,T-shirt,windmill	10	82

Table 3.5: Multi-class Classification: Group B (Non-relevance Class)

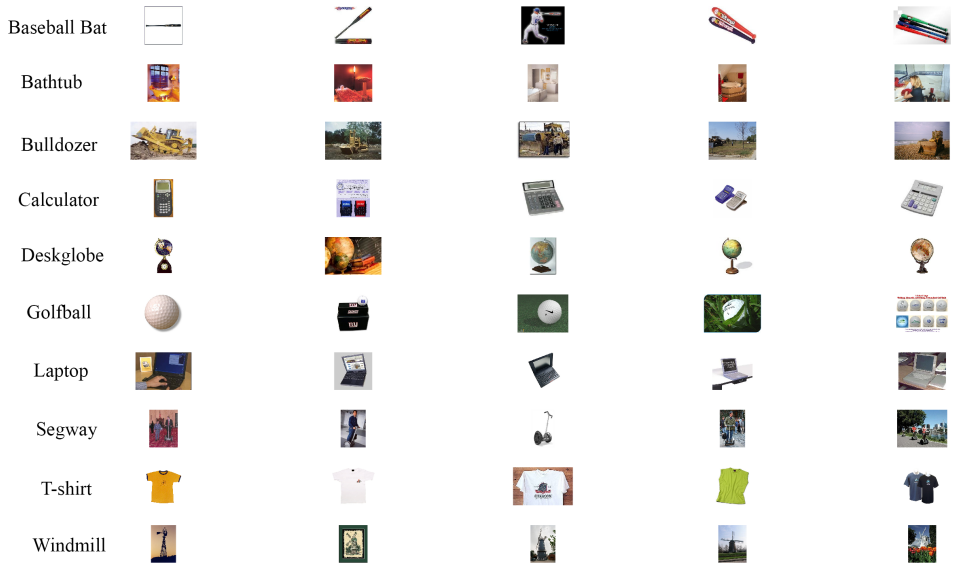


Figure 3.3: Examples of Group B (Non-relevance Class) for multi-class classification

3.3 Evaluation methodology

In this section, the general evaluation methodologies are stated and described in the aspects of approaches, performance measures and statistical test, and these methodologies will be applied to all experiments.

3.3.1 Approach

As the discussion made about the purpose of this research and the description of datasets, the selected datasets will be divided into 2 partitions, treated as the training sets and the testing sets to implement the process of image classification. In the meantime, the proportion of partitions are set as 70% for training sets and 30 % for testing sets.

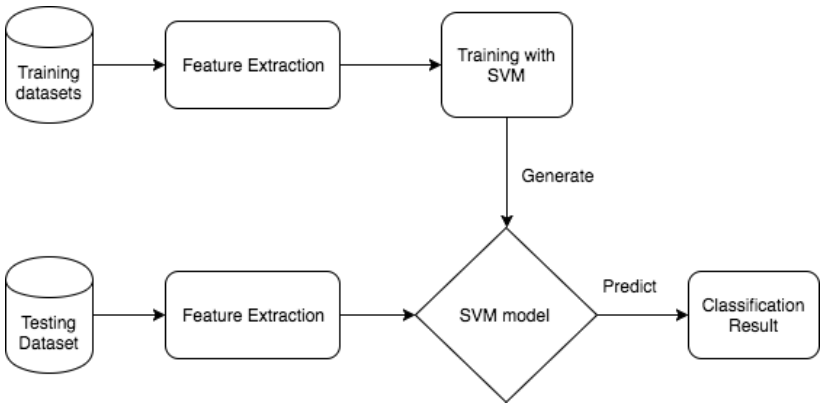


Figure 3.4: The flow chart of the supervised image classification

The flow chart 3.4 is designed to illustrate the fundamental stages during the process of image classification. In general, it contains training stage and testing stage. In the training stage, the images are conducted to the process of feature extraction, such as global feature and BoVW model along with the local feature. Then, the extracted features are trained with SVM algorithm to generate the SVM classifier model. According to the paperwork (Csurka et al., 2004), it empirically proved that the SVM classifier delivers the best performance for the task of image classification over the other supervised machine learning classifiers. In the testing stage, the images to be tested are processed using the same approach of feature extraction, and the

classification results are generated by the corresponding classifier model. Moreover, the linear SVM with one-against-one approach will be used in the experiment of multi-class classification.

Additionally, the number of iterations for all experiments runs is set as three times so that it can provide the robustly experimental results. Besides, the results generated by this way can be conducted to statistical test to prove the correctness of hypotheses given by Chapter 1. Also, some approaches of image processing, introduced and discussed in chapter 2, will be used to present images before entering the training stage, such as raw pixels and color histogram representation, and BoVW pattern with the SIFT and SURF descriptors, respectively. Furthermore, the different designed experiments, given by Chapter 4 in details, will make use of the approaches as mentioned earlier approaches to present images based on the purpose of each experiment.

3.3.2 Performance measures

The most important part of this research is to evaluate the given results by the conducted experiments so that it can achieve and conclude the aim of this research. In the field of image classification, the confusion matrix is the popular and proper method to evaluate the performance of classifiers, even the performance of image processing methods. In another word, the confusion matrix is usually used as the quantitative method of characterizing image classification accuracy. For all designed experiments in this research, the average accuracy will be calculated based on the output of confusion matrix, and it also will be used in the stage of statistical test in order to find whether the significant difference exists during groups.

3.3.3 Statistical test

The statistical test provides a mechanism for making quantitative decisions about processes, and also statistical methodologies are required to make sure that the data is interpreted correctly and that apparent relationship is significant and not merely chance occurrences. Therefore, it is the essential procedure after obtaining the results,

given by experiments in this research. In addition, the average accuracy, obtained by confusion matrix, is not sufficient to prove the defined hypotheses, since it could lead to the problem of different conclusions due to the samples selection during experimentations.

As the designed datasets for experiments, there have some comparisons between the different approaches of image processing over multiple datasets. Therefore, the non-parametric test, called Friedman test, will be applied to find out the differences between each image processing approach without the assumption of normal distribution test (Demšar, 2006). However, the Friedman test only can reveal whether there have the differences over multiple datasets, but it cannot indicate whether the significant differences exist for each approach of image process again the other one. To solve this problem, the post-hoc test is proposed to process if the null-hypothesis is rejected. The Nemenyi test is used when all image processing approaches are compared to each other (Demšar, 2006) in order to explain the significant difference between the performance of two approaches based on the corresponding average ranks differ by at least the critical difference. The formula for calculating critical value is displayed in equation 2.6, and the table of critical value is shown in table 2.2.

3.4 Experimental software design

3.4.1 Development environment

All experiments will be carried out using a MacBook pro with the macOS High Sierra that is version 10.13.2, and the hardware is displayed in table 3.6. Besides, the designed software will be implemented in MATLAB with the version of R2017b 64bit under academic license.

Furthermore, the computer vision toolbox, supported by MATLAB, provides a comprehensive suite of algorithms and tools for object detection and recognition. The system toolbox is a suite of several machine learning, feature-based, and motion-based techniques for object detection and recognition. Also, the VLFeat open source library with the version of 0.9.20, treated as third-party package in this research, implements

Type of Hardware	Description
Processor Name	Intel Core i7
Processor Speed	2.2 GHZ
Total Number of Cores	4
L2 Cache (per Core)	256 KB
L3 Cache	6 MB
Graphics	Intel Iris Pro 1536 MB
Memory	16 GB 1600 MHz DDR3

Table 3.6: Hardware for Development Environment

popular computer vision algorithms specializing in image understanding and local features extraction and matching. It is written in C for efficiency and compatibility, with interfaces in MATLAB for ease of use, and detailed documentation throughout. It supports Windows, Mac OS X, and Linux.

3.4.2 Software design

According to the designed process of image classification above, the relevant coding using MATLAB is designed as follows. The figure 3.5, similar to class diagram, illustrates the structure of software that will conduct all designed experiments in MATLAB. In general, the top level script is main.m for this study, which is proposed to call all functions defined by myself. The functions of extracting features are defined, such as raw pixels extraction, color histogram extraction, SIFT extraction, and SURF extraction. Then, the outputs returned by the function of SIFT extraction and the function of SURF extraction is processed by the function of building the visual vocabularies. Besides, the function of SVM classifier is also designed to provide the high-performance classification. This figure indicates what the variables of input and output are presented.

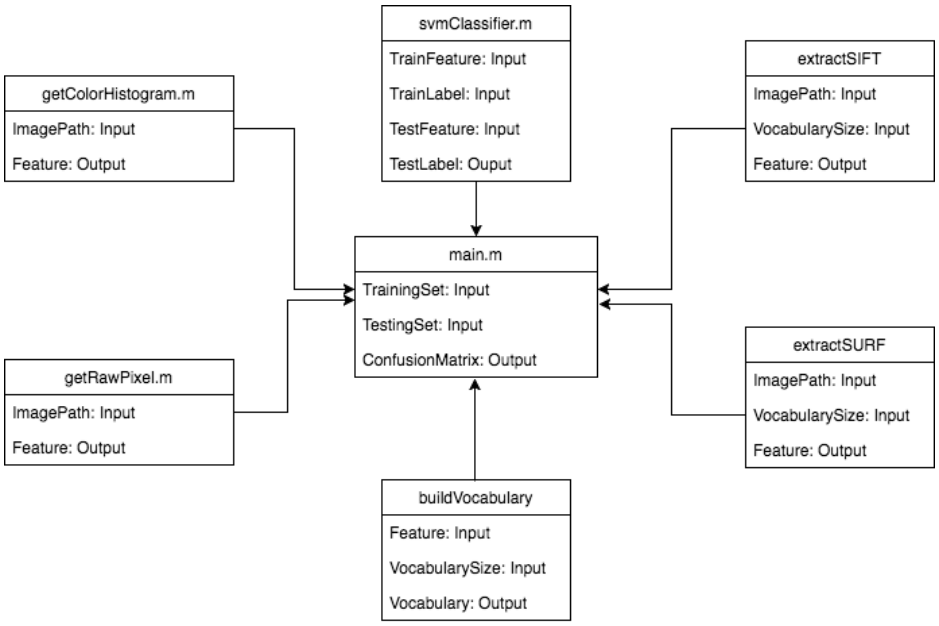


Figure 3.5: The similar class diagram for software design based on MATLAB

3.5 Conclusion

The objective of this chapter is to provide the design of the experiments and methodologies in order to accomplish the objective of this research, discussed in chapter 1. This chapter begins with the explanations of selecting image data source reasons, and the general methodologies of creating the relevant datasets for all designed experiments as well. Then, the evaluation methodologies are designed and discussed, including used approaches, performance measures, and statistical test. Lastly, the experimental software is designed and presented that will be applied to all proposed experiments. The next chapter will present the descriptions, results, and evaluations of each experiment in details, and also, the key findings and analysis will be discussed.

Chapter 4

Experimentation and results

4.1 Introduction

This chapter presents the details and results of all experiments based on the designed methodologies in Chapter 3. In total, there were three experiments conducted. Firstly, the binary classification experiment is to determine whether BoVW is the best approach of image processing over baseline approaches. Secondly, the aim of multi-class classification experiment is to find out how the performance of BoVW model extends to multiple classes. Lastly, the experiment of selecting vocabulary size presents the importance of vocabulary size selection during the process of creating BoVW on classification results. Furthermore, the evaluation and discussion of results and findings will be presented for all experiments.

4.2 Binary classification experiment

4.2.1 Implementation

Considering the overview of approaches of image processing in the field of image classification in chapter 2, the aim of binary classification experiment is to determine which image processing approach can deliver the best performance in the process of image classification using a linear SVM classifier among raw pixels, color histogram,

and BoVW model with SIFT and SURF descriptors, respectively. For performance comparison, on the one hand, the approaches of raw pixels and color histogram are treated as baseline approach, and on the other hand, the BoVW model is treated as state-of-the-art approach with SIFT and SURF descriptors as well, respectively.

The designed dataset to be used in this experiment is illustrated in Table 3.2 in chapter 3. In general, this dataset contains 21 sub-datasets, having only two classes with the same number of instances for each sub-dataset. It is convenient to evaluate the results, generated by this experiment, without the consideration of sample distributions. In addition, due to the selection of different parameters, the paper (Lowe, 2004) provides empirical evidence, such as the number of octaves is 4 and number of scale levels is 5. Therefore, the SIFT parameter in this experiment will be set in the same way. And, the images are shrunk to small square resolution with 16 * 16 blocks for the parameter of raw pixel features. Furthermore, as consideration about the influence of vocabulary size during the process of vector quantization for creating BoVW model (Hou, Kang, & Qi, 2010), the vocabulary size is statically defined to 500 that allows to reduction of the computation cost and time with a still high performance of classification. In the linear SVM algorithm is involved in training stage to generate the predictive models for all selected approaches of image processing.

4.2.2 Results and statistical test

As discussed in the literature review, in the binary classification, the accuracy, calculated by the confusion matrix from SVM classifier, is the most common measurement to evaluate the performance of classification on the various approaches of image representation. Owing to the three iterations during running, the average accuracy is calculated for each sub-dataset in binary classification experiment, where the sum of accuracies and divided by 3. The results of average accuracy for binary classification experiment is shown in table 4.1. Also, for easy viewing results, the multiple histograms are provided in figure 4.1, showing the comparison of average accuracy using the different features for each sub-dataset. Besides, the descriptive statistic is provided as well in table 4.2.

As can be seen, the mean accuracy for SURF, regarded as the common measure, is the highest among all feature extraction approaches, which is 0.916, in contrast, the mean of raw pixel is the lowest, which is 0.657. According to the values of range, standard deviation and its error, it indicates that the SURF has the most accurate and stable performance with the less error on feature extraction in image classification. However, it has not been enough to make conclusion so far, because there has a slight difference in the values of mean accuracy between the approaches of SIFT and SURF. It cannot prove which can yield the best performance in the real world without the inferential test. Furthermore, the difference will be statistically tested among all approaches afterward.

	RawPixel	**Color**	**SIFT**	**SURF**
BD1	0.78	0.81	0.91	0.92
BD2	0.64	0.74	0.89	0.91
BD3	0.66	0.72	0.91	0.90
BD4	0.72	0.80	0.89	0.88
BD5	0.75	0.81	0.90	0.94
BD6	0.68	0.8	0.92	0.88
BD7	0.66	0.68	0.80	0.85
BD8	0.71	0.82	0.9	0.93
BD9	0.51	0.73	0.89	0.91
BD10	0.73	0.74	0.93	0.94
BD11	0.50	0.68	0.90	0.89
BD12	0.58	0.71	0.87	0.90
BD13	0.64	0.74	0.85	0.91
BD14	0.61	0.73	0.79	0.89
BD15	0.63	0.68	0.95	0.89
BD16	0.70	0.81	0.94	0.93
BD17	0.77	0.82	0.91	0.98
BD18	0.69	0.89	0.95	0.96
BD19	0.70	0.80	0.95	0.94
BD20	0.45	0.76	0.94	0.93
BD21	0.68	0.74	0.96	0.95

Table 4.1: The results of average accuracy for binary classification

	N	Range	Min	Max	Mean	Mean Std. Error	Std. Deviation	Variance
Raw Pixel	21	0.33	0.45	0.78	0.657	0.019	0.087	0.008
Color histogram	21	0.21	0.68	0.89	0.762	0.012	0.056	0.003
SIFT	21	0.17	0.79	0.96	0.902	0.010	0.046	0.002
SURF	21	0.13	0.85	0.98	0.916	0.007	0.031	0.001

Table 4.2: The descriptive statistic on accuracy for binary classification

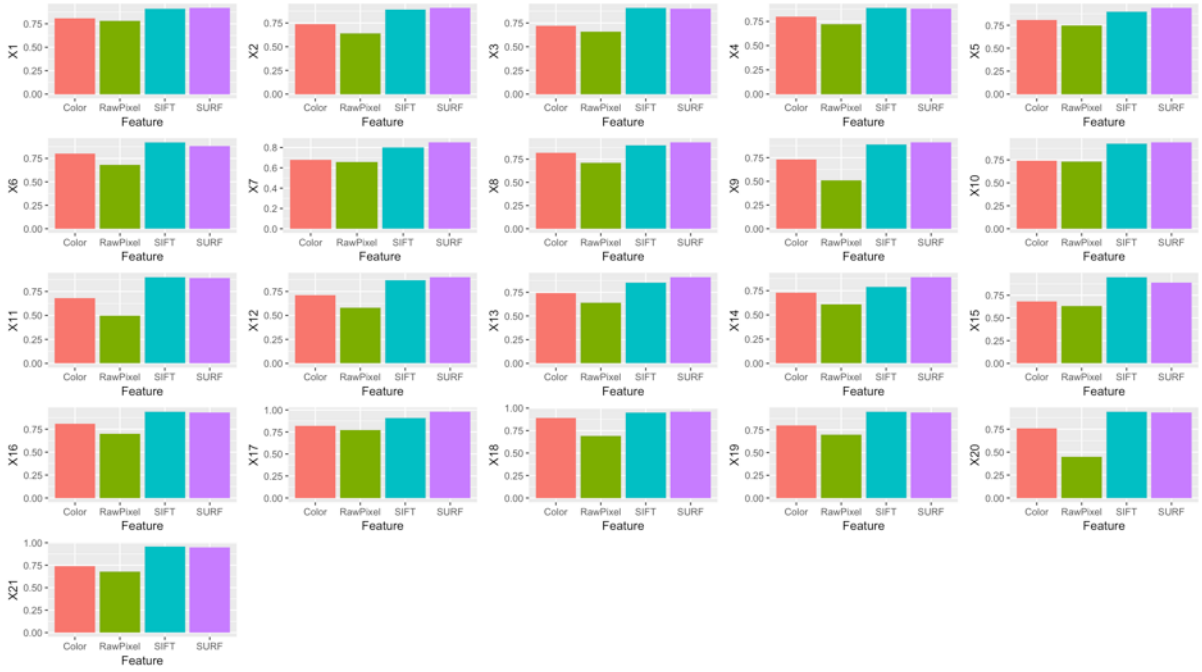


Figure 4.1: Histograms of average accuracy for each sub-dataset in binary classification

As discussed, it is inevitable that the statistical test is conducted to prove the significant difference based on accuracy output. Regarding the literature review, the Friedman test, known as one type of non-parametric test, aims to prove if there has the significant difference over the whole dataset. After that, the Nemenyi test, known as one type of post-hoc test, aims to prove which the significant differences exist between every two approaches. The threshold of p is set as 0.05. For Friedman test, the null hypothesis is defined as there has no significant difference through all approaches, which are SURF, SIFT, color histogram, and raw pixel. Then, χ_F^2 is calculated by the equation 2.5 with the relevant ranks. The results of χ_F^2 is 56.83 and the values of the ranks are 1, 2, 3.43 and 3.57 for raw pixel, color histogram, SIFT, and SURF,

respectively. And the p value is 0.00 that is less than 0.05, then the null hypothesis is rejected. The all original results generated by SPSS are provided in Appendix A. The result indicates that there was a significant difference for the variables of raw pixel, color histogram, SIFT and SURF. After that, the Nemenyi test, called post-hoc test, is conducted to compare all variables to each one, because the null hypothesis is rejected by Friedman test. Regarding to the equation 2.6 and table 2.2, the value of critical difference is calculated as 1.024. Then, the rank of each variable is compared to the other variables so that they have a significant difference if the value of comparison differs by at least the critical difference. The table 4.3 shows that the comparisons of raw pixel and color histogram, and SIFT and SURF are the less than the value of critical difference, that is 1.024, but the other comparisons are greater than critical difference.

Moreover, on the basis of the results in binary classification experiment, the SIFT and SURF are treated as the cutting-edge approach to build the BoVW model for image classification rather than the approach of raw pixel and color histogram for image classification. As known, using raw pixel features as inputs to SVM classifier could yield poor results as even small changes in rotation, translation, viewpoint, and scale, which could significantly impact on the images themselves. The SIFT descriptors are local, based on the appearance of the object on a particular point of interest, and the scale and rotation of the image are invariable. They are also robust to illumination, noise and small changes in the viewpoints. In addition to these attributes, they are highly characteristic and allow accurate object recognition, and the probability of mismatching is very low.

In conclusion, the approaches of feature extraction, SIFT, and SURF, outperform the approaches of raw pixel and color histogram on image classification. Furthermore, regarding the Friedman test, there have no statistically significant differences in the performance of classification between the feature extraction approaches of raw pixels and color features, SIFT and SURF, respectively. Also, there have the statistically significant differences over multiple comparisons, which are raw pixel and SIFT, raw pixel and SURF, color features and SIFT, color feature and SURF, respectively, according

to the results obtained by Nemenyi test.

Comparison Group	Results
Raw pixel, Color Histogram	1
Raw pixel, SIFT	2.43
Raw pixel, SURF	2.57
Color histogram, SIFT	1.45
Color histogram, SURF	1.57
SIFT, SURF	0.14

Table 4.3: The results of the rank differences

4.3 Multi-class Classification Experiment

On the basis of the results from binary classification experiment, the multi-class classification is proposed to concentrate on the performance of BoVW with SURF descriptor for the multiple classes as the extension of binary class. The reason why selecting SURF as feature extraction is that it provides the similar performance and the less computation cost against the cutting-edge approaches, that is SIFT. Also, this experiment aims to study the classification performance of BoVW model on the relevant classes, like animals, and the irrelevant classes, using linear SVM classifier with the same parameters as binary classification experiment.

The data used in this experiment has been designed in the previous chapter, shown in table 3.4 for animal-relevance group (Group A), and table 3.5 for non-relevance group (Group B). And also, the examples of each class for Group A and Group B are presented in figure 3.2 and figure 3.3, respectively. As can be seen, in Group A, some instances seem like that they have the similar shapes between the classes of cormorant, goose, ostrich, and swan, which could be a challenge to classify them to the correct classes. In contrast, the instances seem like to be easily categorized in Group B by a human. Furthermore, the parameters are defined in the same way as the previous experiment, such as 500 sizes of vocabulary and arguments in SURF. However, the

multi-class classification is the more complicated process than binary classification. As the discussion made in chapter 3, the error-correcting output codes (ECOC) model is used to complete the multi-class classification with SVM in MATLAB. By default, it uses K(K-1)/2 binary SVM model, along with the one-against-one approach, where K is the number of unique class labels. Based on the objective of this experiment,another approach for multi-class classification, called one-against-all, will not be considered and tested.

4.3.1 Results

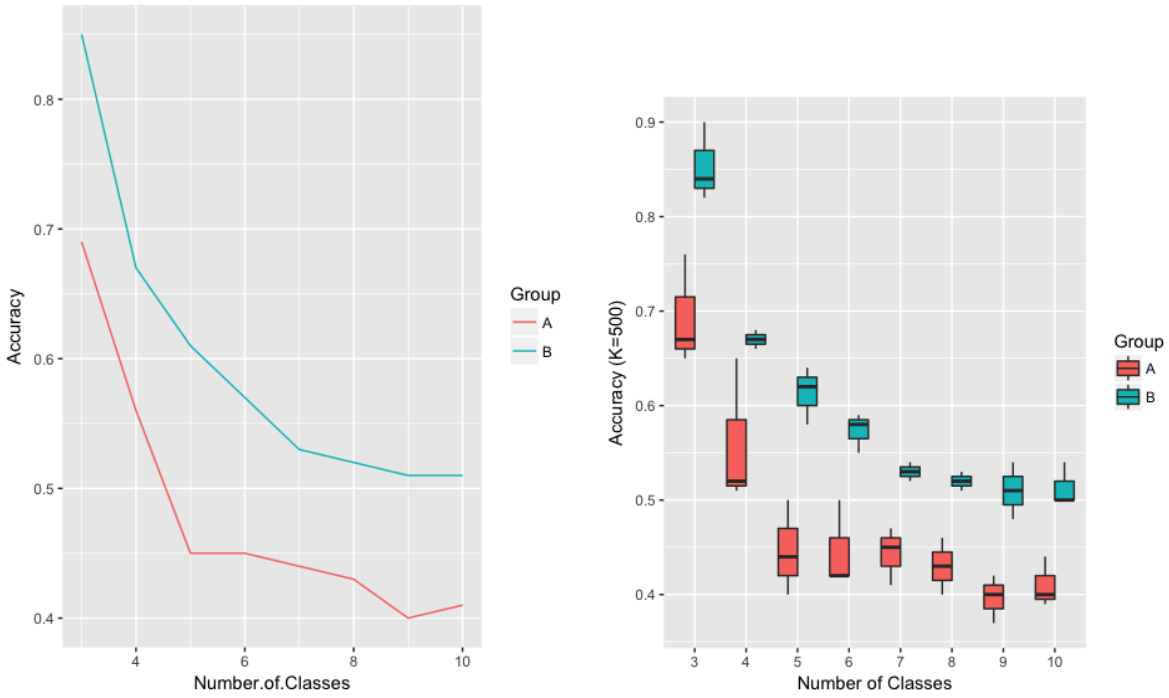
On the basis of implementation in multi-class classification experiment, the results are the form of confusion matrix as output by coding with 3-times iteration runs. In total, 48 confusion matrices are generated. For the convenience of explanation, one of the confusion matrix is shown in table 4.4. According to this confusion matrix, the average accuracy, namely the average per-class effectiveness of a classifier, is calculated as 0.39. Repeating this calculation process, the overall results of average accuracy with the class range from 3 to 10 in both Group A and Group B, shown in table 4.5. Also, the relevant visualization is provided in figure 4.2 with a line chart and a box plot.

Predict / Actual	Bear	Cormorant	Dolphin	Goat	Goose	Horse	Octopus	Ostrich	Penguin	Swan
Bear	70	20	10	30	30	20	50	50	10	20
Cormorant	20	130	30	10	10	20	10	10	30	40
Dolphin	10	20	170	0	0	0	10	40	10	50
Goat	30	10	20	70	0	80	30	40	10	20
Goose	10	50	10	10	50	30	0	40	20	90
Horse	40	10	10	10	10	120	30	70	0	10
Octopus	10	10	20	0	10	30	200	20	0	10
Ostrich	0	0	30	0	0	10	40	190	30	10
Penguin	30	40	20	30	0	20	40	30	50	50
Swan	0	40	20	10	10	20	10	20	20	160

Table 4.4: An example of confusion matrix for 10-class classification

Group A	Average Accuracy	Group B	Average Accuracy
MDA1	0.69	**MDB1**	0.85
MDA2	0.56	**MDB2**	0.67
MDA3	0.45	**MDB3**	0.61
MDA4	0.45	**MDB4**	0.57
MDA5	0.44	**MDB5**	0.53
MDA6	0.43	**MDB6**	0.52
MDA7	0.40	**MDB7**	0.51
MDA8	0.41	**MDB8**	0.51

Table 4.5: The results of average accuracy for Group A and Group B in multi-class classification experiment



(a) Multi-class Classification Mean Accuracy

(b) Multi-class Classification Accuracy Boxplot (Median)

Figure 4.2: The visualization for the results of multi-class classification experiment.

As can be seen, there has a dramatic decrease on average accuracy, that drop from

0.69 to 0.41 in Group A and drop from 0.85 to 0.51 in Group B, with the increasing number of classes from 3 classes to 10 classes. As a matter of fact, the average accuracy of binary classification is 0.916 for SURF descriptors. It indicates that there has a significant influence on the performance of classification on the factor of the number of classes. However, the average accuracy slightly changes when the number of classes is more than 7.

Furthermore, the red line of average accuracy for Group A is entirely under the green line for Group B with the comparison of the same number of classes. It means that the instances that belong to the type of animal, are challenging to be trained and classified by SVM using SURF descriptors due to the similarity of samples. According to an example of confusion matrix in table 4.4, only 50 images of actual gooses are correctly classified to the label of goose, but 90 images of actual gooses are classified to the label of swan in a wrong way. Back to check the entities of images for goose, and swan in figure 3.2, it indicates that the entities of goose and swan have some similar attributes, such as the similar-shape of neck and head, and even color. In addition, the same issue occurs between some categories, such as cormorant, penguin and goose. Also, in the box plot, the black line stands for the median of average accuracy for each sub-group. Compared to Group B, the performance of classification in Group A is not stably change, but a drastic fluctuation. In a word, the smaller boxes in box plot deliver the more stable performance on classification, and the small boxes are desired to generate by experiment.

4.4 Selecting vocabulary size experiment

4.4.1 Implementation

As the consideration mentioned in chapter 2, the size of a visual vocabulary is generated by the number of keypoint clusters in the process of clustering using the k-means algorithm. Therefore, the factor of selecting a suitable vocabulary size is a necessary determination that can greatly impact on the accuracy of classification. Along with a small vocabulary, the visual word is not very discriminative, because the different

keypoints can be treated to the same visual word. With the increasing vocabulary sizes, the feature becomes more discriminative, but in the meantime, the less generalizable and forgiving to noises since similar keypoints can be treated to different visual words. Choosing a large size of vocabulary increases the computational cost of clustering keypoints, computing visual vocabulary, and running SVM classifiers as well. Therefore, the experiment of selecting vocabulary size is conducted to explore what extent it impact on the accuracy of classification, treated as an extension research for multi-class classification.

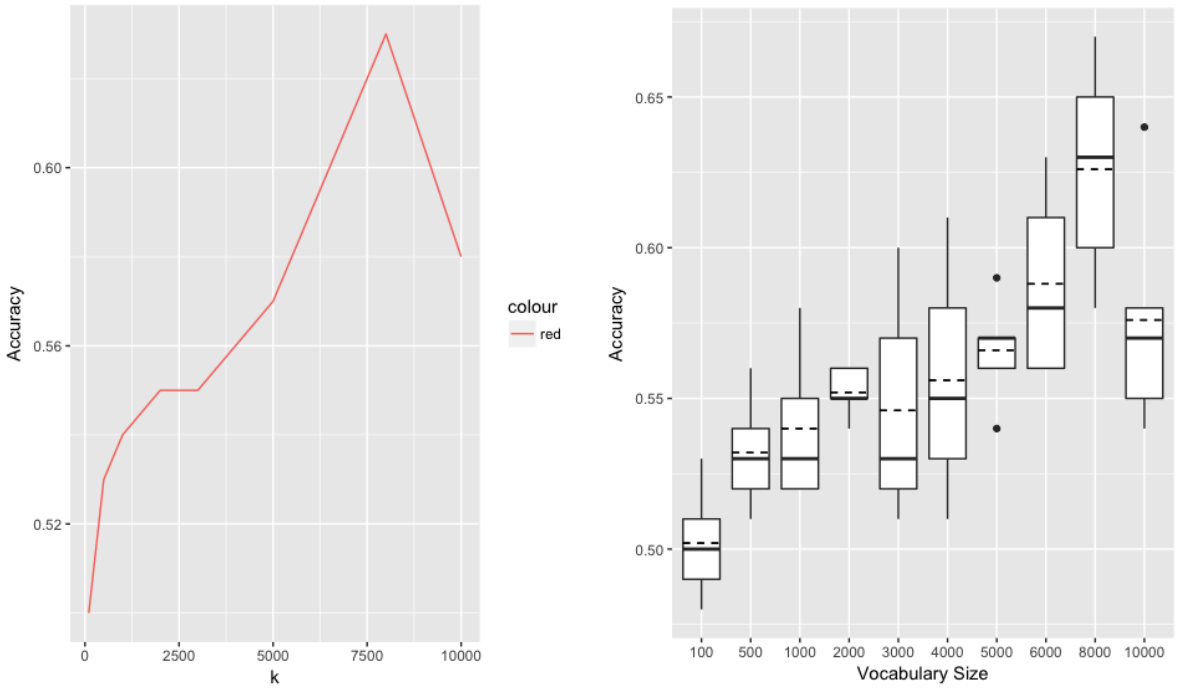
The used data has been designed in table 3.3, which contains ten classes that are bathtub, cormorant, deskglobe, horse, mountain bike, mushroom, mushroom, necktie, penguin, T-shirt and windmill with the same number of instances for each category, that is 82. According to the purpose of this experiment, the vocabulary sizes are defined as 500, 1000, 2000, 3000, 4000, 5000, 6000, 8000, and 10000. Also, the other parameters are the same as the parameters of multi-class classification. In addition, in order to improve the reliability of this experiment, the number of iteration runs increases from 3 to 5.

4.4.2 Results and statistic test

The results of this experiment is shown in table 4.6 and its visualization in figure 4.3, and also the examples of visual word occurrences for an image is provided 4.4. As can be seen, the average accuracy is increased by increasing vocabulary size from 500 to 8000, but there has a slight reduction on average accuracy from 8000 to 10000. As discussed, the fewer vocabulary size could cluster the less SURF descriptors. Therefore, the increase of vocabulary size improves the performance of classification using SURF within the proper range. Nevertheless, the exaggerated vocabulary size could excessively cluster the SURF descriptors during the process of creating BoVW model, resulting in the problem of over-modeling.

Vocabulary Size	Average Accuracy
100	0.50
500	0.53
1000	0.54
2000	0.55
3000	0.55
4000	0.56
5000	0.57
6000	0.59
8000	0.63
10000	0.58

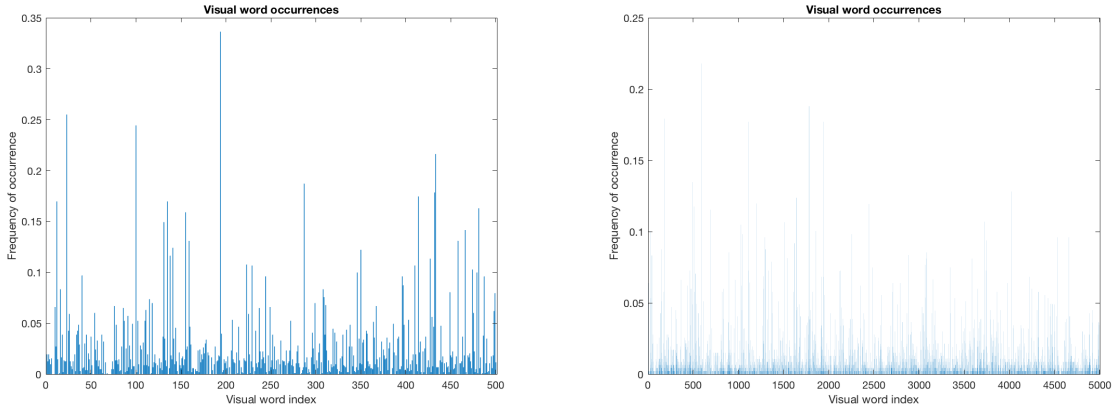
Table 4.6: Accuracy for increasing Vocabulary Size



(a) The trend of changing vocabulary size

(b) Distribution with median and mean

Figure 4.3: The visual results for selecting vocabulary size experiment



(a) 500 vocabulary size

(b) 5000 vocabulary size

Figure 4.4: The comparison of visual word occurrences for an image in bathtub category

On the basis of the box plot 4.4b above, the black lines stand for the median value for each box, and the dashed lines stand for the mean value for each box as well. As we can see, the most of mean values are the more significant than median values. It indicates that the distribution of average accuracy is bias to the positive skew. Furthermore, the dispersion of the average accuracy is slightly sparse at the point of 3000, 4000, and 8000. In a word, it performs the best on classification using SURF descriptors when the vocabulary size is 8000 in this experiment. Besides, the comparison of visual word occurrences for an image that belongs to the category of bathtub is shown in figure 4.4.

The process of statistical test cannot be ignored unless the samples are adequate and the results are explicit. As discussed, the Friedman test is used to test whether there was a significant difference between more than two approaches over multiple datasets, and also it has been conducted in the experiment of binary classification that yields the robust result. Furthermore, Vázquez et al. (2001) was proposed to use Friedman test to find the differences on the single dataset, producing the excellent results as well. Therefore, the Friedman test and Nemenyi test are conducted to find out the statistical significance differences in the results, given by the experiment of selecting vocabulary size. In addition, the threshold of P is set as 0.05 as well.

In Friedman test, χ_F^2 is calculated by the equation 2.5 with their ranks. The result of χ_F^2 is 24.98, and the values of ranks are 1.6, 3.3, 4.3, 5.3, 4.4, 5.4, 6.5, 7.2, 9.4, and 7.4 for the relevantly predefined vocabulary size. Also, the P value is 0.03, which is less than 0.05. Then, the null hypothesis is rejected, and conclude that there was a significant difference in the changing of vocabulary size during the process of constructing the BoVW model in classification. After that, in Nemenyi test, the value of the critical difference is calculated as 6.06. According to the same process of statistical test for binary classification, the result is that there has a significant difference between 8000 vocabulary size and the others based on the value of the critical difference. To sum up, in this experiment, the factor of selecting vocabulary size within the procedure of BoVW model can impact on the accuracy rate of image classification. However, the over-sized visual vocabulary could lead to the issue of the substantial misclassification when BoVW model is used in image classification.

In a nutshell, it is explicit that the vocabulary size has a critical impact on the classification performance. With the vocabulary size increases from 200 to 10,000, the performance grows drastically at first, then peaks at the points (0.63), and after that either levels off or drops mildly. In spite of the optimal vocabulary size dependent on the selection of samples, it suggests exploring for the optimal one among relatively large vocabularies.

4.5 Conclusion

This chapter provided the details of implementation for all experiments that are binary classification, multi-class classification, and selecting vocabulary size in the aspects of descriptions, parameter settings, result, demonstration and evaluation. And the results obtained by experiments are proposed to answer the research question, provided in chapter 1.

Furthermore, the visualizations of the critical results, known as plots and tables, are provided to view in the simple way, such as the accuracy trend of the increasing number of classes, the accuracy trend and distribution of the increasing the size of

vocabulary and some histograms of visual word occurrences. In addition, all results generated by the implemented experiments are involved in the stage of statistical test in order to ensure their reliability and robustness.

The next chapter, namely conclusion, will overview this research at the high level, including problem definition, the results summary, the strength and limitation, and future work.

Chapter 5

Conclusion

5.1 Research overview

This research investigates the comparison of supervised image classification between the SIFT and SURF local descriptors with BoVW representation and more traditional techniques, which are raw pixel and color histogram. At the beginning of the study, the existing research work was viewed through many peer-reviewed scientific papers related to the fields of image processing, BoVW model, supervised machine learning and statistical test.

The experiments were designed and performed to find out the best feature extraction approach that can deliver the most accurate in image classification whatever in binary or multi-class, and also the proper and powerful statistical tests were conducted to prove whether the statistically significant differences appear.

On the basis of the research question, the research attempted to perform the experiment of selecting vocabulary size as an extension of the research in order to explore the insights into the process of constructing BoVW model. Furthermore, the relevant discussion and analysis about the generated results are presented and explained as well.

5.2 Problem definition

As mentioned in the literature review, the main drawback of baseline approaches, namely raw pixel and color histograms, for image classification is that the representation is dependent of the pixel or color of the object being studied, ignoring its shape and texture. Also, the baseline approaches can potentially be identical for two images with different object content which happens to share the information of pixel and color. Conversely, without spatial or shape information, similar objects of different pixel or color may be indistinguishable based solely on pixel or color comparisons. However, a good image classification model must be invariant to the cross product of all these variations, while simultaneously retaining sensitivity to the inter-class variations. However, the BoVW model can fit this standard to yield the better performance for image classification.

5.3 Experiment results and evaluation

In this research, there were three experiments conducted to answer the research question, which is stated in chapter 1. In the first experiment, namely binary classification experiment, the process of image classification was performed by the different feature extraction approaches using the linear SVM algorithm, which are raw pixel and color histogram as the baseline, and SIFT and SURF as state-of-the-art approach. The results reveal that the BoVW model with SIFT and SURF descriptors outperform the baseline in image classification, and also the relevant statistical tests prove that there was a significant difference between the BoVW model and baseline approaches in image classification.

Moreover, the multi-class classification experiment, regarded as an extension to binary classification experiment, was conducted to build the BoVW model with the SURF to classify images over multi-class datasets. The results indicate that the number of classes can impact on the performance of image classification, and also the challenge of similar image classification is still a severe research problem.

Furthermore, the last experiment of selecting vocabulary size was proposed to find

out how the various vocabulary size in the process of building the BoVW model impact on the results of image classification. The results indicate that the slightly increased size of visual vocabulary can improve the accuracy of image classification. However the over-sized visual vocabulary could enlarge the problem of misclassification. In a nutshell, according to the results and findings from the experiments, the research question can be answered as :

The feature extraction approaches of SIFT and SURF with the BoVW model can outperform the feature extraction approaches of raw pixel and color histogram for image classification using the linear SVM algorithm.

5.4 Strength and limitation

The strength of this research is shown as follows.

- The comparison of the performance based on the different feature extraction approaches for image classification was studied by conducting the well-designed experiments and evaluating their results.
- The data source, namely Caltech-256, is reliable and robust for the task of image classification. And also, the experiments were conducted by randomly selecting the sub-datasets from Caltech-256. It significantly reduces the influence in the performance of image classification through randomly sampling. Furthermore, the statistical tests were performed to to prove the correctness of conclusions statistically.
- The research was studied from binary classification to multi-class classification, which can provide the more reliable evidence for image classification using the different feature extraction approaches.
- The factor of vocabulary size during the process of creating BoVW model was considered and proved by conducting the experiment.

Admittedly, there also have some limitations in this research and show them below.

- Only a single data source that is Caltech-256 was used in the empirical experiments. Due to the different methods of data collection for creating the data source, it could impact on the accuracy of image classification using a single data source, not multiple data sources.

- Only linear SVM classifier was performed to the process of training and classifying image features. The other kernel-based SVM classifier and even the other supervised machine learning algorithms could impact on the results in image classification.

- Only SIFT and SURF descriptors were considered to build the BoVW model for image classification. Due to the high similarity between SIFT and SURF, there has no absolute comparison for the performance of local descriptors. Perhaps, the other local descriptors should be considered

5.5 Future work

This research only concentrates on the linear SVM classifier to process the image classification experiments. However, the other kernels within SVM are also well-known to be used in image classification, such RBF kernel and polynomial kernel. The future work could be conducted by investigating how the kernel trick with SVM impacts on the performance of image classification.

Furthermore, in this research, only few classes were involved in binary classification and multi-class classification experiments on the single data source, which is Caltech-256. Therefore, the recommendation of future work is that the various image data sources should be used in the task of image classification in order to enlarge the samples.

Moreover, in this research, only two local descriptors were used in image classification experiments, such as SIFT and SURF. However, one of the well-known local descriptors, called HOG, has been used in the research of image classification based on BoVW model (Dalal & Triggs, 2005). The future work can consider to build the

BoVW model using the HOG descriptors and to investigate the differences among the local descriptors that are SIFT, SURF, and HOG.

On the basis of the limitation of BoVW model, it ignores the spatial relationships among the patches, which are highly vital in image representation. However, the approach of spatial pyramid match performs pyramid matching by partitioning the image into increasingly fine sub-regions, and compute histograms of local features inside each sub-region (Lazebnik et al., 2006). The future work could consider this approach to build the more complex and powerful BoVW model for improving the performance in image classification.

References

Amaratunga, D., Cabrera, J., & Lee, Y.-S. (2008). Enriched random forests. *Bioinformatics*, *24*(18), 2010–2014.

Bay, H., Tuytelaars, T., & Van Gool, L. (2006). Surf: Speeded up robust features. *Computer vision–ECCV 2006*, 404–417.

Bosch, A., Zisserman, A., & Munoz, X. (2007). Image classification using random forests and ferns. In *Computer vision, 2007. iccv 2007. ieee 11th international conference on* (pp. 1–8).

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Chatzichristofis, S. A., Zagoris, K., Boutalis, Y. S., & Papamarkos, N. (2010). Accurate image retrieval based on compact composite descriptors and relevance feedback information. *International Journal of Pattern Recognition and Artificial Intelligence*, *24*(02), 207–244.

Commons, W. (2016). *File:odd-eyed cat histogram.png — wikimedia commons, the free media repository.* Retrieved from `https://commons.wikimedia.org/w/index.php?title=File:Odd-eyed_cat_histogram.png&oldid=223113973` ([Online; accessed 2-January-2018])

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273–297.

Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, eccv* (Vol. 1, pp. 1–2).

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer vision and pattern recognition, 2005. cvpr 2005. ieee computer society conference on* (Vol. 1, pp. 886–893).

Daliri, M. R. (2012). Automated diagnosis of alzheimer disease using the scale-invariant feature transforms in magnetic resonance images. *Journal of medical systems*, *36*(2), 995–1000.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, *7*(Jan), 1–30.

Dey, N., Nandi, P., Barman, N., Das, D., & Chakraborty, S. (2012). A comparative study between moravec and harris corner detection of noisy images using adaptive wavelet thresholding technique. *arXiv preprint arXiv:1209.1558*.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, *10*(7), 1895–1923.

Dollar, P., Wojek, C., Schiele, B., & Perona, P. (2012). Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, *34*(4), 743–761.

Ergene, M. C., & Durdu, A. (2017). Robotic hand grasping of objects classified by using support vector machine and bag of visual words. In *Artificial intelligence and data processing symposium (idap), 2017 international* (pp. 1–5).

Fatahi, M., Speck, O., et al. (2015). Magnetic resonance imaging (mri): A review of genetic damage investigations. *Mutation Research/Reviews in Mutation Research*, *764*, 51–63.

Fei-Fei, L., Fergus, R., & Perona, P. (2007). Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding*, *106*(1), 59–70.

Fisher, R. A. (1956). Statistical methods and scientific inference.

Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on geoscience and remote sensing*, *42*(6), 1335–1343.

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, *11*(1), 86–92.

Griffin, G., Holub, A., & Perona, P. (2007). Caltech-256 object category dataset.

Gualtieri, J. A., & Cromp, R. F. (1999). Support vector machines for hyperspectral remote sensing classification. In *27th aipr workshop: Advances in computer-assisted recognition* (Vol. 3584, pp. 221–233).

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Alvey vision conference* (Vol. 15, pp. 10–5244).

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162.

Hassaballah, M., Abdelmgeid, A. A., & Alshazly, H. A. (2016). Image features detection, description and matching. In *Image feature detectors and descriptors* (pp. 11–45). Springer.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8), 832–844.

Horster, E., & Lienhart, R. (2007). Fusing local image descriptors for large-scale image retrieval. In *Computer vision and pattern recognition, 2007. cvpr'07. ieee conference on* (pp. 1–8).

Hou, J., Kang, J., & Qi, N. (2010). On vocabulary size in bag-of-visual-words representation. In *Pacific-rim conference on multimedia* (pp. 414–424).

Jegou, H., Perronnin, F., Douze, M., Sánchez, J., Perez, P., & Schmid, C. (2012). Aggregating local image descriptors into compact codes. *IEEE transactions on pattern analysis and machine intelligence*, *34*(9), 1704–1716.

Johnson, J. (2017). *Cs231n convolutional neural networks for visual recognition.* Retrieved 25 Dec. 2017, from `http://cs231n.github.io/classification/`

Kamavisdar, P., Saluja, S., & Agrawal, S. (2013). A survey on image classification approaches and techniques. *International Journal of Advanced Research in Computer and Communication Engineering*, *2*(1), 1005–1009.

Khaliq, W., Blakeley, C., Maheshwaran, S., Hashemi, K., & Redman, P. (2010). Comparison of a pacs workstation with laser hard copies for detecting scaphoid fractures in the emergency department. *Journal of digital imaging*, *23*(1), 100–103.

Konstantinidis, K., Gasteratos, A., & Andreadis, I. (2005). Image retrieval based on fuzzy color histogram processing. *Optics Communications*, *248*(4), 375–386.

Kurian, J., & Karunakaran, V. (2012). A survey on image classification methods. *International Journal of Advanced Research in Electronics and Communication Engineering*, *1*(4), pp–69.

Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer vision and pattern recognition, 2006 ieee computer society conference on* (Vol. 2, pp. 2169–2178).

Liu, S., & Bai, X. (2012). Discriminative features for image classification and retrieval. *Pattern Recognition Letters*, *33*(6), 744–751.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, *60*(2), 91–110.

Manjunath, B. S., & Ma, W.-Y. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, *18*(8), 837–842.

Mashao, D. J. (2003). Comparing svm and gmm on parametric feature-sets. In *Proceedings of the 14th annual symposium of the pattern recognition association of south africa.*

Melgani, F., & Bruzzone, L. (2004). Classification of hyperspectral remote sensing images with support vector machines. *IEEE Transactions on geoscience and remote sensing*, *42*(8), 1778–1790.

Mikolajczyk, K., Leibe, B., & Schiele, B. (2005). Local features for object class recognition. In *Computer vision, 2005. iccv 2005. tenth ieee international conference on* (Vol. 2, pp. 1792–1799).

Mikolajczyk, K., & Schmid, C. (2004). Scale & affine invariant interest point detectors. *International journal of computer vision*, *60*(1), 63–86.

Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE transactions on pattern analysis and machine intelligence*, *27*(10), 1615–1630.

Miksik, O., & Mikolajczyk, K. (2012). Evaluation of local detectors and descriptors for fast feature matching. In *Pattern recognition (icpr), 2012 21st international conference on* (pp. 2681–2684).

Moosmann, F., Triggs, B., & Jurie, F. (2007). Fast discriminative visual codebooks using randomized clustering forests. In *Advances in neural information processing systems* (pp. 985–992).

Nadeau, C., & Bengio, Y. (2000). Inference for the generalization error. In *Advances in neural information processing systems* (pp. 307–313).

Nemenyi, P. (1962). Distribution-free multiple comparisons. In *Biometrics* (Vol. 18, p. 263).

Nicosevici, T., & Garcia, R. (2012). Automatic visual bag-of-words for online robot navigation and mapping. *IEEE Transactions on Robotics*, *28*(4), 886–898.

Novak, C. L., & Shafer, S. A. (1992). Anatomy of a color histogram. In *Computer vision and pattern recognition, 1992. proceedings cvpr'92., 1992 ieee computer society conference on* (pp. 599–605).

Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(7), 971–987.

Perronnin, F. (2008). Universal and adapted vocabularies for generic visual categorization. *IEEE Transactions on pattern analysis and machine intelligence*, *30*(7), 1243–1256.

ping Tian, D., et al. (2013). A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, *8*(4), 385–396.

Provost, F. J., Fawcett, T., Kohavi, R., et al. (1998). The case against accuracy estimation for comparing induction algorithms. In *Icml* (Vol. 98, pp. 445–453).

Qin, L., Zheng, Q., Jiang, S., Huang, Q., & Gao, W. (2008). Unsupervised texture classification: Automatically discover and classify texture patterns. *Image and Vision Computing*, *26*(5), 647–656.

Rueda, A., Arevalo, J., Cruz, A., Romero, E., & González, F. A. (2012). Bag of features for automatic classification of alzheimers disease in magnetic resonance images. In *Iberoamerican congress on pattern recognition* (pp. 559–566).

Schaffalitzky, F., & Zisserman, A. (2002). Multi-view matching for unordered image sets, or how do i organize my holiday snaps?. *Computer VisionECCV 2002*, 414–431.

Schütze, H., Hull, D. A., & Pedersen, J. O. (1995). A comparison of classifiers and document representations for the routing problem. In *Proceedings of the 18th annual international acm sigir conference on research and development in information retrieval* (pp. 229–237).

Sinha, U. (2017). *Sift: Theory and practice: Log approximations - ai shack - tutorials for opencv, computer vision, deep learning, image processing, neural networks and artificial intelligence.* Retrieved 17 Oct. 2017, from `http:http://aishack.in/tutorials/sift-scale-invariant-feature-transform-log-approximation/`

Stottinger, J., Hanbury, A., Sebe, N., & Gevers, T. (2012). Sparse color interest points for image retrieval and object categorization. *IEEE Transactions on Image Processing*, *21*(5), 2681–2692.

Stricker, M. A., & Orengo, M. (1995). Similarity of color images. In *Storage and retrieval for image and video databases (spie)* (Vol. 2420, pp. 381–392).

Sural, S., Qian, G., & Pramanik, S. (2002). Segmentation and histogram generation using the hsv color space for image retrieval. In *Image processing. 2002. proceedings. 2002 international conference on* (Vol. 2, pp. II–II).

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International journal of computer vision*, *7*(1), 11–32.

Torralba, A., Fergus, R., & Freeman, W. T. (2008). 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, *30*(11), 1958–1970.

Tsai, C.-F. (2012). Bag-of-words representation in image annotation: A review. *ISRN Artificial Intelligence*, *2012*.

van de Sande, K. E., Gevers, T., & Snoek, C. G. (2011). Empowering visual categorization with the gpu. *IEEE Transactions on Multimedia*, *13*(1), 60–70.

Vázquez, E. G., Escolano, A. Y., Riaño, P. G., & Junquera, J. P. (2001). Repeated measures multiple comparison procedures applied to model selection in neural networks. In *International work-conference on artificial neural networks* (pp. 88–95).

WANG, H.-M., & Shi, P. (2006). Methods to extract images texture features [j]. *Journal of Communication University of China (Science and Technology)*, *1*(008).

Wang, X.-Y., Wu, J.-F., & Yang, H.-Y. (2010). Robust image retrieval based on color histogram of local feature regions. *Multimedia Tools and Applications*, *49*(2), 323–345.

Xu, S., Fang, T., Li, D., & Wang, S. (2010). Object classification of aerial images with bag-of-visual words. *IEEE Geoscience and Remote Sensing Letters*, *7*(2), 366–370.

Zare, M. R., Seng, W. C., & Mueen, A. (2013). Automatic classification of medical x-ray images. *Malaysian Journal of Computer Science*, *26*(1).

Zhang, D., Wong, A., Indrawan, M., & Lu, G. (2000). Content-based image retrieval using gabor texture features. *IEEE Transactions PAMI*, 13–15.

Appendix A

SPSS Output

A.1 Binary classification experiment

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
RawPixel	21	.6567	.08748	.45	.78
Color	21	.7624	.05612	.68	.89
SIFT	21	.9024	.04560	.79	.96
SURF	21	.9157	.03091	.85	.98

Friedman Test

Ranks

	Mean Rank
RawPixel	1.00
Color	2.00
SIFT	3.43
SURF	3.57

Test Statistics[a]

N	21
Chi-Square	56.829
df	3
Asymp. Sig.	.000

a. Friedman Test

Figure A.1: The output of descriptives and Friedman test for binary classification experiment

A.2 Selecting vocabulary size experiment

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
@100	5	.5020	.01924	.48	.53
@500	5	.5320	.01924	.51	.56
@1000	5	.5400	.02550	.52	.58
@2000	5	.5520	.00837	.54	.56
@3000	5	.5460	.03782	.51	.60
@4000	5	.5560	.03975	.51	.61
@5000	5	.5660	.01817	.54	.59
@6000	5	.5880	.03114	.56	.63
@8000	5	.6260	.03647	.58	.67
@10000	5	.5760	.03912	.54	.64

Figure A.2: The output of descriptives for the experiment of selecting vocabulary size experiment

Friedman Test

Ranks

	Mean Rank
@100	1.60
@500	3.30
@1000	4.50
@2000	5.30
@3000	4.40
@4000	5.40
@5000	6.50
@6000	7.20
@8000	9.40
@10000	7.40

Test Statistics[a]

N	5
Chi-Square	24.980
df	9
Asymp. Sig.	.003

a. Friedman Test

Figure A.3: The output of Friedman test for the experiment of selecting vocabulary size experiment