FAKULTÄT FÜR !NFORMATIK
Faculty of Informatics

Invariant Image Representations for Object Category-Based Image Classification

PhD THESIS

submitted in partial fulfillment of the requirements of

Doctor of Technical Sciences

within the

Vienna PhD School of Informatics

by

Hafeez Anwar
Registration Number 1128889

to the Faculty of Informatics
at the Vienna University of Technology

Advisor: Priv. Doz. Dr. Martin Kampel
Second advisor: Dr. Sebastian Zambanini

External reviewers:
Name Lastname. Affiliation, Country.
Name Lastname. Affiliation, Country.

Wien, TT.MM.JJJJ

(Hafeez Anwar)

(Priv. Doz. Dr. Martin Kampel)

Declaration of Authorship

Hafeez Anwar

I hereby declare that I have written this Doctoral Thesis independently, that I have completely specified the utilized sources and resources and that I have definitely marked all parts of the work - including tables, maps and figures - which belong to other works or to the internet, literally or extracted, by referencing the source as borrowed.

(Place, Date) (Signature of Author)

Abstract

The thesis deals with the problem of invariance in image representations for object category-based image classification. The variations in object images caused by various factors such as changes in scale, position and orientation of the objects makes object category-based image classification a challenging task. The so-called invariant local descriptors are widely used to represent the images as a set of image patch descriptions to achieve robustness to such image variations. However, these descriptors are calculated locally and thus are unable to capture the global image structure. To this end, the work presented in this thesis aims to develop discriminative global image representations by deriving the relationships of object features in a way that is highly insensitive to several image variations such as changes in scale, position or in-plane rotations.

These global image representations are applicable to object images having minimal or no background clutter as well as those with severe background clutter. The task of image-based classification of ancient coins is taken as a motivating example for the first group as they are imaged on homogeneous background. Nevertheless, coin images have challenging variations which are differences in object scale, position and orientation. A global invariant representation is developed to cope with such variations where, as a first step, the coin images are automatically segmented, cropped and normalized to acquire scale- and translation-invariance. The segmented image is then partitioned into circular regions from which rotation-invariant local features are sampled, thus achieving rotation-invariance both locally and globally. It is shown that the circular partitioning of the segmented image proves more robust to image rotations than other partitioning strategies such rectangular and radial-polar. In case of severe background clutter in the images, automatic object image segmentation is hard to achieve, as for instance for the natural images of butterflies which may also contain flowers, leaves and trees. In this case, a global image representation is achieved by deriving the invariant geometric relationships of the local invariant features. For this purpose, triangulation is performed among the positions of the local features in the 2D image space. Since the angles and side ratios of a triangle are scale- and rotation-invariant, the global image representation based on the triangulation of local features is also invariant to changes in object scale, position and orientations. The trained object model is made more discriminating by using the local features from the foreground and rejecting the background information. In the presence of image variations caused by changes in object scale, position and orientation the image representation based on invariant geometric relationships of the local features resulted in improved performance.

Contents

CHAPTER 1

Introduction

Images depict contents such as faces of people, objects and scenes. The two dimensional images are represented numerically by the digital images in computers. Thus, the contents of an image, be that the delightful colors of flowers or the astonishing colorful patterns of the Northern Lights, are mere numeric values in the two dimensional matrix unless they are interpreted. Based on such interpretation the image is then classified into a genre.

With advancements in the World Wide Web, it has become the largest global repository of images that belong to various fields such as medical sciences, astronomy, fashion and business etc. On daily basis, the number of images shared by individuals and organizations using the online sources such as the social media, ranges roughly from thousands to millions. The analysis or classification of these images can uncover useful information to the human users about their respective domains. Manual classification of such a huge number of images is not an option because it requires tedious human efforts and tremendous amount of resources and time. Alternatively, effective computer tools can be developed that can automatically classify images based on their contents. Unfortunately, there are insufficient tools to classify such a huge number of images based on their contents given the fact that the contents vary from one another due to image variations caused by factors like geometric transformations.

Due to these reasons, the *image classification* has become a core problem in the field of computer vision. It deals with the contents-based categorization of images into two or more disjoint groups [28]. For instance, a given face image of a person is classified based on a number of categories such as the gender [25], race [47], expression [56] or even the age [48]. The retrieval of images based on their contents from a huge image database (content-based image retrieval) [89] also uses a notion of image classification where images are grouped together into one or more categories based on their contents. Another example is that of the video surveillance such as the one used for ambient assisted living where the fall detection [73] of the elderly is performed by classifying the frames into categories based on the position of the subject. On the industrial side, in order to ensure the quality control in the production, the automatic inspection of products [85] such as the printed circuit boards (PCB) uses images of the circuit boards to classify them as

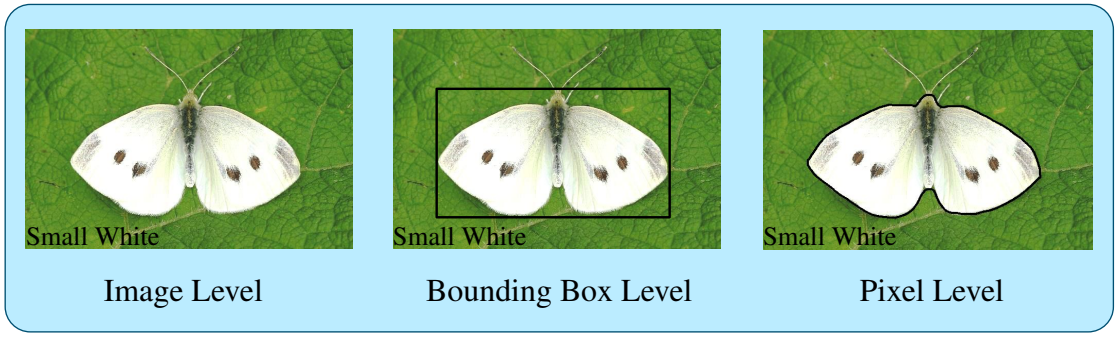


Figure 1.1: Levels of detail achieved by object recognition frameworks

acceptable or non acceptable. These examples demonstrate the wide meaning and technical applicability of the term *image classification* in various domains of computer vision.

Humans have the ability to recognize objects due to their shapes, appearances and their respective contexts. To make computers able to recognize objects in a similar manner within images is the aim of *visual object recognition*. The recognition of an *object category or class* deals with the recognition of instances of objects belonging to the same category. For instance, cars may be of various shapes and appearances but due to their single generic picture in the human mind, all of them are treated under the same category or class. A recognition framework designed to recognize cars of any appearance and shape is an example of *object category recognition* or *object class recognition*. For a given image, various object category recognition frameworks can provide information about the presence of a particular object class at the following levels of detail [28] which are also depicted in Figure 1.1.

- *Image level:* name the object category present in a given image. This is also called object category-based *image classification*.
- *Bounding box level:* coarsely approximate the rectangular area in the image in which the object is present.
- *Pixel level:* detect the pixels of the image that belong to the object area.

Thus the object category-based image classification comes at the *image level*. It deals with *naming* the image according to the category of the object instance depicted in it. Due to this reason, in the rest of the thesis, the term *image classification* is used for object category-based image classification.

Classification of any given scene, object or face by the vision system of humans is faster, effortless and robust to certain challenges such as viewpoint, clutter and occlusion. Achieving such level of image classification for a machine would solve a great number of practical problems. Due to this reason, the problem of image classification in the field of computer vision has gained the attention of researchers since the last few decades [28]. While certain practical problems such as the image-based inspection of printed circuit boards (PCB) [85], image-based sorting of mails and parcels [85] and optical character recognition (OCR) are solved to a greater

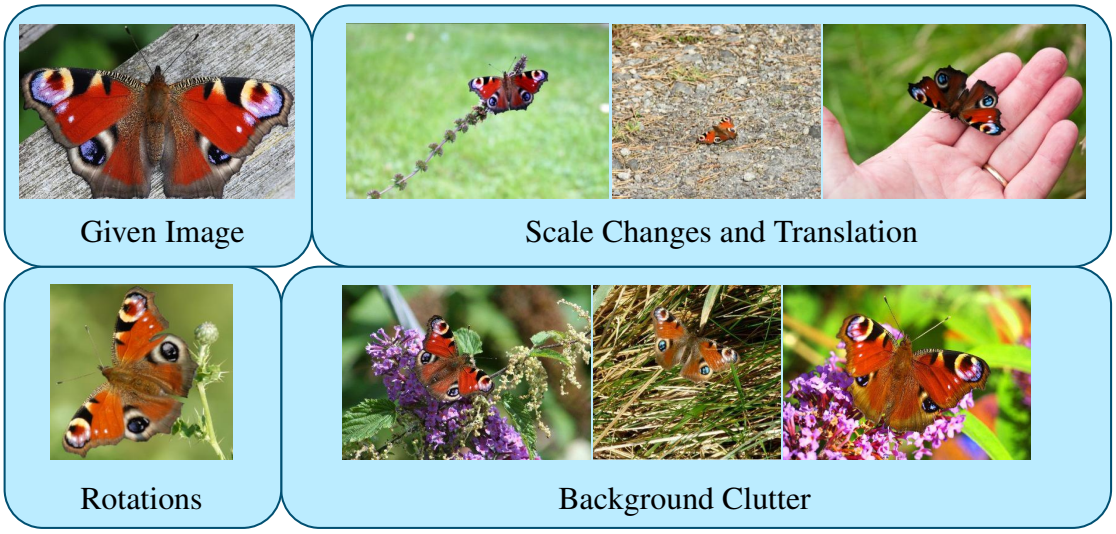


Figure 1.2: Various challenges related to imaging

extent, there are still other image classification problems that need extensive research work to overcome the difficulties and challenges that mainly arise due to the variations in the image contents. Thus, related to the task of image classification is the problem of *image variations* [28].

In the context of object category-based image classification, the variations in the two images that depict instances of the same object mean that they both differ from each other due to certain factors that affect their contents. Being robust to image variations is related to the concept of *invariance*. In computer vision, invariance is the property of an outcome to remain unchanged to a certain degree under a defined set of image variations [24]. In the context of image classification, achieving invariance to a given set of image variations means that the task of classifying the given image is not or least affected by these image variations. Consequently, the set of factors that cause variations in the images for a given application area can be seen as challenges in that particular area. These challenges can affect the discriminating power of any image classification framework and therefore should be identified before hand. The challenges can broadly be divided into the following two groups [28]:

- **Challenges related to imaging**

  Images of the instances belonging to the same object category can vary greatly due to the variance in the imaging conditions. Following are the main challenges related to imaging which are also shown in Figure 1.2 in case of butterflies.

  Scale: Number of pixels covered by the object can vary from one image to another depending on the distance of the camera from the object. If the object is far away from the camera, it will appear smaller as compared to the object that is near to the camera. Due to this variation, images that depict the instances of the same object

category will vary from each other. Therefore, the task of image classification must be robust to changes in scale of the object.

Translation: The change of the object pixels coordinates in the image plane cause translation. Images of the instances of a given object category vary from each other if the coordinates of object pixels in the 2D image space differ.

Rotations: The rotations of objects can vary from one image to another. Here, the rotation means in-plane rotation where the object is rotated around the axis perpendicular to the image plane. The image of an instance of a given object category varies from another image that depict the rotated instance of the same object category. Due to this reason, the invariance to rotations should be achieved by the task of image classification which means that the variations caused by image rotations will have no or minimal effect on the image classification results.

Background clutter: Image content other than the object instance is called the background clutter. Images of the instances of a given object class can vary significantly due to their background. This problem is severe in natural images where objects are imaged along with complex backgrounds. For instance, images of butterflies are taken along with flowers, leaves or trees.

- **Challenges related to objects**

  Apart from imaging, objects themselves present certain challenges and images of the instances of the same object category can vary to a certain degree depending on the following factors.

  Deformations: Images of the instances of the same object category may vary due to the variance in their shapes and textures. For instance, the relative positions of the texture on the wings of butterflies vary due to the movement of the wings. Specimens of a given class of ancient coins can also vary due to the deformations caused by the variations in the hand-made dies used in the process of minting [106].

  Insufficient data: Missing of certain object parts leads to variations in the images of the instances of the same object class. As an example, the wear and tear of the ancient coins caused by various factors results in the absence of certain parts that are vital for visual classification [109]. Such lack of data can also be caused by uncontrollable imaging factors such as partial occlusions in case of natural images, e.g. some part of a wing of a butterfly occluded by leaves or by the other wing.

  Figure 1.3 shows various instances of the same object category that differ from one another due to deformations and partial occlusions.

The thesis aims to perform the task of image classification which is invariant to the challenges mentioned above. However, contrary to the existing methods that solely rely upon the so-called invariant local features [28], the aim is to develop invariant image representations that are built on top of these local features by modeling their configurations. In a sense, the sole purpose of the thesis is to expand the local features into neighborhoods and configurations of local

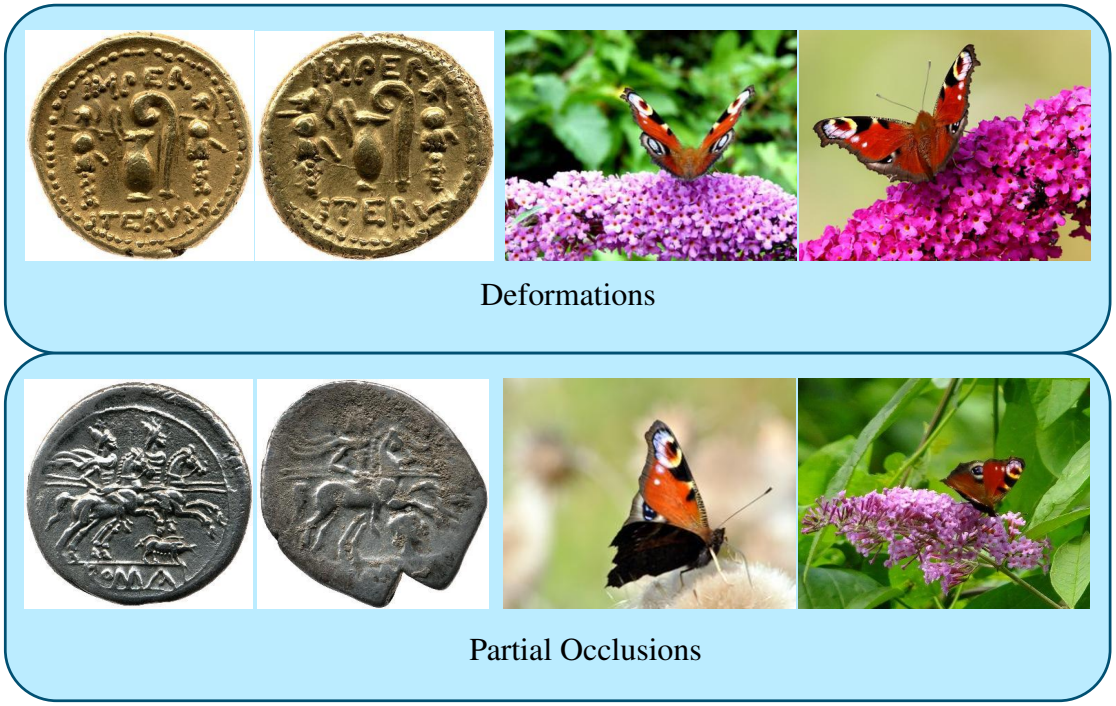


Figure 1.3: Variations caused by deformations and partial occlusions

features thus achieving image representations based on their relative geometry. However, unlike other methods [50] [111] that fail to achieve invariance to the mentioned challenges, the relative geometry of the local features is extracted while still achieving robustness to these challenges. Consequently, individual image representations are developed and evaluated where each one of them either aim at one or more than one challenges simultaneously:

– *Invariance to scale, translation and image rotations is achieved in the **absence** of the background clutter.*
Images of the objects taken with homogeneous or flat background can automatically be segmented to extract the object area [104]. Once segmented, the object image region is then cropped and normalized to achieve *scale-* and *translation-invariance*. The normalized object region is then divided into concentric sub-regions in such away that the center of the concentric sub-regions is aligned with the center of the object image region. Finally, the configurations of the local features in each circular sub-region are modeled to achieve *rotation-invariance* in the resulting image representation.

– *Invariance to scale, translation and image rotations is achieved in the **presence** of the background clutter.*
The severe background clutter in the natural images makes the automatic object segmentation extremely hard. Thus, instead of using object segmentation as a pre-processing step,

the relative geometry of the local features extracted from the whole image is modeled in a triangular manner. Since the angles of a triangle are invariant to changes in *scale*, *translation* and *rotations*, the image representation based on the triangular configurations of the local features will be invariant to such image transformations.

– *To achieve invariance to deformations and insufficient data.*
The local features are computed on local image patches. For the development of the proposed image representation, these local features are densely extracted from the images using a regular grid with a constant pixel stride. The dense extraction of the local features completely captures the underlying structures of the objects. Due to this reason, the feature extraction process achieves adequate robustness to partial occlusions and/or deformations.

– *To reduce the run-time, computationally efficient techniques from computational geometry [12] are used.*
The triangular geometric configurations of local features are achieved by considering identical features only. Since the number of local features densely extracted per image can vary from hundreds to thousands, achieving triangulation among every unique pair of three identical local features is a computationally expensive process. Due to this reason, for triangulation, computationally efficient techniques from computational geometry are used to achieve the image representations in reduced run-time.

1.1 Motivation and Scope of Work

The image classification in this thesis is performed by the predominantly used methodology where an object model is learned from its sample images using machine learning techniques [16]. These training image samples must be labeled carefully under the human supervision due to which they are expensive to obtain [28]. However, as mentioned before, the task of image classification is based on *naming* i.e. assigning a label to an image according to the category of the object depicted in it. Therefore, only weak labeling [28] is required in which the training images are labeled according to the category of the object that they depict. Thus, as compared to pixel-level labeling [28] where each pixel is labeled based on its relationship to the object or to the background, weak labeling is achieved with less effort. In addition to labeling, a lot of training images are needed, typically hundreds or thousands of exemplar images per object class [49]. The object model achieves robustness to image variations by learning from such large number of training images as these images vary from one another due to various factors such as changes in viewpoint and illumination [91] [99]. The online sources such as Google image search, Bing and Flickr made it possible to obtain training images, however, the variations found in these images are not enough to make the learned object model robust to these image variations. For instance, Figure 1.4 shows an example of the Google image search for the "Peacock Butterfly". It can be seen that most of the butterfly images have least in-plane rotation differences. Therefore such images are not enough to learn a model for the image-based recognition of the "Peacock Butterfly" which is robust to larger in-plane rotations.

Due to this issue with the training image samples, the development of the invariant image representations is pursued in the thesis to support the image classification. Such image represen-

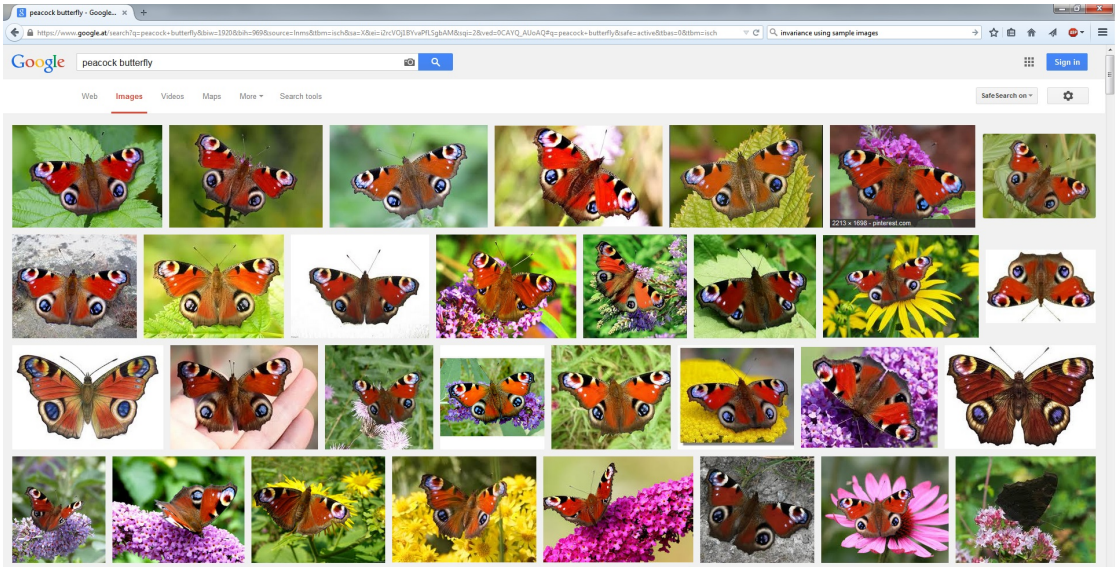


Figure 1.4: An example of the Google image search for the Peacock butterfly showing that the images have minimal differences with respect to the in-plane rotation of the butterfly

tations are derived using certain features of objects such as shape and appearance. These features almost remain unchanged under several image variations such as changes in scale, position or in-plane rotations. Therefore, the image representations based on these features will be invariant to such image variations leading to the invariant image classification. In this context, the so-called local image descriptors have already shown considerable success and achieved invariance to image transformations such as rotations [3] [96], scale [55] and translation [45]. However, these local descriptor capture the local properties of the objects or scenes as they are calculated on local image patches. Instead, due to the proper geometry of the objects and structured scenes, a global representation is preferred that captures the geometric relationship of the local image descriptors.

To acquire an image representation based on such global relative geometry, two lines of research exist where one of them is based on the splitting of the image space [50] [111] and then modeling their relationships. However most of the methods proposed with this idea fail to achieve invariance to certain image variations such as translation and image rotations [28]. The thesis aims to extend this line of research by proposing global image representations which are invariant to such image variations. The other line of research to achieve the global relative geometry is based on the global relationships of the local features [101] [102]. These local features are the representatives of the local image patches on which they are constructed. This method of achieving global geometry is robust to image variations caused by changes in scale, translation and rotation. However, such relationships are often affected by image variations that are caused by deformations and occlusions. In this regard, the thesis aims at combining the best of both the worlds i.e. image space splitting and the geometric relationship of local image

features in each image subspace, to achieve an image representation that is robust to image variations caused by deformations and occlusions.

Another issue is the background clutter found in the images that can affect the discriminating nature of the global image representations developed for object category-based image classification. This is due to the fact that the whole image is considered for the development of the global image representations. The background clutter is very common in natural images. For instance, the images of the butterflies are taken with flowers, trees, leaves and branches etc in the background. In the images taken in controlled environment, the background clutter is minimal. For instance, the images of the ancient coins are taken with homogeneous background. In either case, the effect of the background clutter on the performance of the image representations should be minimized.

In order to evaluate the proposed image representations, both the textured and almost non-textured data is used. The ancient coins are flat objects having least texture. Various objects such as 'dolphin' and 'wolf' are minted on the reverse side of the ancient coins. These objects are non-textured and can only be identified by the contours of their shapes. Based on these objects, the task of the classification of ancient coins is performed using the proposed image representations. The main reason behind the consideration of this task is that the ancient coins can be affected by the aforementioned challenges in the following manner:

- They can be imaged at various places in the image and are thus affected by translation.
- The ancient coins can undergo changes in scale.
- Due to their circular nature, they can exhibit in-plane rotations.
- Due to their manufacturing process and the non-favorable preserving conditions, they can undergo abrasions which cause a high degree of variability within each class.

The task of image-based classification of butterflies is also taken as a motivating example. The butterfly species vary from one another based on the color and texture of their wings. In this thesis, the task of classification of the butterflies is considered due to the following reasons:

- Like the ancient coins, the butterflies are imaged at various places of the image. This causes variations in the images due to translation.
- Based on the distance of the camera from the butterfly, the scale can vary from one butterfly image to another.
- They can exhibit various in-plane rotations thus causing the differences among the images.
- Unlike ancient coins, the butterfly images contain severe background clutter as they are taken in natural environments.

Indeed, apart from the image-based classification of ancient coins and butterflies, the proposed image representations can be used for the image-based classification of other objects. For instance, the image-based classification of fish is one prospective application area where the proposed image representations can effectively be used. Fish are imaged in cluttered background

Figure 1.5: Image-based classification of fish species is a prospective application of the proposed image representations

caused by the objects and other fish in the dense coral reef. The fish species vary from one another due to the texture of their bodies. Similarly, images of the same fish specie vary from one another due to variations caused by changes in scale, position, and in-plane rotations. Examples images of various fish species under different in-plane rotations, at various scales and with different background clutter are shown in Figure 1.5. However, in cases where major portion of the object is occluded, the proposed image representations will not prove helpful. In addition to that, the proposed image representations are robust to in-plane image rotations and do not account for image variations caused by other factors such as affine transformation.

1.2 Research questions

The thesis aims at answering the following research questions.

- The objects imaged on a homogeneous background can be segmented automatically [104]. In this case, how can an image representation be achieved that is invariant to changes in scale, translation and image rotations?

- In case of severe background clutter such as in the natural images, how can the effect of the background clutter be reduced on the image classification results?

– Either in the presence or in the absence of the background clutter, how can the geometry in the appearance and in the shape of the objects be used to achieve an image representation that is invariant to changes in scale, translation and image rotations ?

– Finally, either in the presence or in the absence of the background clutter, how can the deformations of objects and lack of sufficient data be handled?

1.3 Contributions

The common geometry in the shape and appearance of a given object category remains visually consistent amid background clutter, partial occlusions and changes in scale, rotations and translation. In order to cope with all these challenges, the consistency in the geometry of the object category is utilized to derive image representations for invariant image classification. Following are the contributions which are divided with respect to the severe or negligible background clutter.

1.3.1 Negligible background clutter

Background clutter is either completely absent or negligible in images that are taken in controlled environment. As an example, ancient coins are imaged under controlled conditions on a flat background. Therefore the amount of the background clutter in case of ancient coins is almost negligible. In the absence of the background clutter following are achieved.

– *Invariance to changes in scale, position and image rotations:* Automatic segmentation can be used to segment the object in the presence of a flat background. This is due to the fact that most of the information of the image is contained in the object area. For instance, in case of ancient coins the automatic segmentation [104] is used in Anwar et al. [4] to segment the object from the flat background. Such segmentation provides information about the position and scale of the object thus achieving invariance to changes in both scale and translation simultaneously. In addition, for objects that can undergo various rotations, rotation-invariance should be achieved for a better classification rate. For instance, due to the circular nature of ancient coins, rotation-invariance should be achieved both locally and globally after they are segmented from the background. Rotation-invariance is achieved locally by using rotation-invariant local features. The global rotation-invariance is achieved in Anwar et al. [7] by splitting the segmented image into circular regions. The statistics of the local features are then aggregated from each circular region to represent the image. Thus rotation-invariance is achieved both locally and globally. Furthermore, in Anwar et al. [6] [8], the modeling of rotation-invariant geometric relationships of local features within each circular tiling outperformed the circular tiling while still being invariant to image rotations.

– *Invariance to deformations and insufficient data:* In order to deal with the problems of deformations and insufficient data, local features are densely sampled from the image by using a sampling grid with a constant pixel stride in Anwar et al. [7] [8]. Such dense

sampling completely captures the shape of the underlying structure thus accommodating the variations caused by deformations and insufficient data.

1.3.2 Presence of background clutter

Background clutter is severe in natural images such as the images of butterflies which contain leaves, flowers and trees etc. Classification of images is affected by such background clutter. Therefore in the presence of background clutter following is achieved.

- *Reducing the effect of the background clutter:* The effect of the background clutter can be neglected by extracting the object of interest from the image. However, unlike the images of the ancient coins, the severe background clutter in the natural images makes it hard to automatically extract the object of interest from the image. Since the training of the object model is an offline process, to reduce the effect of the background clutter, the objects are manually segmented at the training stage. As an example, the manually generated segmentation masks of butterflies are used. The use of segmentation masks restricts the process of feature extraction thus selecting features from the foreground only. The use of these foreground local features at the stage of training reduces the contaminating effect of the background on the trained model. This results in better classification accuracy as shown by the experimental results of Anwar et al. [5].

- *Local scale- and rotation-invariance:* Objects that are found in natural images can undergo changes in both scale and rotations. For instance, the butterflies can be imaged at various scales with different rotations as shown in Figure 1.4. Due to this reason, the local rotation-invariant features extracted at several scales resulted in improved performance.

- *Global scale- and rotation-invariance:* In addition to local scale- and rotation-invariance, global scale- and rotation-invariance is also desired for an improved performance. Such invariance is implemented in Anwar et al. [5] by using the global geometric relationships of the image patches which are represented by the local features. The positions of the local features in the 2D image space are triangulated and the image is represented using the angles generated by this triangulations. Since the angles and side ratios of a triangle are invariant to changes in scale, rotations and translations, such image representation is scale- and rotation-invariant. Results of the classification experiments performed on the images of butterflies showed robustness to changes in scale and rotations.

- *Reducing the run-time:* The triangulation among the local features is a computationally expensive process as the number of local features can vary from tens to hundreds per image. Therefore, *Delaunay triangulation* is used in Anwar et al. [8] which is an efficient triangulation method from computational geometry. Delaunay triangulation drastically reduces the run-time at the cost of marginal drop in the classification accuracy.

1.4 Innovative and applied aspects of the thesis

On the innovative side, the thesis is concerned with the development of image representations for object category-based image classification that are invariant to changes in scale, translation and image rotations. The effect of the background clutter on these image representations is minimized by using the information from the object image area during the offline training process.

On the applied side, the thesis contributes to the field of *cultural heritage* which deals with the management, analysis and digital preservation of the ancient artifacts. The significance of the field can be observed from the fact that numerous research projects in this field have been funded by reputable funding agencies in order to preserve and analyze the ancient artifacts. For instance, the COINS [103] project was funded by EU to combat the illegal online business and theft of ancient coins. The project was a major success resulting in media coverage and extensive research in the field of digital cultural heritage preservation. Another project related to the ancient coins from the Roman Republican era ILAC [37] was later funded by the Austrian Science Fund (FWF). This project was more concerned with the digital preservation and the image-based classification of the huge collection of ancient coins owned by the Vienna Museum of Fine Arts. Apart from the funded projects, there are several conferences [1] and scientific journals[2] dedicated to this field.

The fundamental work of coin experts is the classification of coins according to standard reference books [14] as this provides additional information such as accurate dating, political background or minting place. However, classifying ancient coins is a highly complex task that

[1]Examples are EUROGRAPHICS Workshop on Graphics and Cultural Heritage, ECCV Workshop on Computer Vision for Art Analysis, ACCV Workshop on e-Heritage and Computer Applications and Quantitative Methods in Archeology Conference (CAA)

[2]Such as the ACM Journal on Computing and Cultural Heritage and IEEE Signal Processing Magazine Special Issue on Signal Processing for Art Investigation

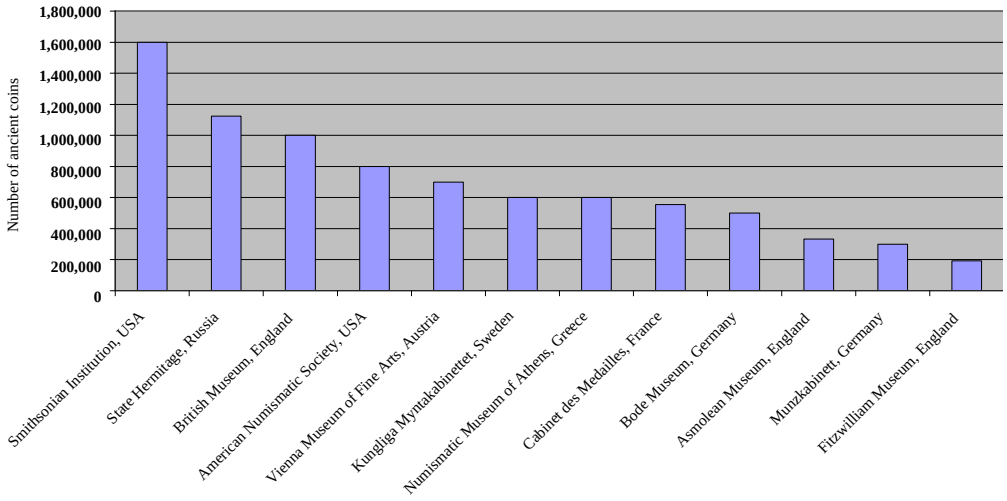


Figure 1.6: Number of ancient coins in some of the world's largest collections. The statistics are given on the official websites of the collections.

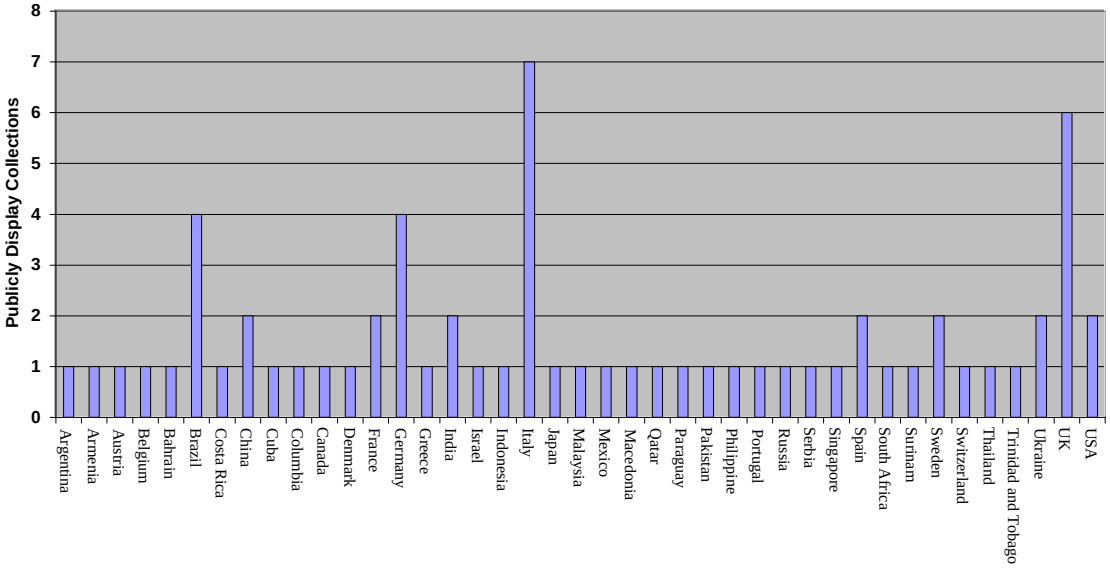


Figure 1.7: Number of publicly displayed ancient coins collection around the world

requires years of experience in the entire field of numismatics. Additionally, efficiency is a vital issue, both among humans and machines, especially when it comes to the processing of large coin collections or may be even larger hoard finds. Therefore, an image-based coin classification framework can support and facilitate a substantial part of numismatic coin analysis and classification

The thesis uses the ancient coins as a motivating example due to the aforementioned reasons. The analysis and classification of ancient coins is based on the recognition of reverse motifs that depict various objects such as animals and ornaments. However, such classification is coarse-grained as any given object might be minted on coins of more than one type. Thus the proposed method can be integrated into an ancient coin classification framework as a first step where it will help to indicate the candidate coin types. The finer classification where the type of the coin is decided, can then be performed by recognizing the obverse side [9] and the legend [36].

A series of techniques proposed by Zambanini et al. [106] [109] deal specifically with the fine-grained classification i.e. image-based classification of ancient coins based on the type or reference number of the coin. The major difference of the research work outlined in the thesis from their work is the degree of classification. They perform fine-grained coin classification based on image matching methods while the work presented in the thesis is more focused on the coarse-grained classification of ancient coins by using offline training methods.

The importance of an image-based coin classification framework can be realized from the number of the ancient coins owned by individuals and museums all over the world. Figure 1.6 shows the number of ancient coins in some of the world's largest collections and Figure 1.7 shows the country-wise number of publicly displayed ancient coins collections around the world. The number of ancient coins is growing day by day with newly discovered coins due to which

it is hard to keep track of every coin type indexed in the reference books [14]. Apart from that, the addition of the newly discovered hoards of ancient coins to the current collection also requires time and expertise. The broad use of digital cameras has led to an exploding number of digitally recorded coins. While computers are extensively used for storing and working on numismatic data, no computer aided classification system for ancient coins, which is based on images, has been investigated so far. Digital images have become the usual way of exchanging information not only for internet auctions but also between scholars, collectors and coin dealers. Classification of a coin from an image is commonplace in numismatics and if the coin is in good condition classification from an image works all the same as from an original specimen. Since specialists are usually able to classify coins from 2D image information only, computer vision methods can be applied to coin images to support numismatists on the examination of coinages as well as to speed up the overall processes significantly.

Due to these reasons, the proposed image-based classification framework can support the manual classification to make it smooth and less time consuming. Figure 1.8 shows the big picture of an application's hardware and software used for image-based classification of ancient coins. The hardware and software of this product are explained:

- **Hardware**

  The hardware of the product consists of a pair of cameras. As a first step, the coin passes between a pair of HD cameras which will take images of both the sides. After classification, the information about the coin such as the issuer, date, reference number and the descriptions of both the sides are displayed on a monitor.

- **Software**

  The main component of the products'software is the trained model stored in the database. The training of the model is performed as an offline process on the exemplar images of the coins. This model is then used to classify the images of a any given coin.

1.5 Structure of the thesis

The thesis is organized as follows.

Chapter 2 outlines state-of-the art on image classification. The Bag-of-visual words (BoVWs) model has been the most prevalent technique used for image classification during the last decade [28]. State of the art image classification results [28] are achieved on various benchmark datasets [18] [20] by modifying the basic BoVWs model. Furthermore, based on the underlying local features, the BoVWs model achieves robustness to image transformations such as scales changes and viewpoint transformations [28]. The thesis extends this technique and proposes enhancements to it in order to achieve the research goals. Thus as a prerequisite, details of each step of the BoVWs model are given with the summaries of various techniques proposed for each step. A basic problem in the BoVWs image representation is the lack of the global spatial information of the local features. It does not considers the spatial information of local features in the 2D image space.

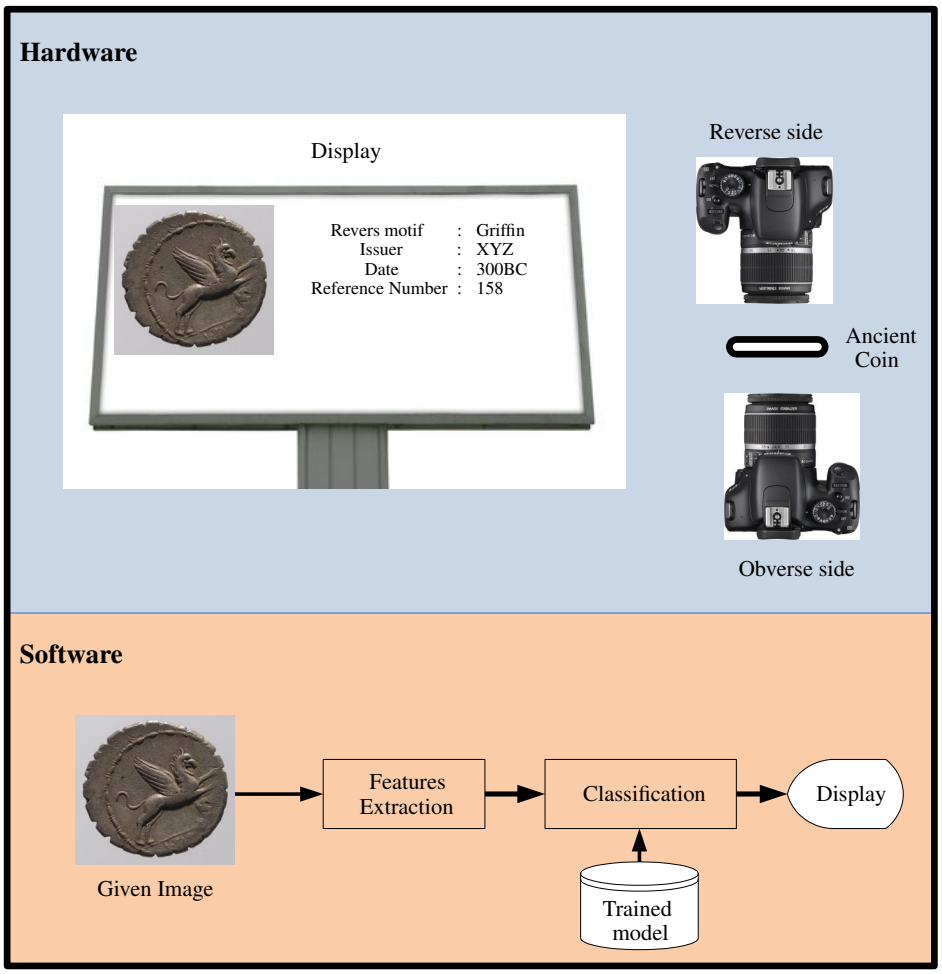


Figure 1.8: Big picture of the hardware and software components of the system for image-based ancient coins classification

Therefore the proposed image representations extend the BoVWs model by inducing this information.

Chapter 3 deals with the development of an image representation which is invariant to changes in scale, position and image rotations in the **absence** of the background clutter. The spatial information is added to the BoVWs model by splitting the image space to enrich the image representation with spatial cues. The image space is split into three types of regions which are rectangular, circular and radial-polar regions [4] [7]. An example is shown in Figure 1.9b where the image space is partitioned into concentric circular regions. Any given tiling strategy is imposed over the BoVWs image representation and then from each individual tiling the statistics of the visual words are extracted. Since the ancient coins are used to evaluate the proposed image representation, as a first step, the automatic segmentation is used to segment the coin from the background using the method proposed

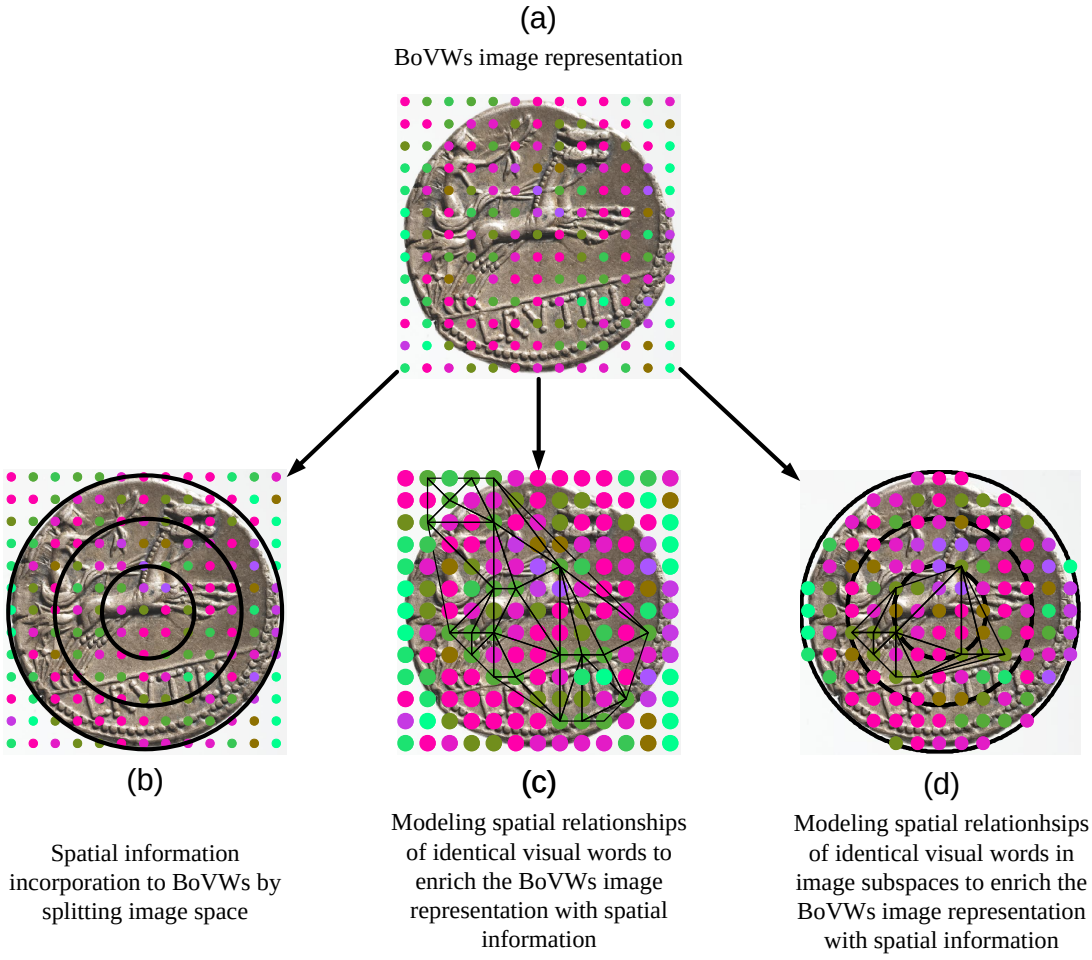


Figure 1.9: Variations in the images of reverse motif due to non-rigid deformations

in [104]. Doing so achieves invariance to changes in scale and position as the segmented coin image is cropped and normalized. Afterward the segmented coin image is split into regions of one of the aforementioned shapes. Then, from each image subspace the statistics of the local features are gathered and used to extend the standard BoVWs image representation. Experimental results demonstrate that the proposed extensions outperforms the BoVWs model for the task of ancient coin classification. In addition, splitting the image space into circular regions proves more robust to image rotations.

Chapter 4 deals with the development of an image representation which is invariant to changes in scale, position and image rotations in the **presence** of the background clutter. The proposed image representation is an extension of the BoVWs model where it adds the spatial information to the BoVWs by modeling the geometric relationships of the local features in a triangular manner [5]. The positions of the local features are triangulated

and the angles between their spatial positions in the 2D image space are used to represent the images. For simplicity, the triangulation among identical visual words is shown in Figure 1.9c in case of ancient coin. The problem of image-based classification of butterflies is used to evaluate the proposed image representation. Experimental results show that the use of the segmentation masks at the stage of training increases the discriminating nature of the proposed image representation. Furthermore, being based on triangulation, the proposed image representation is inherently invariant to changes in scale, position and rotations which is also validated by the experimental results.

Chapter 5 combines the basic ideas of both Chapter 3 and Chapter 4. The proposed image representation is enriched with spatial information by splitting the image space into circular regions and then modeling the triangular geometric relationship of local features in each circular region [6] as shown in Figure 1.9d. Consequently, the angles produced by the triangulation in each circular region are used to represent the image. However, as a first step the coin image is segmented to achieve invariance to changes in scale and position. The proposed image representation is evaluated for the task of ancient coin classification where it outperform the BoVWs model and the circular tiling and proved more robust to coin image rotations.

Chapter 6 concludes the thesis. The achievements of chapters 3, 4 and 5 are outlined. In addition, the shortcomings of each proposed method are also given. Finally, the future directions of the current research are given.

CHAPTER 2

Related work

In this chapter, the preliminary knowledge about the concepts and techniques related to the problem of image classification and specific to the thesis is outlined. In Section 2.1, the basic themes of the techniques that employ the object category recognition-based image classification are given. One of the mostly widely used technique that has become the basic paradigm of the image classification in that last decade is the so-called bag-of-visual-words (BoVWs) model [28]. The work in this thesis also follows this technique and extends it to achieve image representations for invariant image classification. The motivation behind the use of the BoVWs model and its details are given in Section 2.2. State of the art in the image-based ancient coins classification and relevant Numismatic terms are outlined in Section 2.3.

2.1 Image classification

In one of the most prominently used object category-based image classification paradigm, the model of an object category is learned from sample images [16]. Based on such learned model, a given image is labeled according to the category of the object depicted in it. Thus the image classification framework accomplishes the classification task by matching features of a given image against the learned model [57]. Based on these features, the object category-based image classification methods can broadly be divided into three main groups [57]. The appearance-based methods use the features related to the appearance of the objects such as color and textures. The geometry-based methods use the features related to the geometry of the object such as the lines and edges. Finally, the local-invariant features-based methods use the local-invariant features such as SIFT [55].

Appearance-based methods [51] [52] [64] consist of two steps. In the first step, a set of images is used to construct the model. The appearance of the object in the training set varies due to various imaging conditions such as changes in orientation of the object, varying illumination conditions and possibly multiple instances of an object class such as faces. These training images are highly correlated and thus can be compressed efficiently by methods such as Principal

Component Analysis (PCA). In the second step, from a given image, parts are extracted using segmentation (by color, texture, motion) or exhaustive image windows (sliding windows) over the given image. These parts are then compared with the reference images or the learned object model. An example is the recognition of faces using the Eigen faces approach [86] where the variations are learned from the training face images and then a given image is projected on the learned model to recognize the face. Another method that comes under the appearance-based object recognition is the global histogram matching such as the color histogram used by [83] [82]. Color histograms of the image regions are compared with the histogram of object model to recognize the object. Schiele and Crowley [78] [79] then generalized the concept of histogram matching. They use responses of various filters to build the histogram instead of pixels colors which they call receptive field histograms. However there are certain limitations of the appearance-based methods. For instance, they rely on the isolation of object from the input images, thus making them sensitive to occlusions and require good segmentation [57]. Apart from that, the global histogram matching in the appearance-based methods that use the color information [83] [82] are also sensitive to changes in illuminations. Due to these limitations, the appearance-based methods are only suitable for applications with limited or controlled variations in the imaging conditions such as in the case of industrial inspection systems [63].

In the geometry- or shape-based methods, the information of the objects is represented explicitly using primitives such as lines and edges. Ideally, the object and the image should be represented using primitives that are in simple relationship with one another. For instance, if the object is represented as a wire-frame then the image might best be represented as intensity edges where each one of the edge is directly matched to one of the wires in the model. Practically, the intensity edges in the image are affected by changes in illumination and discontinuities in the surface thus making the image representation different from the object model. Considerable amount of work is done to come up with primitives that are invariant to changes in pose and illumination [65] [97]. However, on the downside, geometry-based methods rely on the extraction of primitives such as lines and circles from the images. There might be certain detected primitives in the images that are not modelled thus creating ambiguity in the matching of primitives from the image and those from the object model. Apart from the image representation, there is also a need to construct the models for objects manually.

Local features are used to represent the image patches. The gradient information of the underlying image patch are used to describe it in a manner which is robust to changes in pose, orientation and illumination. These features are extracted from the training images and then used to classify a given image. Following are the main advantages using the local features [57].

- Training images with minimum or no user intervention can be used for learning the object model. The user only has to provide exemplar images that contain the object.

- Explicit modeling of the geometric primitives is not needed to either represent the object model or the image as the local features are calculated on the image patches which are representatives of the object appearance.

- Local features-based methods can handle occlusion and also do not need the segment the object of interest from the complex background.

If the contents of any two images are similar, they are more likely to have similar local features. Thus establishing the correspondences between the local features is a vital step. Many similarity measures such as the L_2 [55],χ^2 [11] and the Earth Mover's Distance [74] have been used. The correspondence among the local features is also established in the form kernel functions with classifiers such as the support vector machine (SVM) [90]. A specialized combination of local features with SVM that has become the standard paradigm for image classification in the past decade is the Bag-of-visual-words (BoVWs). The BoVWs model has shown state of the art performance on benchmark datasets such as the PASCAL Visual Object Class Challenge [18]. Advantages of the BoVWs model are listed in the following:

1. Based on the design of the underlying local image descriptors, the model is invariant to changes in illumination and affine transformation [28].

2. At the stage of vocabulary construction high dimensional local image descriptors are clustered using unsupervised image clustering such as the k-means. The representative cluster centers are kept while the rest of the descriptors are discarded. Afterward, the high dimensional local image descriptors extracted from any given image are mapped to these cluster centers via a similarity measure such as the Euclidean distance. Finally, the image is represented as a single sparse vector of fixed dimensionality allowing the usage of machine learning algorithms which assume the input space is vectorial- either for supervised classification or unsupervised image clustering.

3. As a common practice, the local image descriptors are extracted with some overlap [87] resulting in implicit geometric dependencies in the BoVWs model.

Details of the BoVWs model are given in the next Section:

2.2 Bag of visual words (BoVWs) model

The work presented in this thesis focuses on object category-based image classification where a given image is named according to the category of the object that it depicts. Instances of a given object category such as a "bike" can look different from one another in spite of having certain common features such as the "wheels". Therefore these common features can be used to learn a generalized model for a given object category. Based on this model a given image can be classified according to the instance of the object category that it contains.

Text analysis has a similar problem when it comes to automatically assign a piece of text to a particular topic such as "sport news" or "English literature". A solution is proposed for such a problem based on the so-called code book learning [32]. This code book is learned in the training phase and it consists of words or phrases that are commonly found in a particular topic. Consequently, each topic is represented as a set of words or "bag-of-words". A given piece of text is then assigned to a particular topic based on the learned model.

Images are made up of image patches and the sets of image patches from different images that depict the instance of a given object category have common image patches. Therefore

a vocabulary of patches can be constructed like the vocabulary of words in text analysis to represent the images. This technique is called the bag-of-visual-words (BoVWs).

Csurka et al. [15] proposed the BoVWs model-based image classification for the first time and since then it has been modified by many researchers to achieve competing results on benchmark datasets such as Caltech [20] and PASCAL VOC [18]. In this model, images are first represented as the histograms of the so-called *visual words*. These histograms are then used with some classifier such as the support vector machine (SVM) [90] to perform image classification.

Local features such as SIFT [55] are used to represent the image patches. Due to this reason, most of the works build their BoVWs model on top of the local features. Therefore the image classification based on the BoVWs model consists of the following five steps:

1. **Local features extraction**: Local features such as SIFT are extracted from the given set of images which is also called features sampling [68]. These features are either sampled densely from the images using a regular grid or from prominent image areas which are first located using interest point detectors such as difference of Gaussian (DoG) [55].

2. **Local features clustering**: Once the local features are sampled, they are clustered to construct the vocabulary of visual words using various clustering strategies such as the k-means.

3. **Local features encoding**: Once the visual vocabulary is constructed, it is used to represent the images. From a given image, local image descriptors are sampled and then assigned a visual word from the vocabulary which is called local features encoding.

4. **Spatial information**: The global position information of the local descriptor is encoded to achieve relative geometry of the pure local words [28].

5. **Classification**: Once the image representation is constructed using the vocabulary, classifiers such as SVM are used to label the image.

The details of each of the steps are given in the following:

2.2.1 Local features extraction

Learning mechanisms are highly dependent upon the quality of the image representation where bad representation results in a bad outcome also known as 'garbage in, garbage out' principle [87]. In a contemporary BoVWs model, images are first represented as local features. Extraction or sampling of the local features from the images is an important step in the BoVWs model [68] [87]. The sampling strategies for local features can broadly divided into the following four groups.

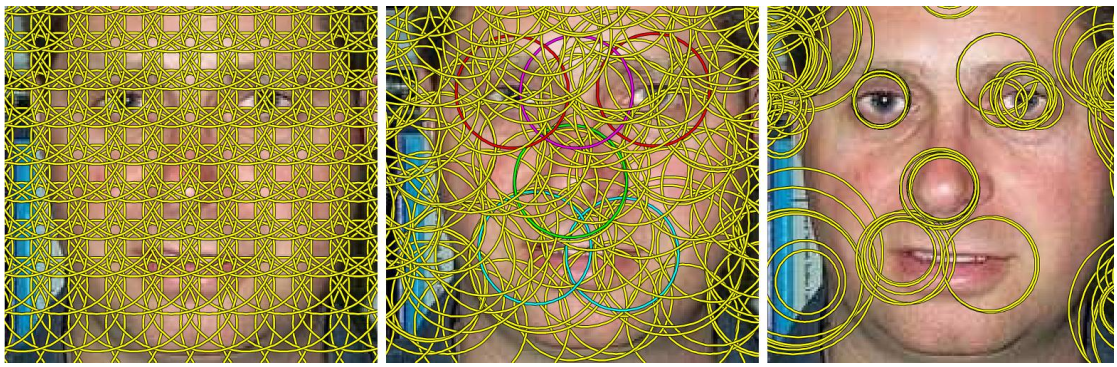
- **Interest points-based sampling:** Regions of images with high information content that can be localized are called *interesting regions* [87]. These regions are detected in images using the interest point detectors [60]. In order to describe the detected interesting regions some methods use SIFT [15] [26] [81] while others use raw patch information [23] [2] [22]. The interest point detectors are invariant to changes in view point and

illumination thus resulting in a high repeatability rate. An example of interest point detectors on face image is shown in Figure 2.1c where it can be observed that they localize well on interesting image regions such as the eyes, nose and mouth. Although interest points perform well on matching applications yet their performance is not satisfactory at the task of image classification [34] [98]. The main reason behind this is that for a given set of parameters, the number of interest points extracted from an image varies substantially based on the content of the image. Sometimes for low contrast images, interest point detectors fail to detect even a single interest point thus making the image representation impractical.

- **Dense sampling:** This kind of sampling strategy provides complete coverage of all the image regions. In this strategy, local image descriptors such as SIFT are extracted on a regular grid with a constant pixel stride [98] [1] [50]. In a given row or column, equally-spaced pixels are selected as feature points and then local descriptors are calculated for each feature. The common practice of overlap of the adjacent regions for descriptor calculation is 50% or less [87]. An example of dense sampling for face image is shown in Figure 2.1a where it can be observed that it completely captures the geometry of face structure. The main advantage of the dense sampling strategy is that even the low contrast regions are given equal importance. Even if they contain little information, they may help to classify the image. Another advantage of such sampling is that it captures the spatial relations among the local features on a regular grid. This an important pre-requisite when modeling the spatial relationship of local features [50] [40] [111]. On the downside, dense sampling is computationally expensive [68]. Secondly, the overlap of the regions for the computation of local descriptor is not enough to guarantee similar descriptors in case of structured scenes [87].

- **Random sampling:** Nowak et al. [68] showed that randomly sampling from regularly extracted multi-scale local features performs better than the interest-points based sampling. However their proposed method does not show any advantage over the dense sampling.

- **Dense interest points:** This hybrid approach is proposed by [87] where it combines the



(a) Dense Sampling (b) Dense Interest Points (c) Interest Points

Figure 2.1: Various strategies for local features sampling (courtesy of Tinne Tuytelaars [87])

best of both the worlds i.e. interest points and dense sampling. Image patches are densely sample on a regular grid and at multiple scales where the amount of the pixel stride on the dense grid is adjusted according to the scale of the patch to minimize the overlap of the adjacent patches. Afterward, each patch is further refined both with respect to position and scale with some measure of interestingness such as the Laplacian. If a true local maximum is found within the patch limits such as an edge, that point is considered as the center of the patch. If no maximum is found over the entire patch area, the center point of the patch is selected. Local descriptor such as SIFT is then calculated for the patch centered on the local maximum. An example of the dense interest points on human face is shown in Figure 2.1b. It can be observed that the dense interest points are densely extracted from the image like the dense sampling. However, unlike dense sampling, it emphasize on and adapt to the interesting regions of the images like the interest points. In the human face, structures such as eyes, nose tip and mouth corners are interesting regions and thus the dense interest points are centered on these regions.

2.2.2 Local features clustering

The local features extracted from a set of images are clustered to construct the visual vocabulary. Clustering is done using various techniques that belong to the following two main groups.

- **Hard clustering**: In this type of clustering, features are assigned to a single cluster based on a similarity measure such as the Euclidean distance. They are also called the partitioning clusters. An example of such type of clustering is the k-means clustering which is frequently used in the BoVWs model [72] [81]. In such type of clustering, the feature space is quantized into informative regions and each region is represented by the mean of all the points located in that region. Once the algorithm produces stable regions, the features in each region are discarded while their means are kept. These means are called the *visual words* and the collection of visual words is called the *visual vocabulary*. The time complexity of k-means is $O(kN)$, where N is the total number of data points or descriptors in the feature space and k is the size of vocabulary. For small vocabulary sizes this feasible but for larger vocabularies this results in increased time. Several methods are proposed to reduce the time complexity and achieve the quantization with complexities of $O(N)$ [88] and $O(N \log k)$ [66].
- **Soft clustering**: In this type of clustering, features are assigned to various clusters in a weighted manner. A given descriptor in the feature space can be assigned to more than one cluster based on the probability distribution. Such a vocabulary is called a *soft vocabulary* and is built using the Gaussian Mixture Model (GMM) [39]. The parameters of the GMM are learned using expectation maximization (EM) [59] over a training set of descriptors that are sampled from the training images.

2.2.3 Local features encoding

Local features extracted from the images are encoded or assigned visual words from the visual vocabulary using various techniques depending on the type of vocabulary. Histogram encoding

is the simplest of all where the features of a given image are assigned visual words based on some similarity measure such as the Euclidean distance [15] [50] [81]. The image is then represented as a histogram which counts the instances of the visual words. The size of this histogram is equal to the size of the visual vocabulary. Recently proposed advanced techniques perform such assignment in a more sophisticated manner either by representing features as a combination of visual words e.g. soft quantization [72] and local linear encoding [94] or by modeling the difference between the local features and the visual words such as the super-vector encoding [114] and the Fisher vector encoding [71]. In the following, a brief summary of soft quantization [72] and Fisher encoding [71] is given to show the differences between the two types of encoding methods.

- **Soft quantization**: The local descriptor such as SIFT calculated for identical image patches are identical. However, unwanted factors such as changes in illumination, image rotation and noise etc, result in abrupt changes in the descriptors calculated for the same physical image patch. Therefore, during the process of hard clustering like k-means, they are assigned to two different clusters. The features that fall within the same cluster are considered identical and different from those assigned to other clusters. Consequently, features that are close to each other in the feature space but assigned to different clusters are considered different. The purpose of the soft assignment is to assign multiple visual words to a given features instead of assigning the feature to a single visual word. Such assignment is based on a weighted technique that is based on the distance of cluster center and the given feature in the feature space.

- **Fisher encoding**: In this encoding scheme, soft vocabulary is used which is built using the Gaussian Mixture Model. Each image descriptor is then represented as vector whose entries are the first and second order differences between the descriptor and the center of a GMM.

2.2.4 Spatial information of the visual words

A basic problem in the BoVWs image representation is lack of spatial information of the quantized local features [69] [50] [6]. To highlight this problem and summarize various solutions proposed for it, the simplest case of histogram encoding is considered here. However, the spatial information incorporation methods can also be performed on other encoding methods [46]. The histogram counts the number of visual words without considering their spatial positions in the 2D image space. On one hand it is an advantage because this property of BoVWs makes them flexible to viewpoint and pose changes [28]. However, the BoVWs model enriched with the spatial information outperform the contemporary BoVWs model in cases of image classification based on scene and object recognition [50] [111] [70]. The spatial information incorporation methods capture the underlying geometry present in objects and structured scenes and thus perform better than the BoVWs model. Based on the underlying principles, the spatial information incorporation methods can broadly be divided into the following two groups.

- **Splitting the image space**: The methods in this group use image space splitting to enrich the BoVWs image representation with spatial information. After encoding the local

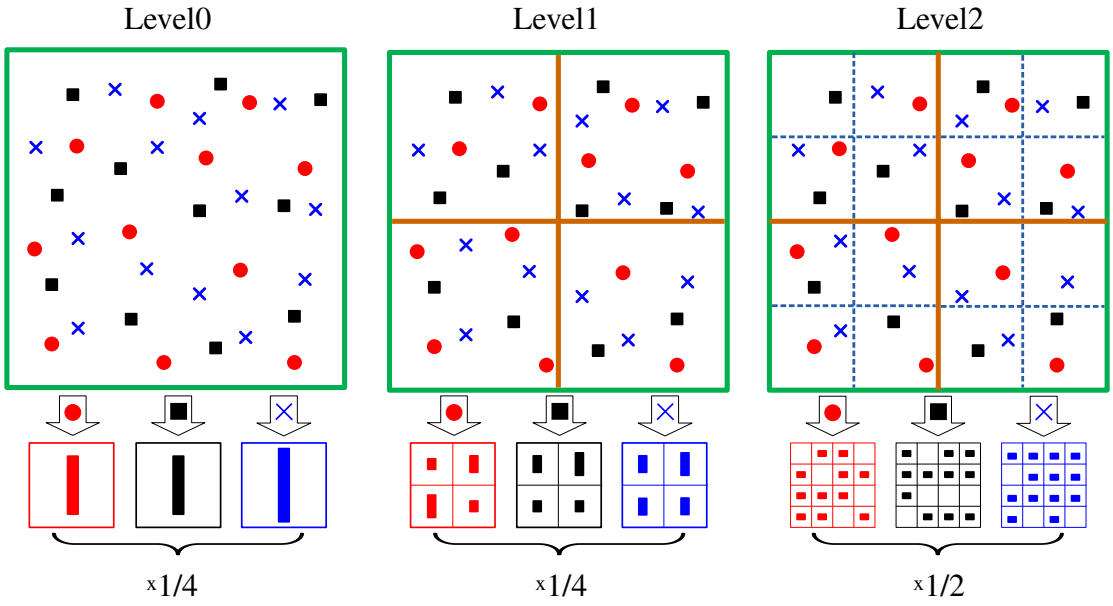


Figure 2.2: Spatial pyramid matching (SPM) [50]. Image space is split into rectangular tilings of various granularities. Level0 has not tiling, Level1 has four rectangular tilings which Level2 has eight rectangular tilings. Level2 is given more weight than Level0 and Level1. The weighted histograms of visual words from each level are concatenated to represent the image.

features using the visual vocabulary, the spatial relationships of visual words are modeled by splitting the image space into subspaces of various shapes and scales. From each subspace, the statistics of visual words are calculated and used to represent the image. In this way, the spatial information of encoded local features is added to the BoVWs image representation.

A notable work from this group is the Spatial Pyramid Matching (SPM) [50] which is based on the spatial pyramid match kernel [27]. SPM divides the image space into hierarchical rectangular tilings. Figure 2.2 shows three levels of the pyramid where Level0 has not tiling at all, Level1 has 4 rectangular tilings and Level2 has 8 rectangular tilings. Weighted statistics of visual words from tilings at each level are then aggregated to construct the final image representation. Inspired by the shape matching [11], log-polar tiling is used by Zhang and Mayo [111]. Single, multiple and multi-scale log-polar tilings are imposed on image space as shown in Figure 2.3. From each sector of the log-polar tiling, statistics of visual words are extracted. They report improved performance over SPM on three benchmark datasets. Another recent method also splits the image space into rectangular tilings which is called *word spatial arrangements* [70]. Making the position of a given visual word as the origin, the image space is divided into four tilings. From each tiling, the information about the visual words is collected. This process is repeated for

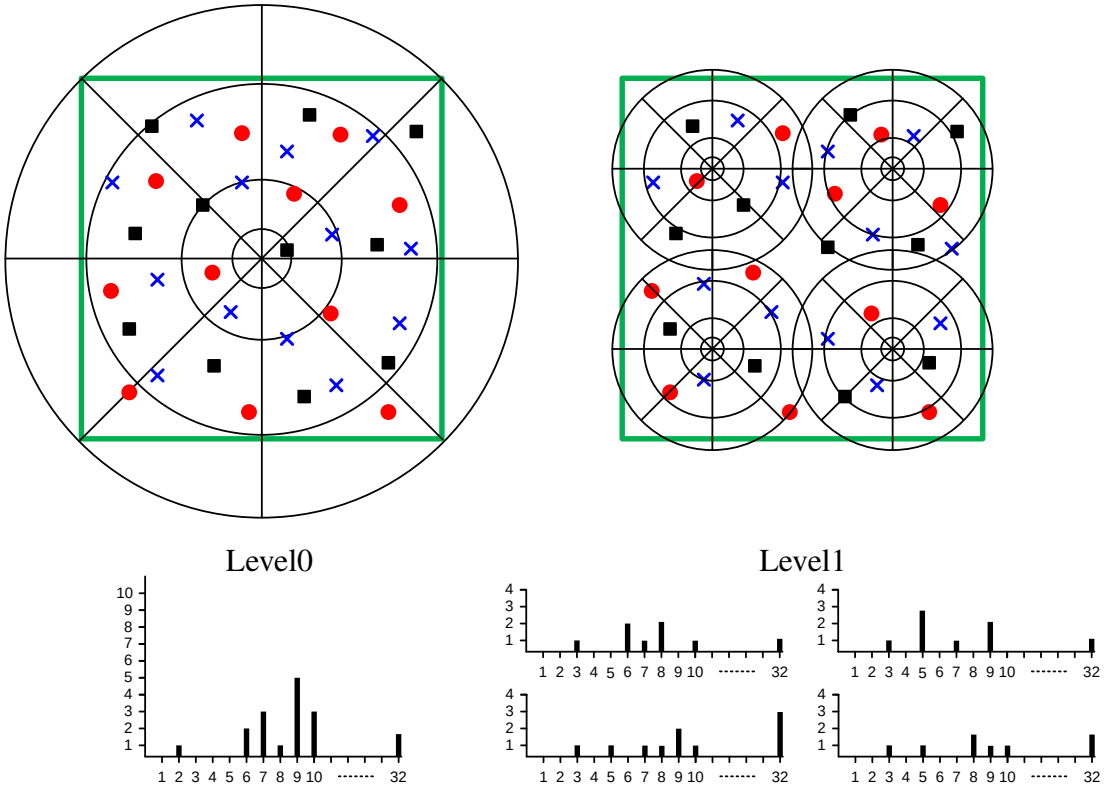


Figure 2.3: Splitting image space into log-polar tiling [111]. Image space is divided into sectors of log-polar tilings where Level0 has one tiling and Level1 has four log-polar tilings. From each sector of the tilings, statistics of visual words are extracted to represent the image.

all the visual words of a given image. Finally, the information of all the four tilings is aggregated to represent the image. These methods are invariant to scale changes as the image subspaces are of various scales. However they are not robust to image rotations and translation. For object category based-image classification, being invariant to changes in rotations and translation is important as objects can undergo various rotations and they can be found at various locations in the image. A recent method for ancient coins classification [4] accounts for these problems by proposing a circular tiling imposed over an automatically segmented and scale adjusted coin image. Thus they achieve robustness to changes in image rotations, scale and translation all at the same time. However their method is only suitable for ancient coins because they are imaged over homogeneous background and thus are automatically segmentable.

- **Visual words spatial co-occurrence**: In this group of methods, spatial information is incorporated to the BoVWs model using the relative positions of the visual words which is also named the visual words co-occurrence [53, 54, 100]. These methods are mostly inspired by the so-called correlograms of quantized colors for indexing and classifying

images. Instead of color, Savarese et al. [77] use correlogram of visual words to model the spatial relationships of the quantized local descriptors. Correlograms are three dimensional structures that count the number of times two visual words occur together at a particular distance from each other. The elements of the correlogram related to a particular pair of visual words are quantized to form correlations. Consequently, the histograms of correlations are constructed to represent the image for image classification. Liu et al. [54] use circular regions of varying radii to calculate the co-occurrence-based spatial histograms. Spatial pyramid co-occurrence is proposed by Yang and Newsam [102] to combine the best of both the worlds i.e. SPM and co-occurrence. The method of Khan et al. [42] is more focused on modeling the geometric co-occurrence using the angles made by identical visual words in the 2D image space. Their approach achieved competitive results against SPM [50] on four benchmark datasets. They represent images using normalized angles histograms which are built from angles between two pairwise identical visual words. However modeling the co-occurrences of visual words in the case of large vocabularies proves to be computationally expensive [113]. The methods based on geometric relationships of visual words such as the one proposed by Khan et al. [42] are also unable to achieve invariance to image rotations.

2.2.5 Classification

For the BoVWs model, most of the works [50] [111] show state-of-the-art results on benchmark datasets [18] [20] by using the support vector machine (SVM) [90] for classification. Image representations based on any of the local features encoding method can be used with a linear kernel [39]. Although linear kernels are efficient in training [33] yet their classification accuracies are inferior to those of the non-linear kernels [112]. A group of kernels that are almost as efficient as linear ones but more accurate are the additive homogeneous kernels [92]. For instance, the Helinger kernel is pre-computed on the histogram representation of the image and then fed to the linear SVM that results in better classification than the linear kernel [92] but almost as efficient as the linear one.

2.3 Image-based Analysis of Ancient Coins

The proposed image representations are practically evaluated on the problem of image-based classification of ancient coins. Since Numismatics deals with the study of coins and currency, in this section, an overview of the knowledge from Numismatics that is applicable to the image-based classification of ancient coins is presented along with the challenges in the image-based ancient coins classification. In addition, an overview of the techniques and methods that deal with image-based ancient coins classification is also given.

2.3.1 Introduction to ancient Numismatics (The Roman Republican coinage)

In the ancient times coins had become the central embodiment of money, starting from around the 7^{th} century BC in Greece and spreading over other civilizations like the Roman Empire, Byzantium, India or China [29]. Expert craftsmen known as *engravers* had been minting the

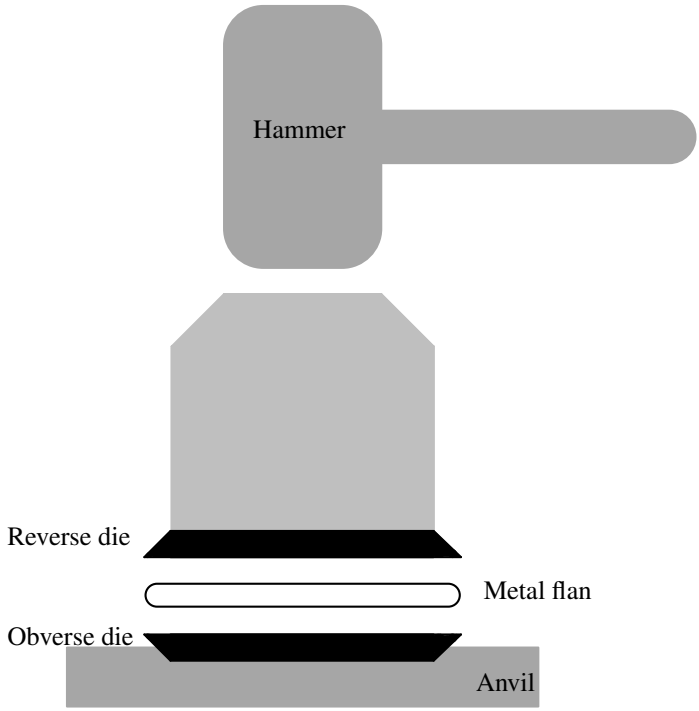
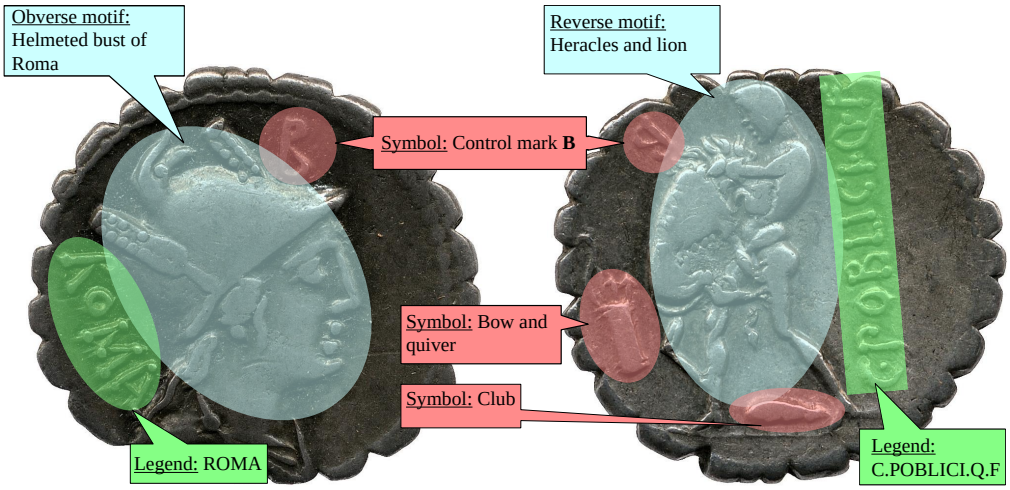


Figure 2.4: Process of minting the coin

ancient coins until the late 18^{th} century AD. They used to design the dies and then utilize them for striking the coins with the help of a hammer and anvil as shown in Figure 2.4. They used to put a hot metal flan on the die of the obverse side clamped by the anvil and strike the die of the reverse side on to the metal flan by a hammer. While being everyday objects in the past, nowadays ancient coins are also considered as pieces of art which reflect the individualism of the engravers who manually cut the dies used for minting the coins [14]. Roman coins, for instance, often depict portraits of gods, influential persons as well as historical events, in a similar manner as sculptures or paintings from this era do [14]. The Roman Republican coins are properly indexed in the reference book of Crawford [14] where each coin class is given a reference number called the *type* of the coin. Under each type, elements of the coin type depicted on the obverse and reverse sides are described. An example of the basic elements of Roman Republican coin along with description is depicted Figure 2.5. On most of the coins, the obverse coin sides show a portrait of the goddess Roma as central *motif* or, at a later stage, that of a historic person, whereas the reverses show some kind of significant objects or scenes. The main motifs are mostly combined with a *legend*. In addition to that, there are minor images, such as a tiny club, or numerals, which are referred to as *symbols* in classical numismatics. They split up otherwise large and uniform coin issues of a specific year.

In contrast to existing works, where the image-based classification is grounded either on obverse portraits [9, 43, 44] or legends [10, 38], reverse-side motifs are used. Accordingly, the presented method closely mirrors the human approach, as the reverse images usually hold a



Cra 380/1

Obv. ROMA Bust of Roma r., draped, wearing helmet with plume on each side; above, control mark B.

Rev. C. POBLICI. Q. F. Hercules strangling Nemean lion; at his feet, club; on l., bow and quiver; above, control-mark B.

Figure 2.5: The anatomy of a Roman Republican coin as given in the description of a reference book [14]

critical amount of information that is required for successful numismatic classification. This is also true for other periods than the Roman Republican time. Naturally the central (reverse) motif is most prominent in all descriptions and, compared to legends or minor symbols, larger and thus more resistant against the inevitable degree of wear that is generally found on ancient coins. If manual numismatic classification is not possible for exactly that reason, it is common agreement to give a description of the visible parts (usually the central motif) and, if possible, the resulting probable reference numbers from the reference books [14].

2.3.2 Challenges in the Image-based Analysis of Ancient Coins

Unlike the modern day coins, image analysis of the ancient coins has a number of challenges which are briefly outlined as follows:

1. **Incomplete data:** Ancient coins are worn to a certain degree because of their age. In most of the cases, ancient coins are discovered in ruins and fields where they are buried in soil. The corrosion caused by such preserving condition induces variations in the images of the same reverse motif class. A few examples are shown in Figure 2.6 where images of the same reverse motif vary from one another due to the incomplete data caused by corrosion.

2. **Variations in symbols (club and numerals):** The main object of the reverse motif such as *triga* as shown in Figure 2.6 or *Griffin* as shown in Figure 2.7 are used for image-based classification of ancient coins. In addition to these objects, the reverse motifs also

Figure 2.6: Variations in the images of reverse motif due to incomplete data

depict small symbols such as club along with the main objects. In a number of cases, these symbols vary significantly causing variations among the images of the same reverse motif. This is shown in Figure 2.7 where various clubs are depicted beneath the belly of the main object i.e. *Griffin*. These clubs capture a significant part of the reverse motif due to which images of the same reverse motif significantly vary from one another.

3. **Non-rigid deformations:** The dies used by the engravers for minting the coins were not struck with the same force on every coins. The flan was not precisely centered between the dies. The flans were also hand-crafted thus differing in their size and shape. Furthermore, the coins dies themselves wear off by the time as it is estimated that the number of coins possibly struck using the same pair of dies can range between 1,000 to 40,000 [17]. These factors cause non-rigid deformations that lead to the variations among motifs. An example is shown in Figure 2.8 where the non-rigid deformations can be observed. The main object i.e. *dolphin* is not minted precisely at the same place of flan resulting in the variation of the distance between the object and the coin border. Furthermore, the heads of the *dolphins* are not of the same dimension.

4. **Imaging conditions:** The images in the coin image dataset are collected from three different sources which are the Vienna Museum of Fine Arts, The British Museum London and the online auction website acsearch.info. Sample coin images from all the three sources are shown in Figure 2.9 where the differences among the images can be observed. The coin images acquired from the Vienna Museum of Fine Arts differ from those of the

Figure 2.7: Variations in the images of reverse motif due to variations in symbols

British Museum and acsearch due to the background and illumination conditions. The images from the British museum vary from one another due to strong illumination changes. Similarly, the dirt on the coins in the images obtained from acsearch also cause variations among coin images of the same class. Lastly, a significant variation can be observed among the coin images from the Vienna Museum due to rotations.

2.3.3 Prior works and contributions on image-based analysis of ancient coins

In contrast to methods dedicated to present-day coins [67, 75], image-based classification of ancient coins has become a research interest more lately, owing to the higher complexity of the problem due to the challenging conditions of ancient coins. Ancient coins do not have a rigid shape and are worn to a certain degree because of their age. Consequently, it was experimentally shown by Zaharieva et al. [103] that the success of classification methods for present-day coins cannot be transferred to the domain of ancient coins. Various works and contributions that deal with image-based analysis of ancient coins are summarized based on their common methods.

- **Local features correspondence:**The first method exclusively dedicated to ancient coins was proposed by Kampel and Zaharieva [35]. The method compares two coin images by establishing correspondences between them and taking their total number as similarity value. Given such a similarity function, classification is achieved in an exemplar-based manner where a query image is compared to class exemplar images in the dataset. Dense

Figure 2.8: Variations in the images of reverse motif due to non-rigid deformations

correspondence search is used by [107] to derive visual similarity among the coins. This visual similarity is then used in a coarse-to-fine search to retrieve the most similar image from the database. Accuracy rate of 82.8% is reported on a test set that contains 60 classes of Roman Republican coins. This idea was later extended to more sophisticated similarity functions that take also geometric constraints into account [110]. The main motivation behind the use of local features correspondence is the lack or scarcity of the coin images for training a machine learning algorithm such as the SVM. However, these training images can synthetically be generated according to the descriptions of the reference books [14].

- **Ancient coins classification based on legend recognition:** Alternatively or in addition to comparing the local features, one can leverage a-priori known semantic background information for classification. Arandjelović [10] and Kavelar et al. [38] proposed to exploit the legend shown on ancient coins. The background information here is a lexicon of known legend words linked with coin classes, which are detected by means of graphical models that integrate likelihoods of letter appearances to words. However, the two methods are designed for two different Roman coin periods: the method of Kavelar et al. [38] is designed for Republican coins and thus has to consider more degrees of freedom in the word search than the method of Arandjelović [10] for Imperial coins. On Imperial coins the legend is arranged along the coin border and thus the image can be normalized for orientation by means of a log-polar transformation. The main disadvantage of legend recognition is that the alphabets and numerals minted on the coins are most vulnerable to corrosion leading to the absence of legends on most of the coins.

- **Ancient coins classification based on obverse motif recognition:** Apart from the legend, for Roman Imperial coins the person portrayed on the obverse coin side has been used for classification [9, 43, 44], as this is typically the issuer defining the coin class [80]. The method of Arandjelović [9] relies on detected SIFT features [55] which are quantized into a fixed vocabulary of visual words. The spatial configuration of visual words is

Figure 2.9: Sample coin images from the three sources i.e. The Vienna Museum of Fine Arts, The British Museum London and acsearch.info

then encoded by Locally-Biased Directional Histograms (LBDHs) which are again subject to vocabulary creation and the histogram of LBDH words serves as final image feature. However, two recently proposed methods by Kim and Pavlovic [43, 44] showed to outperform this method. These approaches are more adapted to the field of face recognition and achieve recognition rates of $82-86\%$ on a 15-class-problem by using spatially augmented features [43, 58] and the deformable parts model [21, 44].

A semantic information other than the legend or the person on the obverse is exploited for classification, namely the motifs on the reverse of Roman Republican coins. This is an essential part of coin descriptions and thus highly practical for classification. The method for reverse motif recognition integrates semantic knowledge from expert sources of Numismatics and therefore

can be extended to a comprehensive classification framework. This approach is very close to the human intuition as humans would rely on the motif minted on the reverse side of coins for classification. This is due to the fact that motifs vary significantly in their visual structures and therefore can easily be distinguished from one another. On the other hand, the portraits of historical personalities on the obverse side show a higher similarity [9, 43, 44]. In addition to their advantage over the obverse motifs, the reverse motifs are more resistant to corrosion than the legends and thus presenting themselves as a better choice for ancient coin classification.

2.4 Technical outline

The BoVWs model for image classification is utilized due to the aforementioned reasons. Focus is upon the design and development of techniques to encode the spatial information of the visual words. Dense sampling is used to extract the features from the images which are later clustered using the k-means to construct the visual vocabulary. Histogram encoding is used to encode the local features where the histogram of visual words represents the image. For classification, SVM is used with linear and homogeneous kernels [92]. However, histogram encoding of the image in the BoVWs model does not consider the geometric distribution of the local features in the 2D image space. Such geometric information provides discriminating information in the image-based classification of objects and structured scenes [50] [111]. Three methods are proposed to incorporate the spatial information to the BoVWs model which are based on the two principles i.e. splitting the image space and modeling the geometric relationships of local features. These methods are robust to certain geometric transformations such as rotations, translation and scale changes and thus can be applied to the image-based classification of objects that can be affected by these geometric transformations. The evaluation of the proposed methods is performed on flat objects such as ancient coins as well as textured objects such as butterflies. However the proposed techniques can be applied to the image-based classification of other classes which is prone to the mentioned geometric transformation. Summary of the three techniques used for the incorporation of spatial information to the BoVWs model is given as follows:

1. **Spatial information incorporation by using image subspaces:** The image space is divided into subspaces of various shapes [50] [111]. The subspaces are also called *tilings*. Rectangular, log-polar and circular tiling are used to incorporate the spatial information to the BoVWs model for the problem of reverse motif-based ancient coins classification. It is shown that the circular tiling not only performs on par against the other two tiling schemes but also more robust to image rotations.

2. **Spatial information incorporation by encoding the geometric co-occurrences of identical local features:** The geometric relationships of the identical visual words are encoded using angles that are made by their spatial positions in the 2D image space. These geometric relationships are achieved by modeling the triangular relationships of identical visual words. Since the angles of a triangle are invariant to rotations, translations and scales changes, the spatial information using the triangular relationship is inherently robust to such transformations. The proposed method is applied to the problem of image-based classification of butterflies where it outperformed the BoVWs model.

3. **Spatial information incorporation by encoding the geometric co-occurrences of identical local features in image subspaces:** As a first step, the circular tilings [4] is used to add spatial information to the BoVWs model in a rotation invariant way. Furthermore, the geometric relationships of identical visual words in each circular tiling are modeled using their triangular relationships which not only achieve increased classification rate but also reduce the calculation complexity. The proposed method is evaluated on the ancient coins dataset where it not only outperform the BoVWs model but also proved more robust to image rotations.

CHAPTER 3

Capturing spatial arrangements of local features via image space division

In this chapter a scale-, translation- and rotation-invariant image representation is developed for image-based classification of objects that are imaged with homogeneous background. To this end, the object of interest is first automatically segmented to achieve invariance to changes in scale and object position. Afterward, the configuration of the local features are captured in a manner that is invariant to image rotations.

The problem of image-based classification of ancient coins is taken as a motivating example as images of these coins have homogeneous background. Various objects, animals and historical personalities are depicted on the reverse side of the ancient coins which are called *reverse motifs*. Ancient coins are indexed in reference books [14] with a reference number which is called the 'type' of the ancient coin. Coins of more than one type may depict the same reverse motif due to which the image-based classification of ancient coin based on these motifs is coarse-grained. Figure 3.1 shows the reverse motifs used in experiments.

The development of the proposed image representation is given in Section 3.1. Since the coins are imaged on a homogeneous background, automatic coin image segmentation can be used as a pre-processing step thus making the resulting image representation invariant to translation. A brief overview of the automatic coin image segmentation proposed by Zambanini and Kampel [104] is given as this technique is specifically proposed for the ancient coin images. An overview of the BoVWs model is outline because the proposed image representation is an extension of this model. The results of the experiments performed on the coin image dataset are shown in Section 3.2. These experiments are performed to evaluate various parameters of the proposed image representation such as the density of the regular grid on which local features are extracted from the images. The proposed image representation is an extension of the BoVWs model where spatial information is added using various image tiling schemes such as the rectangular tiling, radial-polar tiling and the circular tiling. Experiments are performed to evaluate the number of partitions in each tiling scheme. The size of visual vocabulary is a basic parameter in

Figure 3.1: Reverse motifs used for coarse-grained ancient coin classification

the BoVWs model and thus is evaluated for the given dataset. Finally Section 3.3 summarizes the chapter by listing its contribution and the future directions.

3.1 Methodology

The details of the proposed image representation is given in this section. It is used for the problem of motif recognition minted on the reverse side of ancient coins. Based on such recognition, the coarse-grained classification of ancient coins is performed. This can be used as a first step towards fine-grained classification [106,109] based on the type of the coin. The pipeline adapted for the development of the proposed image representation is presented in Figure 3.2.

1. **Automatic coin image segmentation:** The ancient coins are imaged on homogeneous background and can thus be automatically segmented as shown in Figure 3.2a. After such segmentation, the coin image is cropped and normalized to achieve invariance to changes in scale and translation.

2. **Image representation using BoVWs model:** Local features densely sampled from the segmented coin image are then mapped to a visual vocabulary resulting in a representation of the image as visual words as shown in Figure 3.2b. The image is then represented as a histogram of visual words.

3. **Spatial extensions to the BoVWs image representation:** Since the BoVWs image representation counts the visual words without considering their location in the image space, adding such information leads to improved performance [50]. Three spatial information adding techniques are evaluated which are circular tiling, rectangular tiling and radial-polar tiling as shown in Figure 3.2c. However, the method for the incorporation of such spatial information should be robust to certain image transformation such as rotations. In the development of the proposed image representation, circular tiling is used to incorporate the spatial information to the basic BoVWs model thus making it rotation-invariant which is also supported by the experimental results.

The details of each step are given in the following:

3.1.1 Automatic coin image segmentation

Coin image segmentation is used as a pre-processing step to achieve invariance to changes in coin position and scale. The method proposed by [104] is used as it is specifically proposed for ancient coins. Their method is based on two assumption which are:

- Coin is the most circular object present in the image.

- Due to the fact that ancient coins are imaged on flat homogeneous backgrounds, the area of the image depicting the coin contains more information contents than the rest of the image.

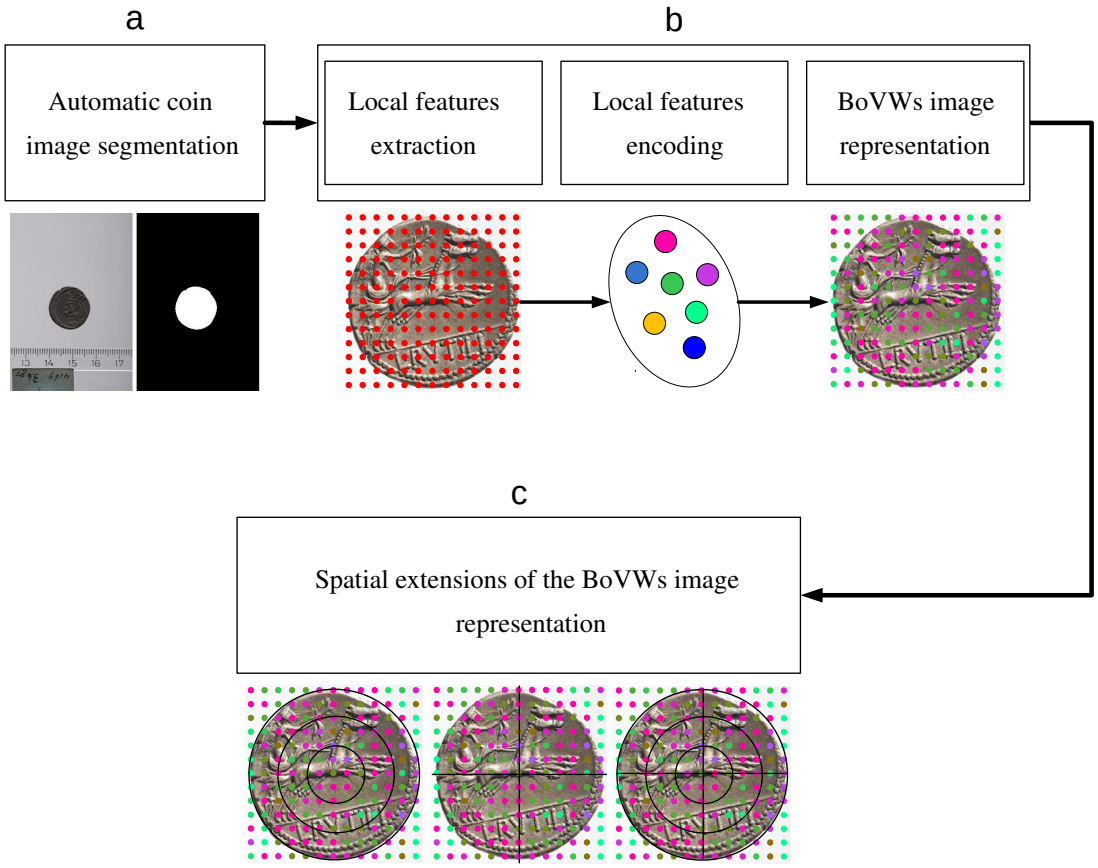


Figure 3.2: Steps for the development of the proposed scale- translation- and rotation-invariant image representation used for the coarse-grained ancient coin classification

Consequently, the automatic coin image segmentation method is a two step process. In the first step, two filters are applied to the image in order to measure the local information content. The first filter measures the local entropy while the second one measures the range of gray values. A circular neighborhood of 3 is used for both filters. The outputs of both filters are summed up to obtain the intensity image which is then normalized to the range 0 and 1. To achieve the final segmentation, global thresholding can be used to remove the holes in the binary mask caused by homogeneous flat regions in the coins. However instead of global thresholding a more sophisticated approach is used in the second step. Seven empirically defined threshold values are applied to the normalized intensity image and a confidence score is calculated for each achieved segmentation. This confidence score is based on the formfactor [76] calculated from the area and perimeter of the binary mask. Formfactor is sensitive to both the elongation and jaggedness of the border. It is highest for a circle equal to 1 and less for any other shape. Since the shape of the coins is supposed be circular or nearly circular, the formfactor is a good choice for confidence measurement. Finally a segmentation is only accepted if the area of the segmented region is less than 90% of the image area. Figure 3.3 depicts ancient coin images and their respective

Figure 3.3: Images of ancients coins along with their segmentation masks

segmentation masks showing the effectiveness of the segmentation method for near circular and degraded coin borders. Figure 3.4 shows the automatic segmentation method for coin images that depict coins of various sizes and at various image locations. To summarize, the coin image segmentation method proposed by Zambanini and Kampel [104] can automatically segment images of the irregular but near circular coins depicted at various scales and positions.

3.1.2 Image representation using the histogram/bag of visual words

Images are represented by performing the following steps:

1. **Dense rotation-invariant local features extraction:** Local features such as SIFT [55] are extracted from each image using a dense grid of constant pixel stride. Since there are

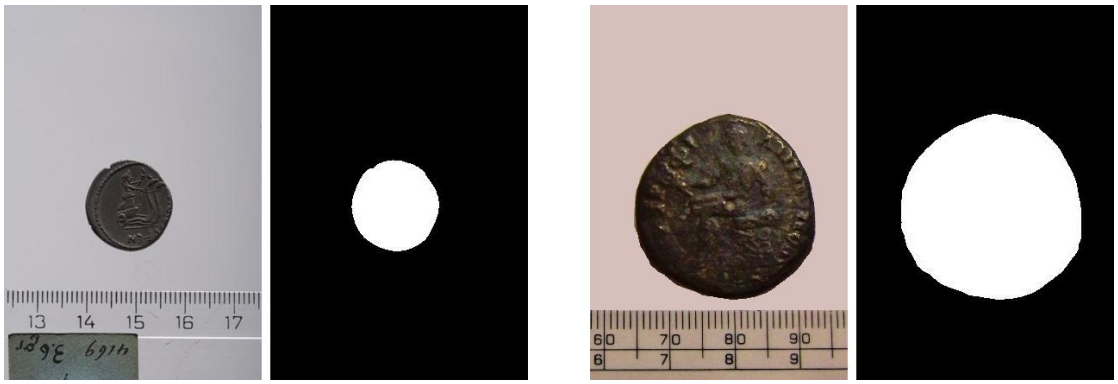


Figure 3.4: Automatically segmented coin images of various scales and at various image positions

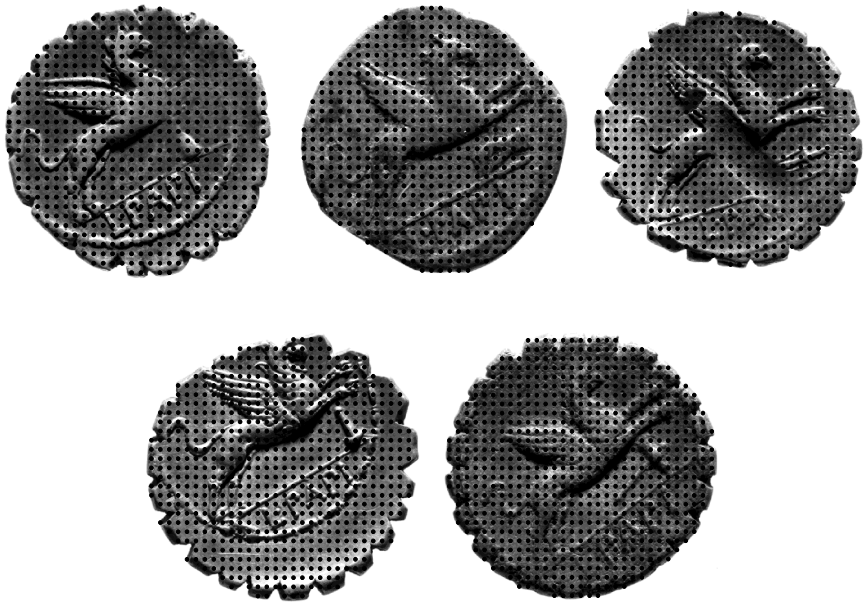


Figure 3.5: Dense rotation-invariant local features extracted from segmented coin images

variations in the reverse motifs due to abrasions, dense sampling is helpful to completely capture the underlying structures of the motifs. This shown in Figure 3.5 where instances of the same reverse motif vary due to aberrations but dense sampling completely covers the underlying structures. The dense sampling accommodates for the missing parts of the object minted on the reverse side. For instance, various parts of the objects are missing as depicted in Figure 3.5 but the dense sampling from the whole set of images accommodates for such missing data. Furthermore, to achieve invariance to rotations locally, dominant orientation of each local feature is also calculated.

2. **Local features quantization:** From a given set of images features are densely extracted using a regular grid. The descriptors are these features in a feature space of 'd' dimensions are shown in Figure 3.6a. The local features densely collected from the images are quantized using a clustering strategy such as the k-means clustering. The number of cluster centers is an important parameter for which empirically defined values are evaluated. Once the k-means clustering becomes stable, the cluster centers become the representatives of all the features in their respective clusters as shown in Figure 3.6b. Since image features are calculated using the underlying image patches, the features that are clustered together in a given cluster represent identical or almost identical image patches. Therefore the cluster centers are kept while the rest of the features are discarded. These cluster centers are the *visual words* their collection is the *visual vocabulary* as shown in Figure 3.6c.

3. **Image representation using the visual vocabulary:** A given image is represented using the visual vocabulary of size M. Local features are collected from the image using a dense grid with a constant pixel stride as shown in Figure 3.7b. Each feature is then mapped to a

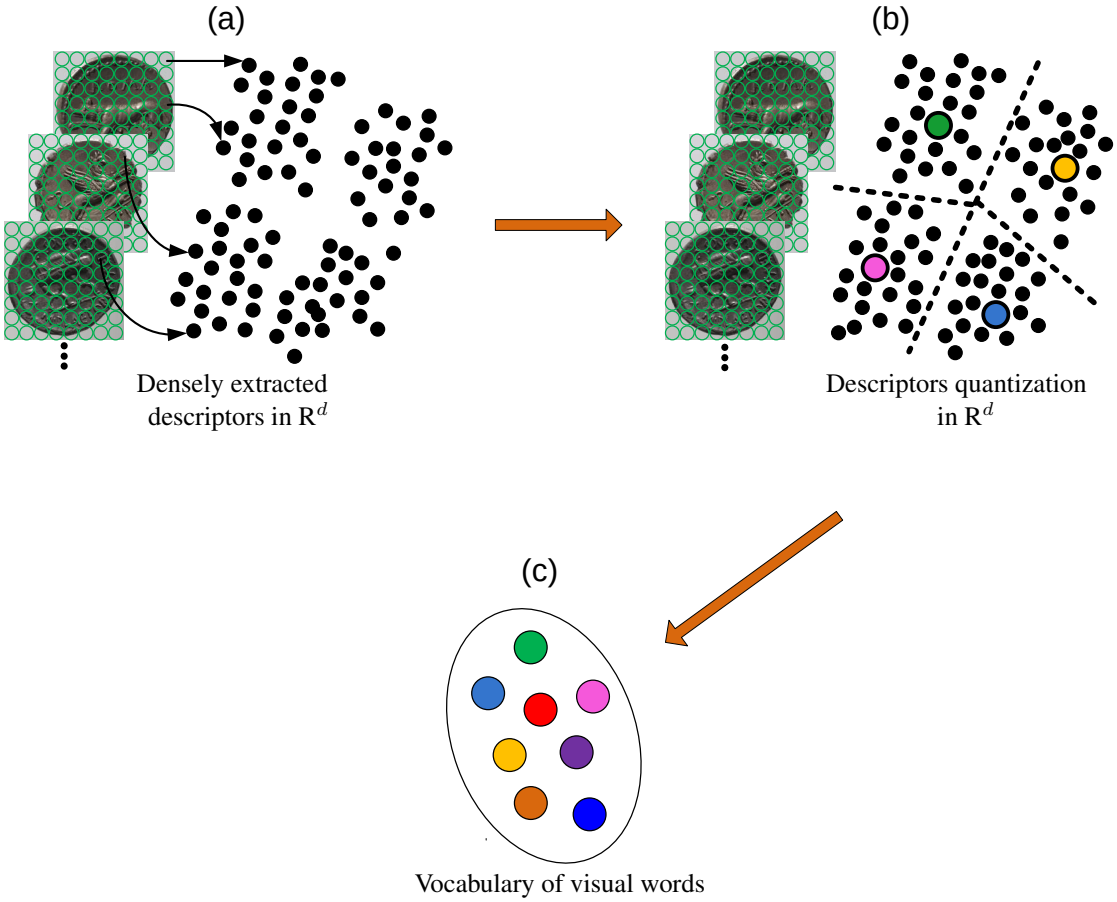


Figure 3.6: Process of local features quantization for vocabulary construction

specific visual word from the vocabulary using a similarity measure such as the Euclidean distance. This is shown in Figure 3.7c and 3.7d where are given descriptor is mapped to the visual vocabulary. Figure 3.7e shows all the the descriptors assigned to the visual words. Once the descriptors are mapped to the visual words, the image is represented as the histogram of visual word as depicted in Figure 3.7f. Such representation is also called the *bag-of-visual words* (**BoVWs**). The size of this histogram is also M, where each bin of this histogram represents the count of the respective visual word in the image.

The given image dataset is divided into two disjoint sets for experiments. The set that is used at the stage of training is called the *training set* and the set used at the testing step is called the *test set*. Densely extracted features from the training set are used for vocabulary construction. Figure 3.8 summarizes the process of the BoVWs representation for both the training and the test sets. The local features extracted from the training set are clustered to form the visual vocabulary. This visual vocabulary is then used to encode the local features extracted from both the training and test sets to represent the image as the collection of visual words which are then summarized in histogram of visual words to give the final image representation.

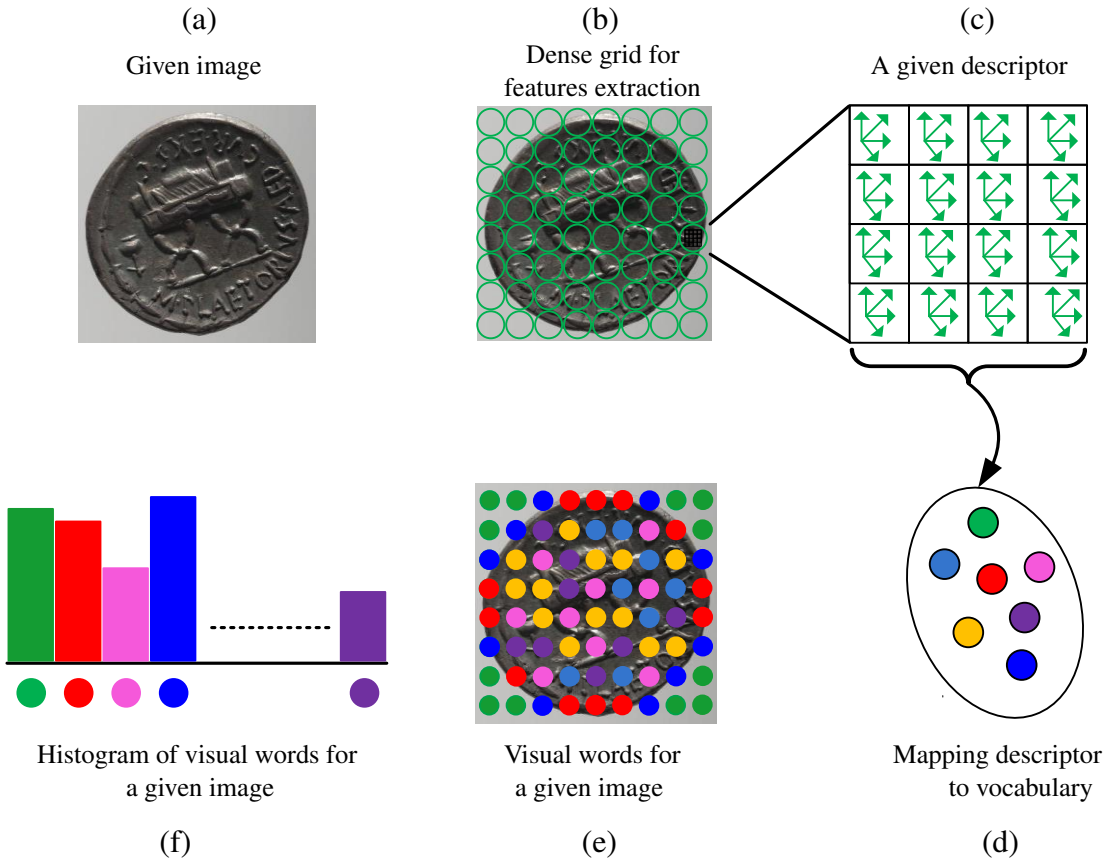


Figure 3.7: Steps of image representation with visual vocabulary

3.1.3 Spatial Extensions to the BoVWs representation

In the BoVWs representation of images, the words are counted and accumulated into the bins without considering their spatial locations in the image space. This is advantageous as this property makes the representation flexible to viewpoint and pose changes. However, it can be disadvantageous in problems where the spatial locations of visual words provide discriminating information. For instance, the classification of ancient coins is based on reverse motif recognition and these motifs have specific geometric structures due to which the spatial information of visual words provides discriminating information for symbol recognition. In this section, the following techniques are used to incorporate spatial information to the BoVWs:

- **Rectangular tiling:** A given image is partitioned into $x \times y$ rectangular tilings. Using a visual vocabulary of size M, a histogram of M visual words is generated for each tiling resulting in $x \times y$ histograms. These histograms are then concatenated into a single feature vector of length $M \times x \times y$. An example is shown in Figure 3.9a where the rectangular tiling of 2×2 is imposed over the image whose the local features are represented as visual

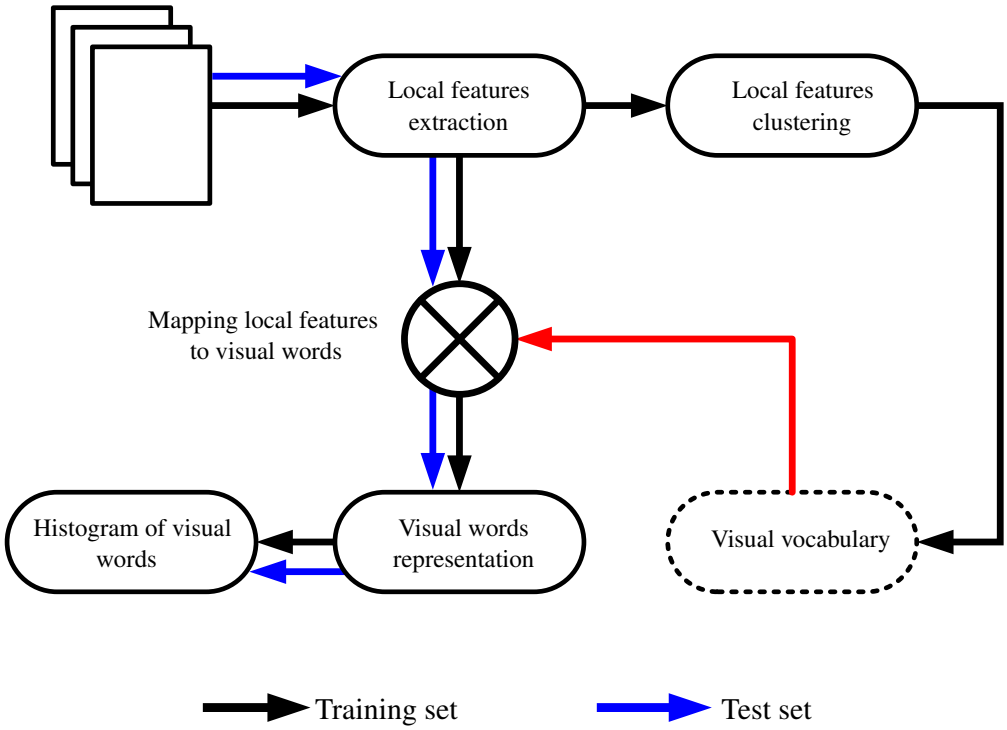


Figure 3.8: Summary of the BoVWs representation

words. The image space is divided into four subspaces of rectangular shapes and equal dimensions. For each of the four tilings, histogram of visual words of size M is calculated. These four histograms are then concatenated to form the final histogram of size $4 \times M$.

- **Radial-polar tiling:** The log-polar binning scheme was used by Belongie et al. [11] to develop their shape matching descriptor. Zhang and Mayo [111] used log-polar binning to incorporate spatial information to BoVWs. Their proposed method outperforms Spatial Pyramid Matching (SPM) [50] on three benchmark datasets. Here an approximation of the log-polar tiling is proposed which is called the *radial-polar* tiling. In log-polar tiling the radii of the concentric circles increase on a log scale while in the radial-polar scheme the concentric circles are linearly equidistant as shown in Figure 3.9b. When radial-polar tiling is applied to a given image, this image is partitioned into sectors of various scales and orientations. Such sectoring helps to capture the distribution of image features both in distance and orientation. A radial-polar spatial scheme of r radial tilings and θ polar tilings is imposed on the image resulting in $r \times \theta$ sectors. For each sector, a histogram of size M is generated and a total of $r \times \theta$ histograms are calculated for the whole image. These histograms are then concatenated in a single feature vector of length $M \times r \times \theta$. Figure 3.9b shows an example of the radial-polar tiling with $r = 3$ and $\theta = 4$. It can be seen that the sectors near the center of the image are smaller than those far from the center. This results in a spatial configuration where the sectors near the center of the image

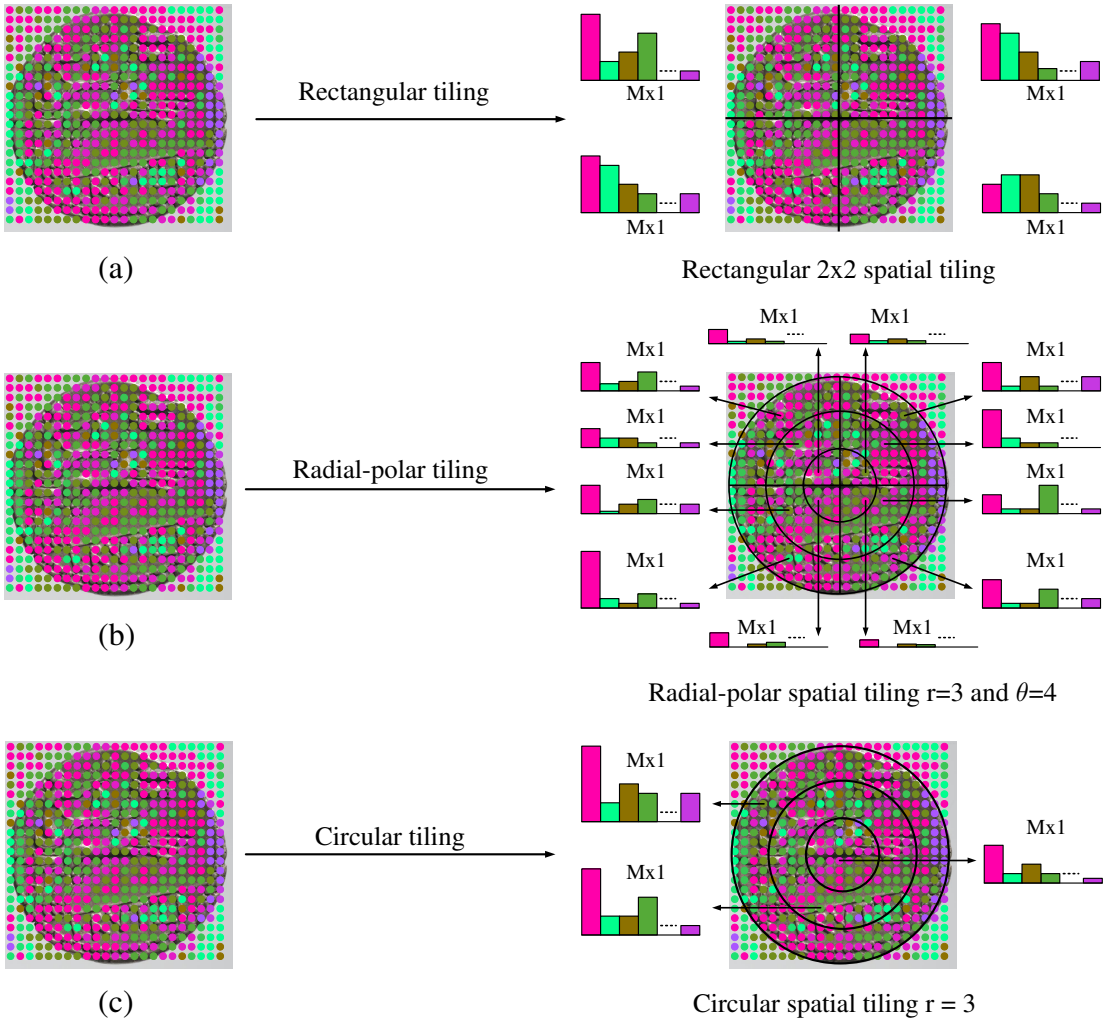


Figure 3.9: Various tiling schemes over the BoVWs image representation

contain fewer visual words that those far from the center.

- **Circular tiling:** Circular tiling scheme is proposed due to the circular shape of the ancient coins. In this scheme as shown in Figure 3.9c, concentric circular tilings are imposed on the BoVWs representation of an image. For a vocabulary size of M, a histogram of M visual words is generated for each circular tiling. If the number of concentric circles is r, then r number of M sized histograms are calculated for the whole image. These histograms are then concatenated in a single feature vector of length $M \times r$. Here it is worth mentioning that the radii of the concentric circles are calculated based on the smaller dimension of a rectangular image. For instance, if the width of the image is smaller than the height, it is utilized to calculate the radii. Doing so keeps the outermost circle well inside the boundaries of the image.

3.2 Experiments and results

Experiments are performed on 1900 images of 21 different reverse motifs of the ancient coins. This dataset is divided into two disjoint training and test sets. Since the image classification is based on the BoVWs model, optimization is performed on the current dataset for its basic parameters. In the proposed model, features are sampled densely from the images on a regular grid with a constant pixel stride. Experiments are performed to optimize for the value of the pixel stride. The size of vocabulary is a basic parameter in the BoVWs model which need to be optimized on the current dataset. In the proposed image representation, the BoVWs model is extended by partitioning the image space into sub-regions or tilings of various shapes. Thus in the second set of experiments, the number of partitions in each of the spatial tilings are evaluated. Finally, the image representation based on different tiling schemes are evaluated for rotation-invariant. Since the image dataset lacks the rotated images for such evaluations, synthetically rotated images are generated and all the tiling schemes are evaluated on these images for rotation-invariance.

The one-vs-all approach of the support vector machine (SVM) is used for classification. The package used is LIBSVM [13]. Helinger kernel [92] is used where the feature vectors are precomputed and then fed to the linear SVM thus reducing the time for training. To obtain the best value of the regularization-loss trade off parameter C, n-fold cross validation is performed on the training set.

3.2.1 Experiments and results for the BoVWs image representation and its spatial extensions

In the experiments, a number of parameters of the BoVWs image representation are evaluated. The first parameter is the pixel stride which defines the granularity of the dense grid for features extraction. The second parameter is the number of features collected from each image to construct the vocabulary. The third parameter is the size of the visual vocabulary. For these experiments, rectangular tiling scheme has $2 \times 2 = 4$ tilings, circular tiling scheme consists of 3 concentric tilings and radial-polar tiling has 4 angular and 3 radial bins. The overall result of these three parameters for each spatial information incorporation scheme is shown in Figure 3.10. For each spatial tiling, three matrices are shown where each matrix represents the results for features densely extracted at a given pixel stride. In each matrix, the x-axis represents the size of vocabulary and the y-axis represents the percentage of features collected per image to construct the visual vocabulary. For instance, in case of "No Tiling", the first matrix from the left shows the results for experiments where local features are densely extracted on a regular grid of pixel stride 10. The first value in this matrix is 49.97% which is the classification accuracy of the BoVWs model with a vocabulary of size 10 constructed from 2% features collected per image. The following results of the three parameters i.e. granularity of the dense grid, size of vocabulary and the number of features are based on these matrices:

1. **Granularity of the dense grid:** Features are collected from the images on a dense grid. In this approach every n^{th} pixel in a given row or column is considered as a feature and descriptor is calculated for that pixel. The pixel stride in dense grid approach is an

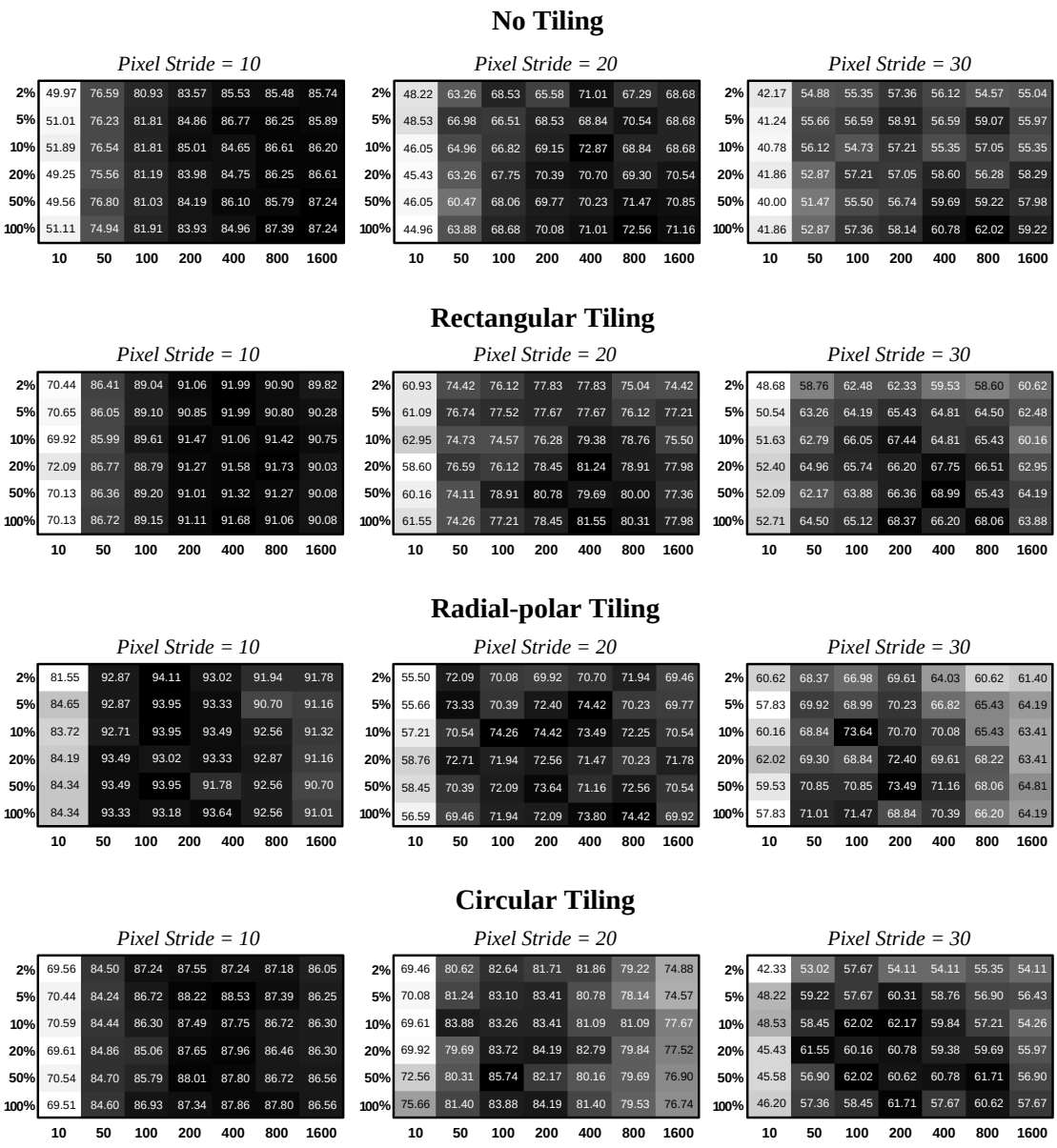


Figure 3.10: Experimental results of various tilings

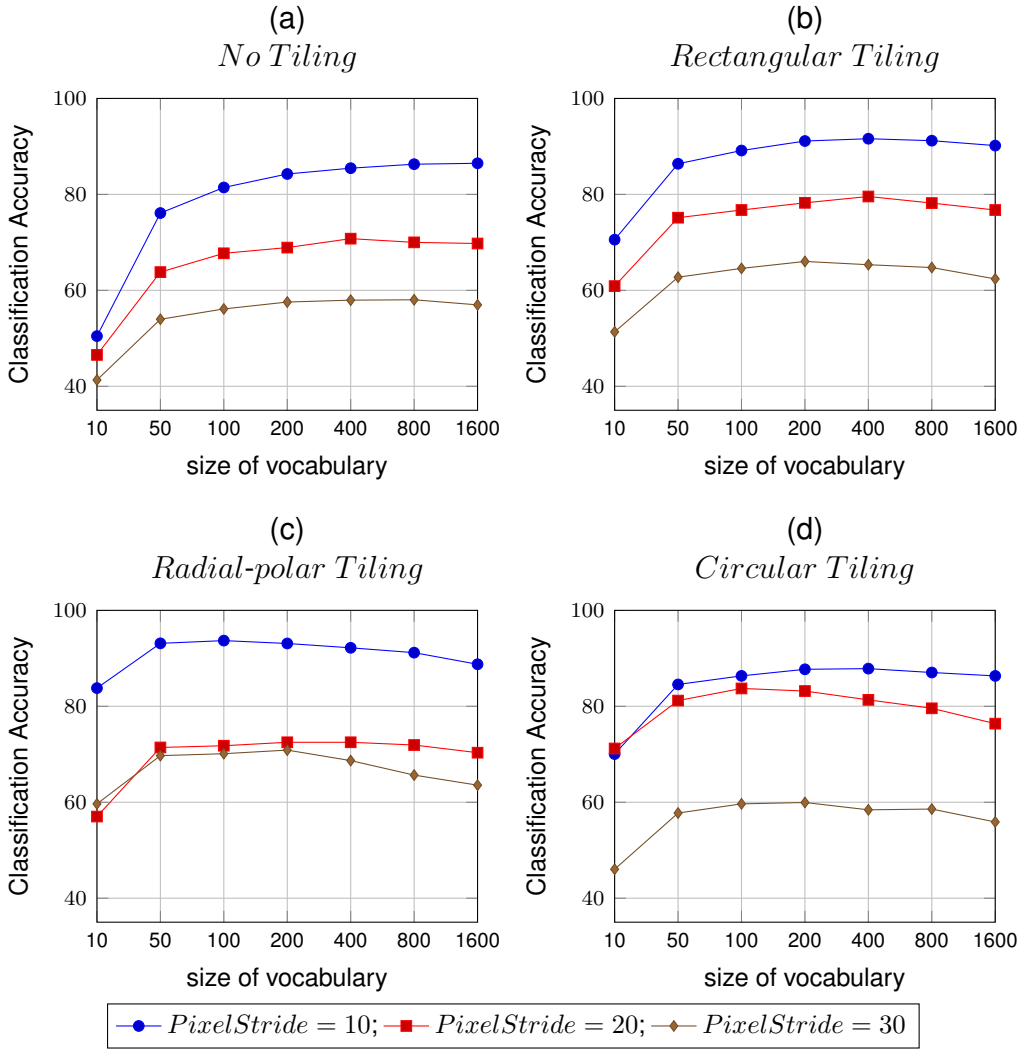


Figure 3.11: The effect of the granularity of the dense grid on the classification accuracy for the BoVWs image representation and its spatial extensions

important parameter that needs to be evaluated. Therefore the values for the pixel stride are empirically defined as $\{10,\ 20,\ 30\}$. For instance, if the pixel stride is 10, then the descriptor is calculated for each 10^{th} pixel in a given row or column. To show the overall performance of each pixel stride, in Figure 3.10, the column means of each matrix is calculated in each spatial setting. The column mean shows the mean performance of a given vocabulary size over all the percentage values in each spatial setting and for each stride value. These mean values are shown in Figure 3.11a-d in case of BoVWs model and the three spatial tiling schemes. In each case, it can be observed that the performance of the dense grid with pixel stride of 10 is better than those with pixel strides of 20 and

30. It means that the feature extraction process performed at a finer grid enhances the performance of the BoVWs model and its spatial extensions.

2. **Number of features used for feature quantization:** The performance of the image classification based on the BoVWs representation increases with the increase in the number of features used for the construction of visual vocabulary [68]. However, the increase in performance becomes negligible after a certain limit. Thus the number of features used for vocabulary construction is evaluated on empirically defined values which are $\{2\%, 5\%, 10\%, 20\%, 50\%, 100\%\}$. For instance, in case of 2% features, from a given image, features are extracted on a dense grid and 2% of these features are randomly selected for the features quantization step to construct the visual vocabulary. In Figure 3.10, the row means of each matrix is calculated in each spatial setting to observe the overall performance of percentage value. The row mean shows the mean performance of a given percentage value over all the vocabulary sizes in each spatial setting and for each stride value. These mean values are shown in Figure 3.12a-d from which it can be observed that the pixels stride with the value of 10 performs better than 20 and 30. However, all the values for the percentage of features collected per image to construct the visual vocabulary perform equal. This is due to the flat nature of the ancient coins where homogeneous patches or patches with relatively low gradient are found. Furthermore, the surfaces of the objects minted on the ancient coins are less textured and their contours are more pronounced. Since the features for vocabulary construction are randomly selected from any given coin image and not from particular areas such as high entropy image regions, the value of 2% perform on par with 100%.

3. **Number of cluster centers:** For the construction of visual vocabulary, features are extracted at dense grid from a given set of images. A quantization scheme is then applied to the descriptors of these features in the descriptor space using a clustering strategy such as the k-means clustering. The number of clusters which are supplied by the user as an initial step in the k-means clustering represents the number of visual words or the size of vocabulary. Therefore the size of vocabulary is a major parameter which needs to be evaluated [15, 68]. The sizes of visual vocabulary that are evaluated here are empirically defined as $\{10, 50, 100, 200, 400, 800, 1600\}$. To show the overall performance of the vocabulary size, the column mean of each matrix in Figure 3.10 is taken. The results are shown in Figure 3.13 where the vocabulary size of 200 performs best in almost all cases of spatial representations i.e. rectangular, circular and radial-polar and the simple BoVWs image representation.

To summarize, the radial-polar tiling dominates the rest of the two schemes and the simple BoVWs image representation. The spatial schemes outperform the visual words representation with no spatial information as shown in Figure 3.13. The performance also decreases as the regular grid for features extraction becomes coarser as shown in Figure 3.11 where the dense grid with pixel stride of 10 outperformed the grids with pixel strides of 20 and 30. The optimal vocabulary size for all the settings is 200 and the percentage of features collected per image to construct the visual vocabulary is 2%.

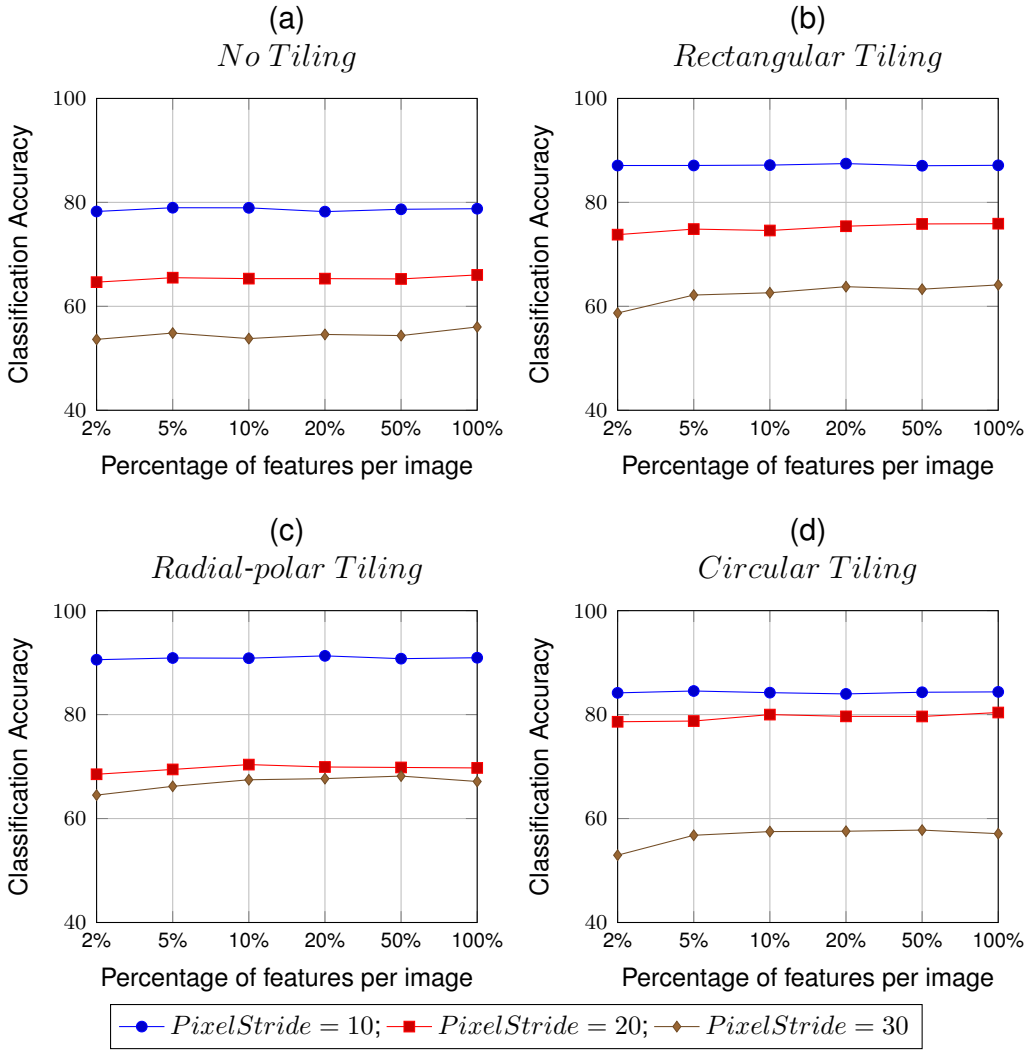


Figure 3.12: The features densely extracted per image and the percentage of features then collected from these features are used for vocabulary construction. The effect of various percentage values on the classification accuracy of the BoVWs image representation and its spatial extensions are shown

3.2.2 Experiments for the number of tilings in the spatial extensions to the BoVWs image representation

Experiments are performed to evaluate the number of tilings or partitions in the spatial extension schemes of the BoVWs image representation. Features are densely extracted on a dense grid with pixel stride of 10. Vocabulary construction is done with 2% features collected per image. The number of tilings in each tiling scheme is an important parameter which has to be

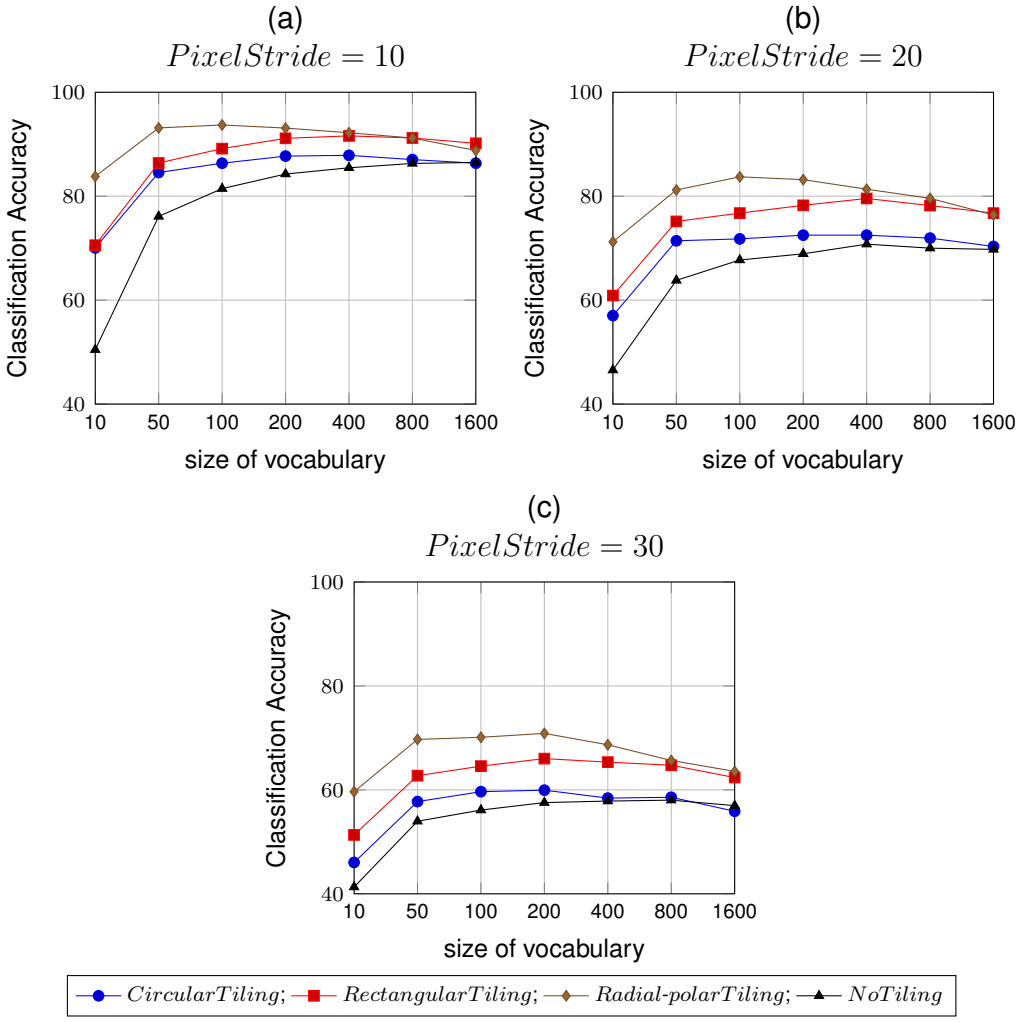


Figure 3.13: Classification accuracy as a function of vocabulary sizes

investigated thoroughly. Following are the number of tilings in each tiling scheme.

1. **Rectangular tiling:** The image space is divided into 1×1, $2 \times 2 = 4$ tilings, $3 \times 3 = 9$ tilings, $4 \times 4 = 16$ tilings and $5 \times 5 = 25$ tilings. The special case 1×1 means no tiling at all.

2. **Circular tiling:** The image space is divided into concentric circles where the number of circles are defined as $\{1, 2, 3, \ldots, 10\}$. The special case of one circle means no tiling at all. Furthermore, the adaptation of the dense grid is also proposed where the dense grid for features extraction is adapted to the circular tilings as shown in Figure 3.14. The features are densely sampled by using equidistant circular grids within each circular tiling denoted by "Adapted-Circular tiling scheme".

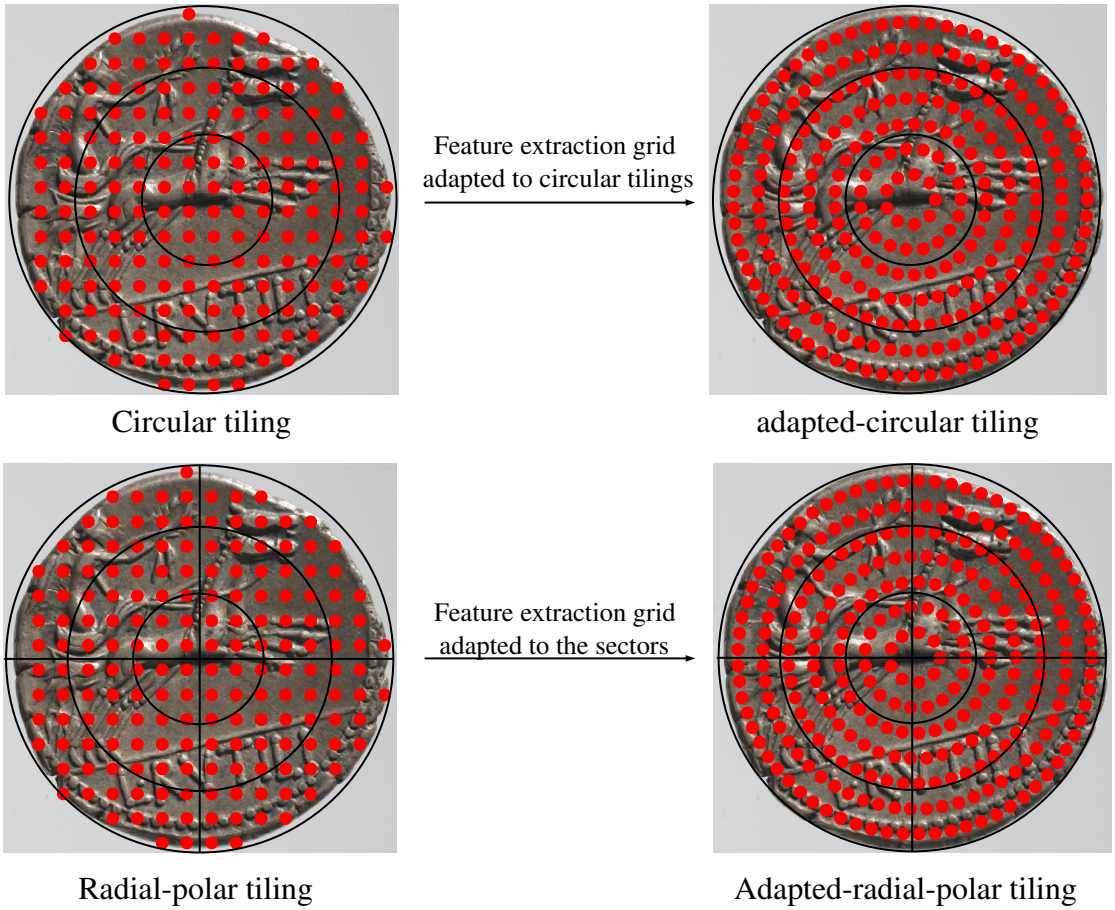


Figure 3.14: Dense grid adaptation to circular and radial-polar tilings

3. **Radial-polar tiling:** The image space is simultaneously divided into both angular and radial tilings. The number of angular θ and radial tilings r are empirically selected as $\{2, 3, \ldots, 10\}$ and $\{4, 6, 8, 12\}$, respectively. For instance, for $\theta = 6$ and $r = 12$, the total number of tilings in the radial-polar tiling scheme is $6 \times 12 \ = 72$. It is also proposed to adapt the dense grid in the radial direction where features are densely sampled by using equidistant circular grids in each radial tiling as shown in Figure 3.14. Such setting is denoted as the "Adapted-radial-polar tiling scheme" because the features extraction is adapted to the sectors of the radial-polar tiling scheme.

Results of the experiments conducted on the number of tilings in each tiling scheme are shown in Figure 3.15. Rectangular and radial-polar tilings achieve the highest classification results. For rectangular tiling 4×4 achieves the best result. The adaptation of the dense grid in the radial direction of the radial-polar tiling achieves marginally improved performance. Due to the non-industrial manufacturing of ancient coins, their image structures are only coarsely aligned among the specimens of a class. Consequently, fewer tilings are favorable and the performance decreases for $r > 3$. Therefore, the best results are achieved with 3 radial and 8 angular tilings.

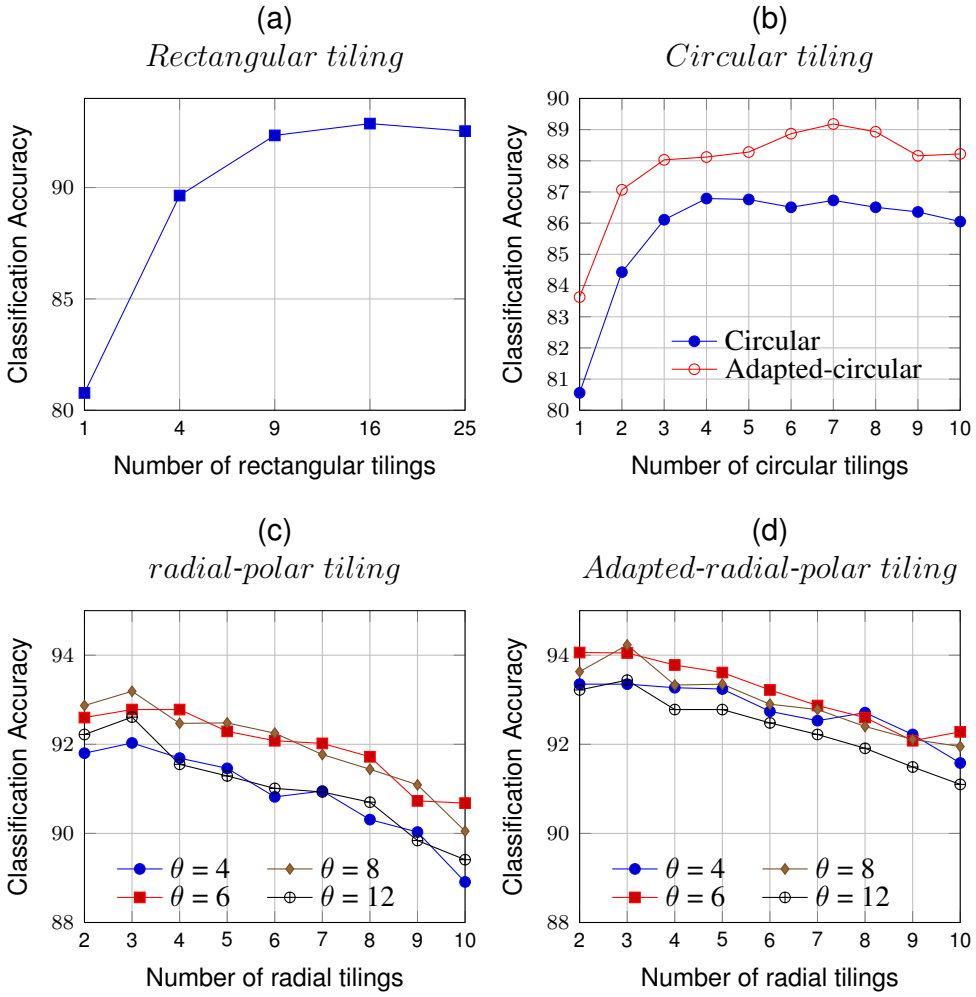


Figure 3.15: Evaluation of the number of tilings in various tiling schemes

In case of circular tiling scheme, the adaptation of the grid to the circular tilings increases the performance by around 3%. It is concluded from these results that due to the detailed nature of rectangular and radial-polar tiling schemes, they achieve better performance than the circular tiling scheme and in case of circular tiling scheme the adaptation of the dense grid to the circular tilings for features extraction achieves improved performance.

3.2.3 Rotation-invariance evaluation

The basic motivation behind the use of tiling schemes is to achieve an improved performance over the BoVWs image representation for problems like object category recognition and scene classification [50, 111]. In case of ancient coins, all the tiling schemes achieve better classification rates than the BoVWs image representation without spatial information. However being

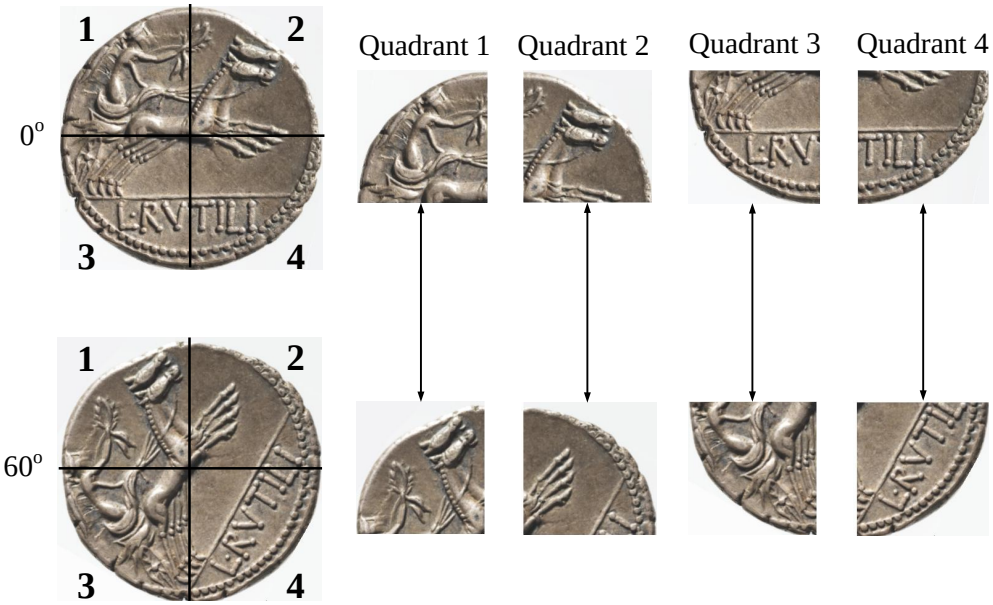


Figure 3.16: Differences among the quadrants of a 2×2 rectangular tiling due to image rotation. The image parts of each quadrant differ from one another due to image rotations

a circular object, a coin can undergo rotations. Thus the tiling scheme that is more robust to rotations is most favorable for the problem of ancient coin classification. In theory, due to rotations, the distribution of visual words in the adjacent rectangular tilings will change. Figure 3.16 shows an example where a 2×2 rectangular tiling is imposed over a coin image and its rotated version. The coin image is rotated through $60°$. The differences among the image parts of each quadrant caused by rotation can be observed. These differences lead to the differences of the distributions of the visual words for each quadrant. Consequently, the histogram of visual words of a given image and that of its rotated version differ from each other. In a similar manner, rotation will cause a change in the distribution of visual words in the adjacent sectors of the radial-polar tiling scheme. In circular tilings such change will not take place as shown in Figure 3.17. It can be observed that the image parts in each circular tiling remain almost constant before and after rotations. Consequently, the distribution of visual words in each circular tiling will remain constant leading to almost unchanged classification results.

Synthetically rotated images are produced to evaluate the spatial tiling schemes for rotation-invariance. The rotated coin images are produced by rotating the images in the test set in $30°$ steps. An example is shown in Figure 3.18. For each image of the test set, a set of images is synthetically produced by rotating it through the aforementioned angles. Therefore our test dataset consists of 12 subsets.

To evaluate the contribution of the particular tiling schemes, the other parameters are kept constant. The size of vocabulary is 200 and the dense grid stride is 10. Rotation-invariant local features are used to achieve rotation-invariance locally. In the experiments SIFT features are used for which the dominant orientations are calculated. However other rotation-invariant local

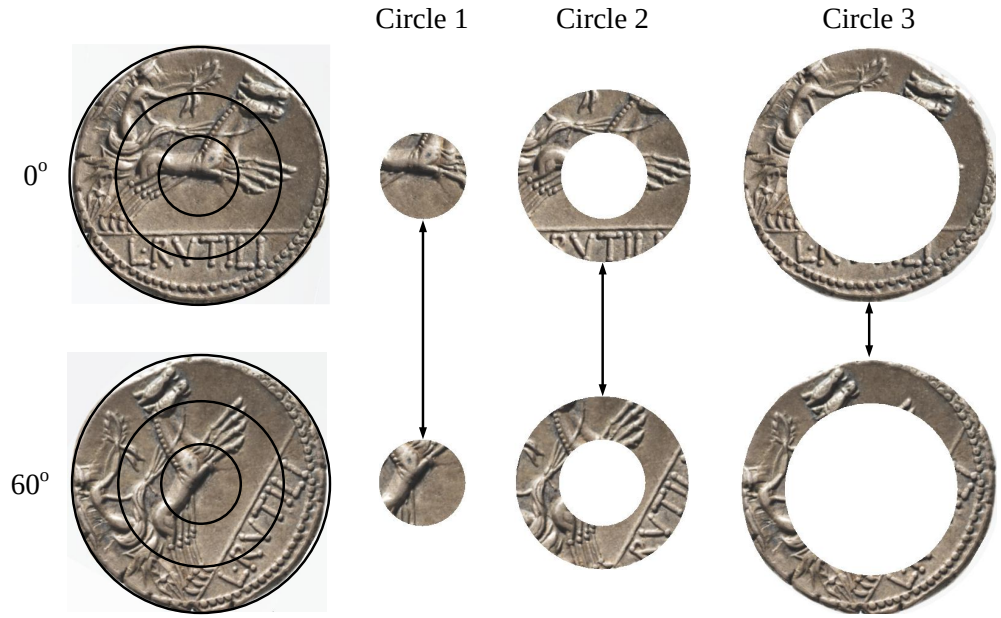


Figure 3.17: Differences among the circular tilings due to image rotation. Unlike rectangular tiling, there are minor differences among the image parts of each circular tiling

features [19, 96] can also be used. Figure 3.19 shows the classification results which validate our hypothesis. Both the circular tiling schemes are more robust to rotations than the rectangular tiling scheme. The variations in the performance of circular tiling is far less than that of rectangular tiling. The performance of the rectangular tiling degrades under severe rotations to such an extent that even the BoVWs representation without spatial tiling performs better than rectangular tiling. The performance of the adapted-circular tiling scheme is better than the circular tiling scheme. It is concluded from these experiments that in the presence of rotations, the dense grid for features extraction adapted to the circular tiling scheme achieves the best results for motif-based ancient coin classification.

3.3 Summary

In the BoVWs model, as a first, the local features are extracted from the images. These local features represent the image patches. The extraction of the local features is performed using a dense grid with a given pixel stride. As a second step, these features are clustered to construct the visual vocabulary. The clustering is performed using the k-means where k is the number of cluster centers. When the k-means clustering stabilizes, identical features are clustered together and their mean represents the cluster center. These centers are kept and the rest of the feature are discarded. The centers are called the *visual word*s and their collection is called the *visual vocabulary*. This visual vocabulary is then used to represent novel images. From a given image, features are sampled using a dense grid with a given pixel stride. These features are then assigned to the the visual words using a similarity measure such as the Euclidean distance. Once the

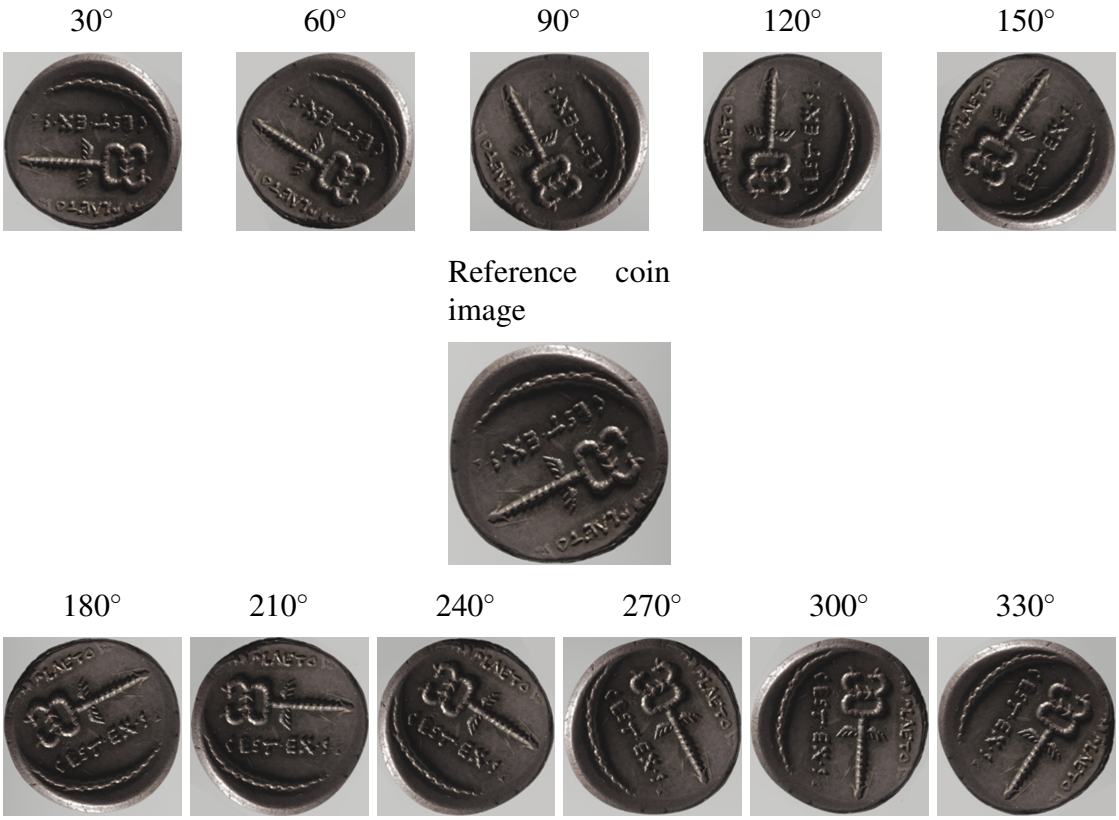


Figure 3.18: Synthetically rotated coin images

feature are represented using the visual words, a histogram is built to represent the image. The number of bins of this histogram is equal to the size of the visual vocabulary and each bin represents the number of times the visual word is occurred in the image. This representation is called the histogram of visual words of the bag of visual words (BoVWs).

However, such representation is orderless because the histogram is built without considering the spatial positions of the local features in the image space. Such information proves discriminating for the classification of regular objects and structured scenes. In the chapter, three schemes are proposed to incorporate such information which are rectangular tiling, circular tiling and the radial-polar tiling. When the image is represented as visual words, before the construction of the histogram any one of these tilings is imposed over the image space. The histogram of visual words is constructed for each partition of the tiling. Finally, the histograms of all the partitions are concatenated to represent the image. However, the main focus of the chapter is on the robustness of the tiling schemes to the image rotations. Since the classification of the ancient coins based on the reverse motif is taken as a motivating example, the ancient coins undergo various in-plane rotations. Due to this reason, the tiling scheme robust to image rotation is most preferred. The main findings of the experiments are summarized in the following.

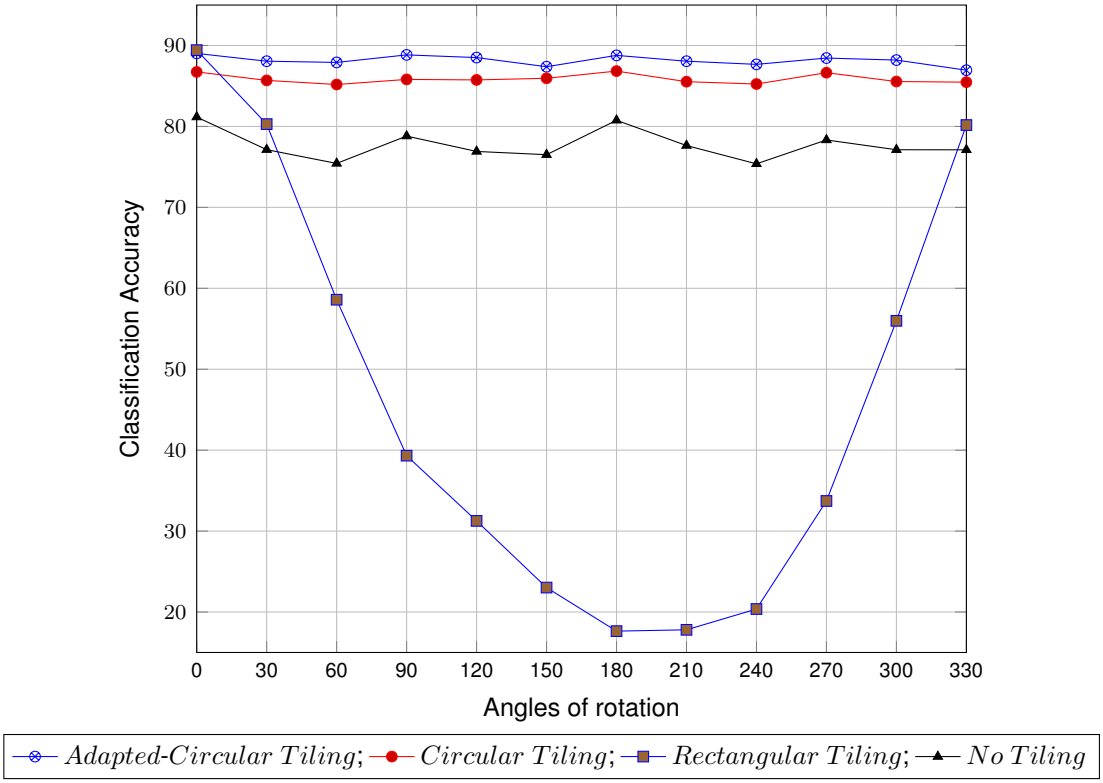


Figure 3.19: Classification accuracy of various tiling schemes on rotated dataset

1. The granularity of the dense grid for features extraction has a significant impact on the classification results. It was found that on the current dataset, a pixel stride of 10 performs better than coarser grids with pixel strides of 20 and 30.

2. The tiling schemes performed better than the BoVWs model without spatial information thus showing that the addition of spatial information to the BoVWs image representation cause a performance boost.

3. The radial-polar tiling performed better than the rectangular tiling and the circular tiling. This is due to the detailed nature of the radial-polar tiling.

4. The variants of the circular tiling and the radial-polar tiling where the dense feature extraction grid is adapted to the circular tilings is also used. It is shown that such adaptation performs better than the non-adapted grid.

5. Finally, it was shown experimentally that both the variants of the circular tiling i.e the adapted and the non-adapted settings are more robust to coin image rotations than the rectangular and radial-polar tiling.

Following are some future directions of the currently proposed image representation.

1. Local rotation-invariance is achieved by calculating the dominant orientation of each local feature. SIFT is used as a local feature due to its popularity and easily available code. However other rotation-invariant local features can also be used.

2. The ancient coins are imaged with non-uniform illumination in most of the images. Zambanini and Kampel [108] proposed the so-called LIDRIC features which are more robust to illumination changes than SIFT features. They showed that their proposed local features outperform SIFT for ancient coin classification. LIDRIC can be used in the BoVWs model instead of SIFT for a better classification rate.

3. The current BoVWs model uses the k-means clustering for the construction of the visual vocabulary. Later, using this vocabulary the image is encoded as the the histogram of visual words. Other encoding methods such as the Fisher vectors have shown greater classification accuracy of state of the art datasets [39]. As a future direction, fisher vectors can be used for local features encoding.

However, the circular tiling is most feasible for circular objects imaged with homogeneous background such as the ancient coins. It is preceded by the automatic segmentation which is also a relatively easier task for objects with homogeneous background. The automatic segmentation helps to achieve invariance to scale changes and translation while the circular tiling imposed over the segmented image achieves robustness to image rotations. These two steps can not be applied to objects imaged with severe background clutter because they can not be segmented automatically. As an example, the butterflies can undergo various in-plane rotations like the ancient coins but they are imaged with severe background clutter. Due to this reason, the circular tiling accompanied by the automatic segmentation can not be used for the image-based classification of butterflies. In the next chapter, an approach is presented to induce the spatial information to the BoVWs model based on the geometric relationship of visual words which is also invariant to changes in scale, translation and image rotations.

CHAPTER 4

Encoding spatial arrangements of local features by modeling their geometric co-occurrences

This chapter presents the methodology to obtain the spatial information of the local features by modeling their geometric relationships in the 2D image space. Based on such information, efficient image representations invariant to scale, rotations and translation are developed to support the image classification.

The proposed method is evaluated on textured objects such as the butterflies. Various species of the butterflies differ from one another based on the texture of their wings which can be used for their image-based classification. Images of the butterflies are mostly taken with flowers, trees and leaves in the background. Such presence of the background clutter makes the image classification for butterflies more challenging. In addition, butterflies are imaged on various scale, with various in-plane orientations and at different positions.

Section 4.1 presents the methodology for the development of a scale-, translation- and rotation-invariant image representation based on the geometric co-occurrences of the local features. To this end, triangulation is performed among the spatial locations of triplets of identical local features. The angles produced by such triangulation are then used to represent the image. Since, angles of a triangle are invariant to changes in scale, translation and rotations, the image representation based on these angles is scale- translation- and rotation-invariant. Results of the experiments performed on the butterfly dataset of 15 classes are shown in Section 4.2. In order to increase the discriminative power of the image representation on a local level, the local features are extracted at several predefined scales. Experiments are performed to optimize for the number of scales on the give dataset. In addition to that, experiments are also performed to evaluate the proposed image representation for rotation-invariance. Finally, Section 4.3 concludes the chapter by outlining its achievements and future directions.

Figure 4.1: Identical patterns on the wings of a butterfly

4.1 Methodology

Identical image patches are found in abundance in symmetrically textured objects such as the butterflies. Identical patches represent identical textural patterns such as those found on the wings of butterflies. Furthermore these identical patterns have proper relative geometry as shown in Figure 4.1 for various species of butterflies. The blobs and patterns are symmetrically found on both the wings at specific positions. This geometric property of the identical image patches can be exploited to induce the spatial information into the image representations to support the image classification.

In this chapter, the problem of classification of butterflies is taken as a motivating example. Figure 4.2 depicts the steps to achieve the image representation based on the geometric relationships of identical local features. As a first step, dense rotation-invariant features are collected at several scales. Experiments are performed to optimize the number of scales for the give dataset. Extracting rotation-invariant features at several scales achieves invariance to image rotations and changes in scale locally. These features are then used to construct the visual vocabulary. Unlike ancient coins in Chapter 3, where the coins are imaged on a homogeneous background, butterflies are imaged in natural environment that presents the problem of the background clutter. Butterflies are imaged with severe background clutter as can be observed from Figure 4.3 that depicts the butterflies species used to evaluate the proposed method. Thus for the construction of the visual vocabulary, the segmentation masks are utilized to use the local features from the object area only. Doing so reduces the effect of the background clutter on the discriminating nature of the visual vocabulary. Lastly, the scale-, translation- and rotation-invariant geometry of the identical visual words is captured using their triangular relationships. Angles and ratios of sides of a triangle are invariant to changes in scale, translation and rotations. The triangular geometric relationship of identical image patches such as those found on the wings of the butterflies is used in this chapter to represent the image. Such image representation is invariant to changes in scale, translation and rotations. Image patches are represented by local features such as SIFT. These local features are extracted from the images using a dense grid. The spatial position of a local feature is given by its position on the dense grid. Therefore the geometric relationships of the spatial positions of the identical local features are modeled to obtain the image representations which is robust to changes in scale, translation and image rotations.

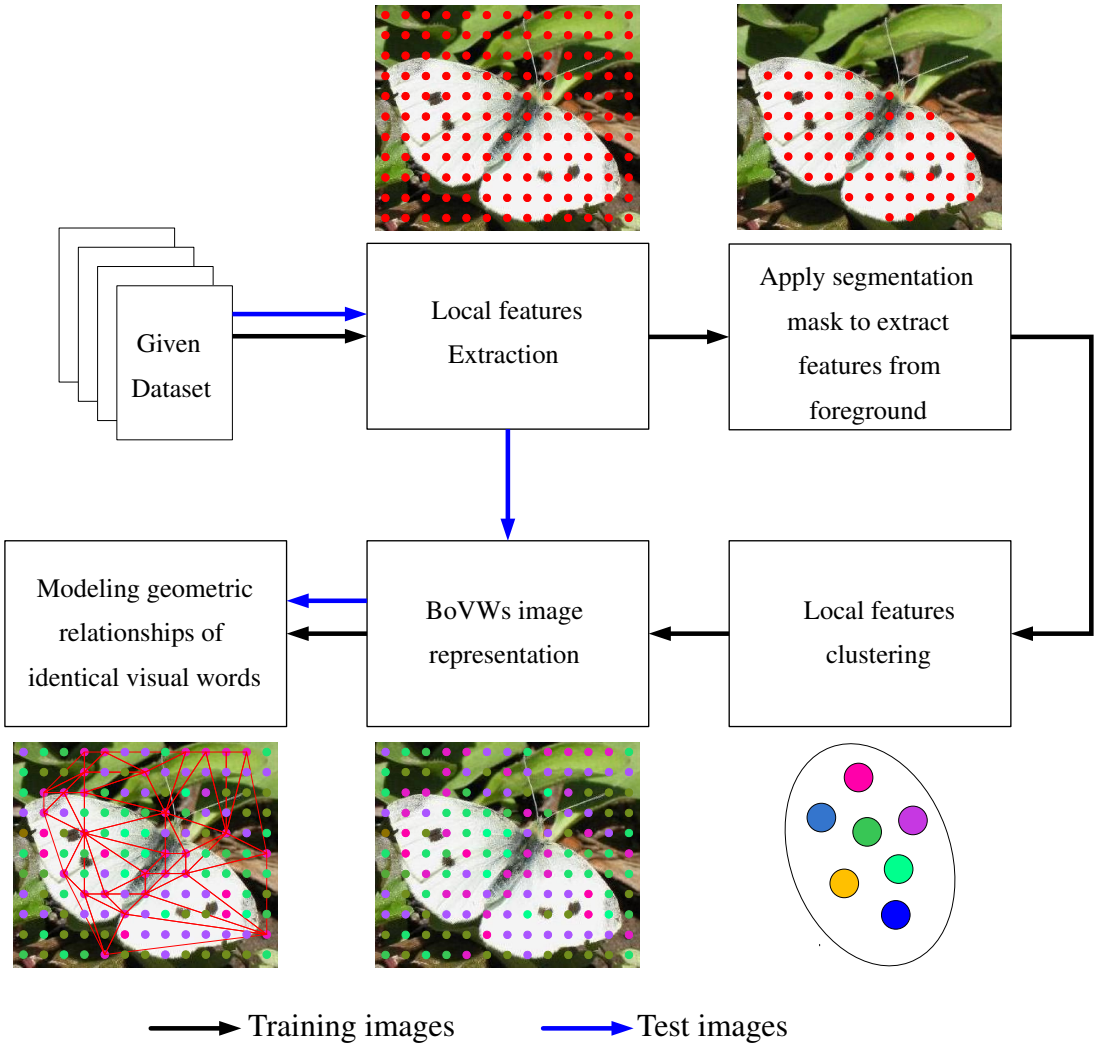


Figure 4.2: Steps

4.1.1 Vocabulary construction and BoVWs image representation

Features are collected from a set of images and quantized using a clustering strategy such as the k-means to form the visual vocabulary. A visual vocabulary $voc = \{v_1, v_2, v_3, \ldots, v_M\}$ consists of M visual words. An image consists of image patches and these patches are represented by local descriptors such SIFT, the given image is first represented as a set of descriptors

$$I = \{d_1, d_2, d_3, ..., d_N\} \tag{4.1}$$

where N is the total number of descriptors. A given descriptor d_k is then mapped to a visual word v_i using some similarity measure like the Euclidean distance as follows:

$$v\left(d_k\right) = \underset{v \in voc}{\arg\min}\ \mathrm{Dist}\left(v, d_k\right) \tag{4.2}$$

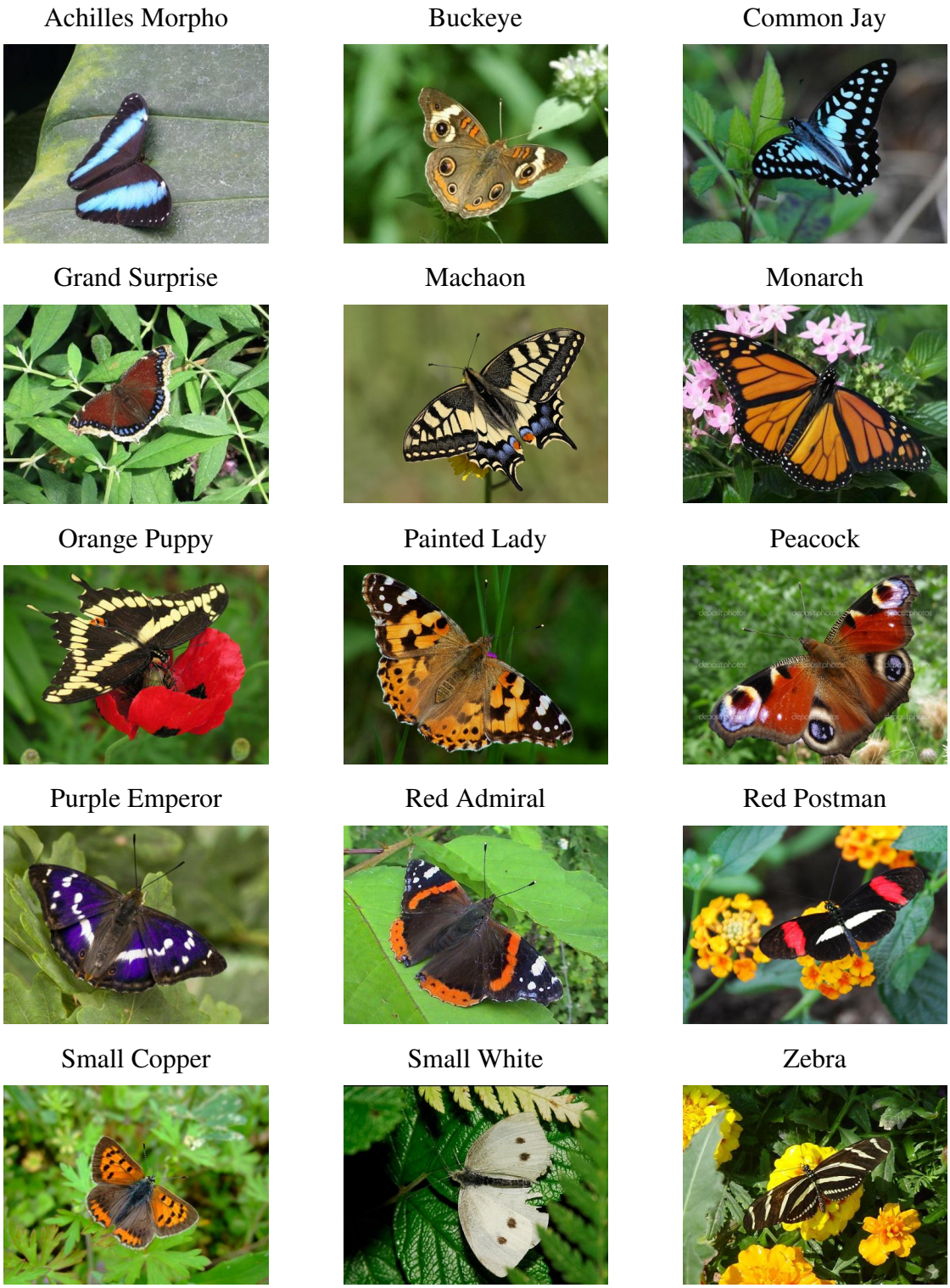


Figure 4.3: Exemplar images of butterflies species used for experiments

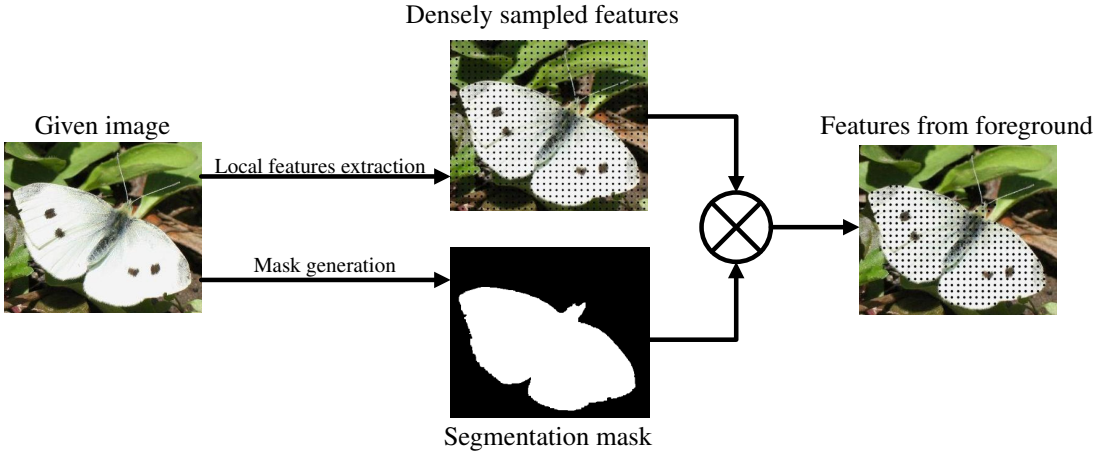


Figure 4.4: Segmentation masks utilized for feature extraction at the stage of vocabulary construction

where d_k is the k^{th} descriptor in the image and $v\left(d_k\right)$ is the visual word assigned to this descriptor based on the distance Dist (v, d_k). The given image is then represented as the histogram or the bag of visual words (BoVWs) where the number of bins of this histogram is equal to the size of the visual vocabulary which is M.

Local features are densely extracted from images to construct the visual vocabulary. These densely extracted features consist of features from the background and the object area or the foreground. Visual vocabulary is prone to contamination due to the presence of the features from the background. For instance, the butterflies are imaged under severe background clutter and it is more likely that the features from the background affect the discriminating nature of the visual vocabulary in a negative manner. There exist specialized methods [61, 62] to learn a discriminating vocabulary however these methods are computationally expensive. For the sake of simplicity, in case of butterflies, segmentation mask are manually generated for the process of vocabulary construction. These segmentation masks are used to extract the features from the foreground for the construction of visual vocabulary. Figure 4.4 shows the process of the extraction of dense features from the foreground with the help of a segmentation mask.

4.1.2 Geometric modeling of identical visual words for spatial information incorporation to BoVWs

The value of the bin b_i of the histogram of visual words gives the number of occurrences of a visual word v_i in an image. Let D_i be the set of all descriptors mapped to a visual word v_i, then the i^{th} bin of the histogram of visual words b_i, is the cardinality of the set D_i.

$$b_i = Card\left(D_i\right) \text{ where } D_i = \{d_k, k \in [1, \ldots, M] \mid v\left(d_k\right) = v_i\} \tag{4.3}$$

The set D_i consists of all descriptors that are assigned the same visual word v_i. In order to incorporate the spatial information to the BoVWs representation of image, Khan et al. [42] proposed a method using identical visual words. All the pairs of the two distinct descriptors from a

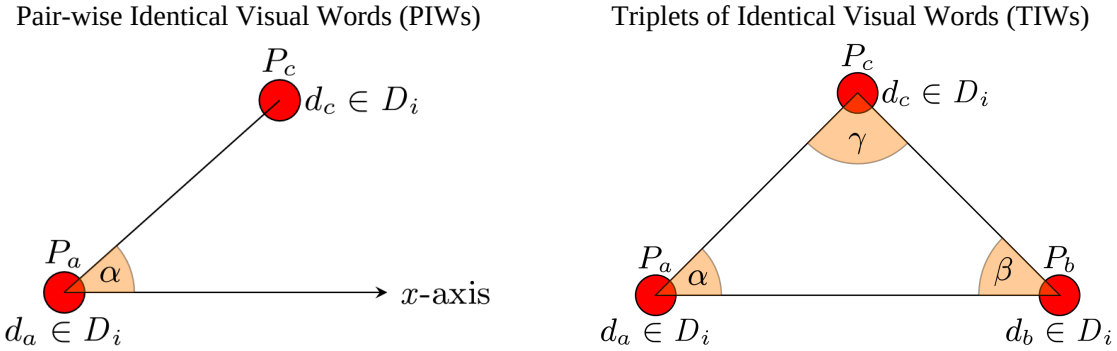


Figure 4.5: *PIWs* and *TIWs*

given set D_i are taken which are called pair-wise identical words (*PIWs*) as shown in Figure 4.5. Angles made by members of each pair of *PIWs* with respect to x-axis in the 2D image space are calculated. The spatial positions of the descriptors on the dense grid are used for angles calculation. An angles histogram is then built from these angles to represent the image which they call pair-wise identical words angle histogram (*PIWAH*). *PIWAH* representation of images is invariant to changes in scale and translation. However it is not rotation-invariant as angles are calculated with respect to the x-axis. Therefore their proposed method is modified by considering triplets of identical words *TIWs* to calculate angles. This will make a rotation-invariant triangular relationship among the words of a given pair of *TIWs* as shown in Figure 4.5. The triangular relationship among the members of a pair of *TIWs* is invariant to changes in translation, scale and rotation. However such a relationship is global for which the local features must also be scale- and rotation-invariant. Therefore the underlying BoVWs image representation is built upon rotation-invariant local features that are extracted at several scales.

The set of all *3-PIWs* related to a visual word v_i is defined as:

$$TIW_i = \left\{(P_a, P_b, P_c) \mid (d_a, d_b, d_c) \in D_i^3, d_a \neq d_b \neq d_c\right\} \tag{4.4}$$

where P_a, P_b and P_c are the spatial positions of the descriptors d_a, d_b and d_c respectively. The value of the i^{th} bin of the histogram shows the frequency of the visual word v_i. Therefore the cardinality of TIW_i is ${}^{b_i}C_3$ which is the number of all possible subsets of three distinct elements among the elements of D_i. However, those pairs of *TIWs* in which all the three words are collinear are not considered for angles calculation. In order to suppress very small angles, triangles made by pairs of non-collinear *TIWs* where one of the three angles is less than 5° are also ignored for angles calculation. Consequently, for a word v_i, the number of candidate *TIWs* for angles computation is most likely less than ${}^{b_i}C_3$. Such triangulation is termed as the *combinatorial triangulation* which is computationally expensive. For instance, if the cardinality b_i of the set D_i is 80 then the number of unique three combinations is 82160. Therefore, to compute the angles among the members of *TIWs*, the *Delaunay triangulation* is also evaluated where the number of triangles is much smaller. In *Delaunay triangulation*, the three points should not be collinear and the circumscribed circle defined by the three points should not contain any

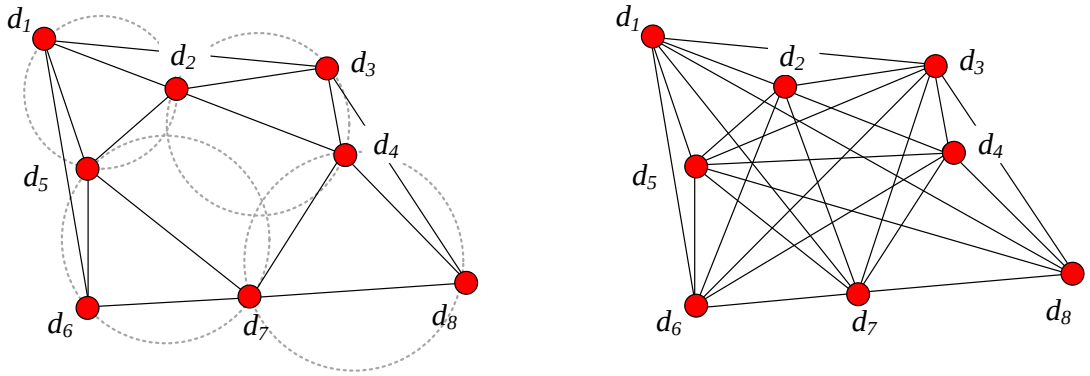


Figure 4.6: Various triangulation methods

other point. The principles of the *Delaunay triangulation* significantly reduce the number of the candidate *3-PIWs* for angles computation. Figure 4.6 shows both the *Delaunay* and the *combinatorial* triangulations. It can be observed that for 8 descriptors belonging to a visual word, *combinatorial triangulation* results in 56 triangles while the *Delaunay triangulation* results in 9 triangles. The three angles shown in Figure 4.5 made by members of a given *TIWs* are calculated according to the law of cosines. For simplicity in Equation 4.5, sides of the triangle shown in Figure 4.5 are named as $a = \overline{P_b P_c}$, $b = \overline{P_a P_c}$ and $c = \overline{P_a P_b}$.

$$\alpha = \arccos\left[\left(a^2 + c^2 - b^2\right)/2ac\right] \tag{4.5a}$$

$$\beta = \arccos\left[\left(a^2 + b^2 - c^2\right)/2ab\right] \tag{4.5b}$$

$$\gamma = \arccos\left[\left(b^2 + c^2 - a^2\right)/2bc\right] \tag{4.5c}$$

The angles histogram is built from these angles for which the bins are empirically chosen between 0° and 180°. The angles histogram for a specific word v_i is named as $TIWAH_i$. Khan et al. [42] proposed a 'bin replacement' technique to combine the $TIWAH_i$ from all the visual words. The bin b_i of the histogram of visual words associated with visual word v_i is replaced with $TIWAH_i$ in such a way that the spatial information is added without altering the frequency information of v_i. Finally $TIWAH_i$ of all the visual words are combined to represent a given image.

$$\begin{gathered} TIWAH = (\psi_1 TIWAH_1, \psi_2 TIWAH_2, \ldots, \psi_M TIWAH_M) \\ \text{where} \quad \psi_i = \frac{b_i}{\left\| rPIWAH_i \right\|} \end{gathered} \tag{4.6}$$

where ψ_i is the normalization coefficient. For a visual vocabulary of size M, if the number of bins in angles histogram is θ, then the size of the *TIWAH* is $M\theta$. This histogram is invariant to changes in scale, translation and rotation because it is built upon the triangular relationship of the *TIWs*.

4.2 Experiments and results

Experiments are performed on all the classes of the Leed's butterfly dataset [95] and five more classes which are 'Achilles Morpho', 'Common Jay', 'Machaon', 'Peacock' and 'Purple Emperor'. Google image search is used to obtain images of these five classes. In total 1171 images of butterflies are used that belong to 15 different classes. The training set consists of 479 images while 692 images are used for validation. Experiments are performed in a step-by-step manner to optimize for a number of parameters of the BoVWs model. As a first step, the effect of using the segmentation mask for vocabulary construction is observed. Since the local features are densely sampled from the images, they may contain local features from the background. The images of the butterflies contain background clutter such as leaves, trees and stones etc, they may affect the discriminating nature of the visual vocabulary in a negative manner. Since the images contain the butterflies at various scale, experiments are performed to optimize for the number of scales at which the local features must be extracted from the give dataset. Another issue worth experimenting is the triangulation method as it is an integral part of the proposed image representation. Both the combinatorial and the Delaunay triangulations are evaluated for efficiency in terms of time. The size of the visual vocabulary is also an important parameter [68] and needs to be optimized for the given dataset. Experiments are also performed for the size of the training set. Finally, since the butterflies can undergo various in-plane rotations, the proposed image representation is evaluated for rotation-invariance on synthetically rotated butterfly images.

Classification is performed using the one-vs-all mode of SVM. The images are represented as histograms of angles based on the triplets of identical visual words (TIWAH). These histograms are then used as feature vectors for classification with an SVM. However, before feeding them to SVM, the feature vectors are explicitly transformed using a Helinger kernel as proposed by [92]. They call it an approximation of the Helinger kernel. Once the feature vectors are transformed, they are used with a linear SVM for classification. The best value of the regularization parameter 'C' is calculated using the n-fold cross-validation.

4.2.1 Segmentation for vocabulary construction

TIWs are used for angles calculation and then use these angles to form the angles histogram for image representation. Similar or identical image patches should be assigned to identical visual words from the vocabulary. Therefore the visual vocabulary should be discriminative enough to assign identical words to identical image patches. For vocabulary construction, dense sampling is used where features are extracted from the foreground as well as the background. These features are quantized by the k-means clustering to form the visual vocabulary. Due to the unsupervised nature of k-means, the visual vocabulary is more likely to be non-discriminative and noisy. Therefore the utilization of segmentation is proposed at the stage of vocabulary construction as shown in Figure 4.4. Segmentation will help to select features for vocabulary construction from proper areas of the images.

The segmentation masks for the butterfly images are used which are generated by the a semi-automatic segmentation method proposed by [93]. The backgrounds of images in the butterfly dataset consists of trees, grass, stones, sky and flowers as shown in Figure 4.3. Therefore,

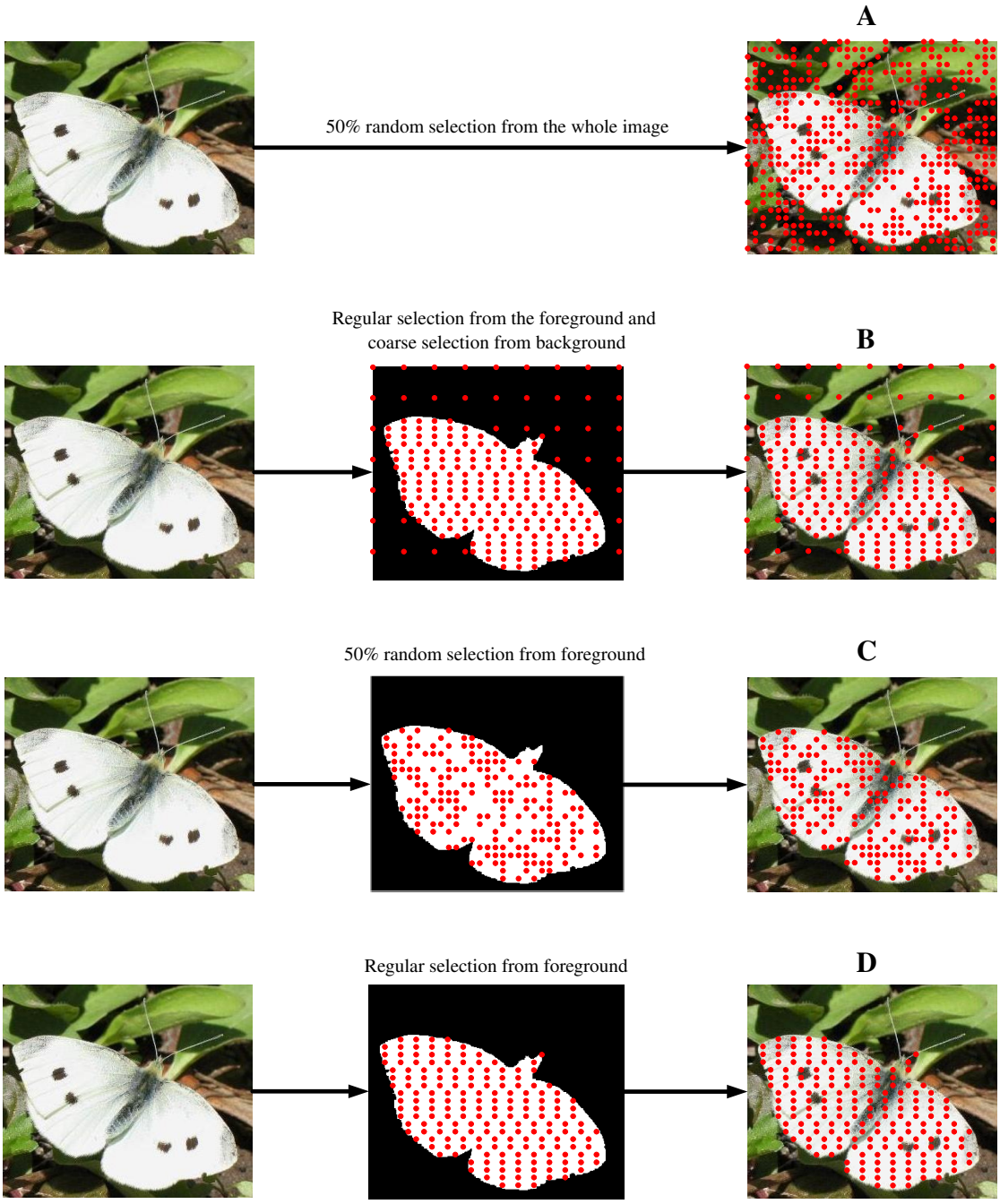


Figure 4.7: Various settings of feature selection from images with and without the use of segmentation masks

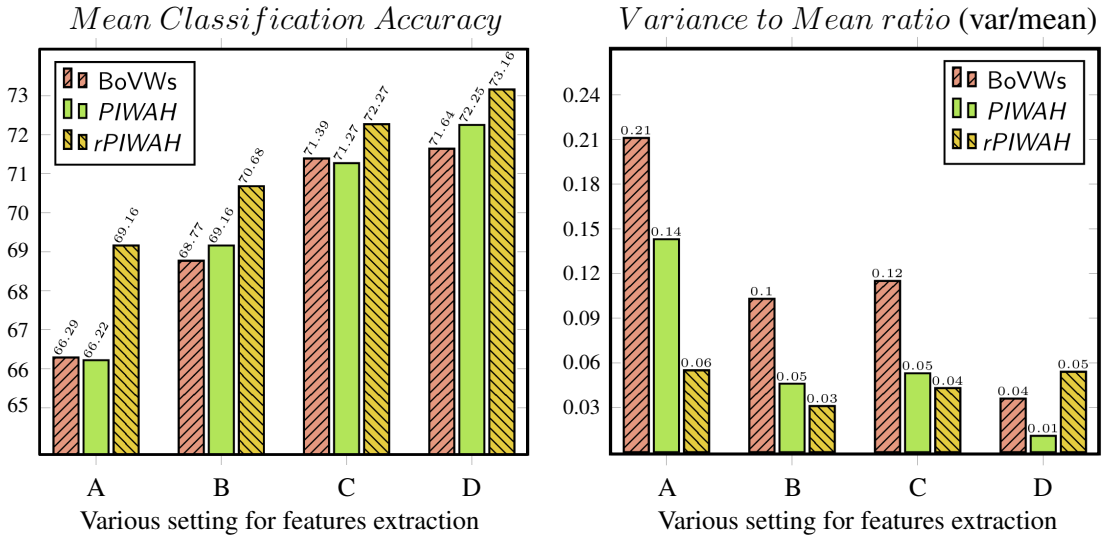


Figure 4.8: Mean classification accuracy and variance to mean ratio

the use of segmentation will have a significant effect on the discriminating nature of the visual vocabulary. Local features are extracted from the images using a dense grid with a constant pixel stride of 10. The following settings are used to sub-select features for vocabulary construction using the segmentation masks.

1. **50% random selection from the whole image:** From each image, 50% features are randomly selected among the densely extracted features. These include features from the background as well as the foreground as shown in Figure 4.7A.

2. **Regular selection from the foreground and coarse selection from the background:** From each image, features are regularly selected from the foreground and coarsely extracted from the background. It can be observed from Figure 4.7B that features from the foreground are extracted by selecting every second feature in each row. However from the background, local features are selected using a larger value of the pixel stride.

3. **50% random selection from the foreground:** 50% features are randomly selected among the densely extracted features from the foreground only as shown in Figure 4.7C.

4. **Regular selection from the foreground:** From each image, features are regularly selected from the foreground only. In each row, every second feature is considered for vocabulary construction as shown in Figure 4.7D.

The features for vocabulary construction are quantized using k-means clustering where the cluster centers are initialized randomly. Apart from that, in settings 'A' and 'C', features are randomly selected for vocabulary construction. Due to these reasons, for each of the above settings, experiments are performed ten times and the mean classification accuracies of all the three methods i.e. BoVWs, *PIWAH* and *TIWAH* are reported in Figure 4.8. The variance to mean ratio of

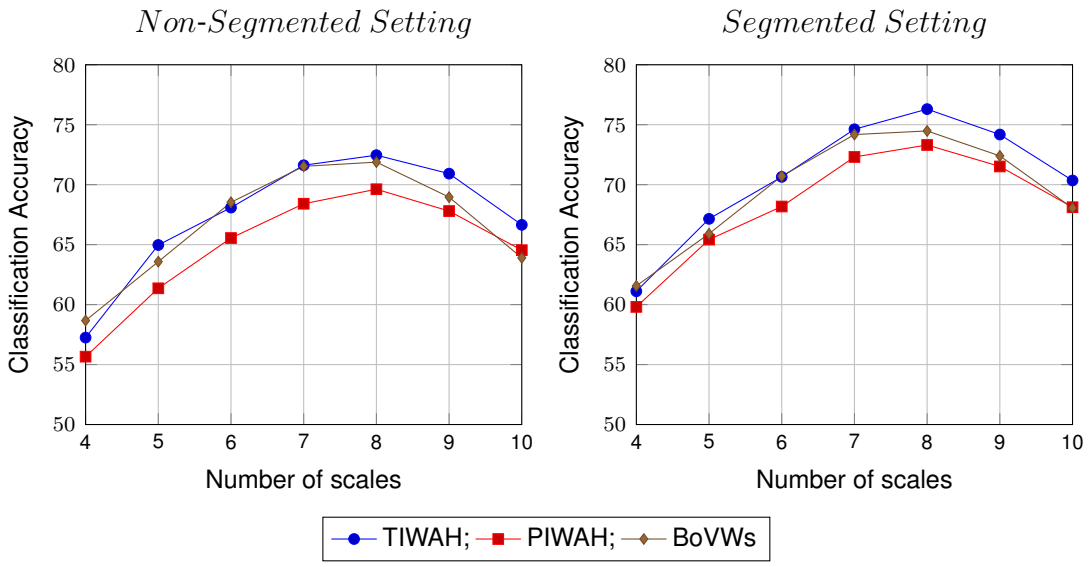


Figure 4.9: Results for the number scales in both the segmented and non-segmented settings. Both have a maximum at 8 scales and TIWAH performs better than PIWAH and BoVWs

each method for all the settings are also reported in order to show their robustness to the randomness of the k-means and features selection. Each time, a visual vocabulary is constructed for image classification. For the image representation of training and test sets, segmentation is not used. For 'C' and 'D' in Figure 4.8, the classification rates are higher because the vocabularies are only constructed from the foreground i.e. the object area. Therefore the visual words are more discriminating resulting in better performance than 'A' and 'B'. It is noteworthy that the proposed image representation not only outperforms the other methods but also it has the least variance to mean ratios on settings 'A', 'B' and 'C'. In case of 'A', features are randomly selected from the whole image while in case of 'C', they are randomly selected from the foreground only. The proposed image representation has the least variance to mean ratio on both 'A' and 'C' which shows that it is robust to the randomness of feature selection. In addition, our proposed image representation outperforms the other methods on setting 'A'. This shows that *TIWAH* performs better even if the segmentation masks are not used for vocabulary construction. In method 'B', features are coarsely selected from the background as well. Due the variance in the background, the variance to mean ratios of BoVWs and *PIWAH* are higher. The proposed image representation has the least variance to mean ratio in 'B' which shows that it is also robust to the variations in the background. Therefore it is concluded that the use of segmentation masks during vocabulary construction results in higher classification rates. Over all the proposed image representation method outperformed BoVWs and *PIWAH* on both segmented and non-segmented images and it is more robust to the randomness of both the k-means and the feature selection.

4.2.2 Number of scales for local features extraction

The images depict butterflies at various scales due to which the local features such as SIFT are extracted at several scales to achieve robustness to such changes in scale. Optimization on the predefined number of scales is done for the given dataset. Starting from a single scale of 2, features are extracted at 10 scales where a given scale is $\sqrt{2}$ multiple of its predecessor. Predefined scales are $\{2, [2\ 4], [2\ 4\ 6], \cdots, [2\ 4\ 6\ 8\ 12\ 16\ 22\ 32\ 45\ 64]\}$. From a given image, local features extracted at these scales are concatenated in a single feature vector. Vocabulary size is 200 and is constructed with local features that are extracted from the training set. The segmentation masks are manually generated and used to extract features from the image regions that contain butterflies. These features are then utilized for vocabulary construction thus reducing the contaminating effect of the background features on vocabulary. Since the method depicted in Figure 4.7D performed the best as shown in Figure 4.8, it is utilized to select the features for the construction of the visual vocabulary. Results for the experiments on the number of scales for non-segmented and segmented settings are shown in Figure 4.9 from which four main conclusions can be drawn. First, the performance of each method increases with increase in the number of scales in both segmented and non-segmented settings. Second, each method has a maximum at 8 scales. Third, the maxima of the segmented setting is greater than those of the non-segmented setting. Fourth, in segmented setting, *TIWAH* clearly outperforms *PIWAH* and the simple BoVWs image representation. Therefore it is concluded that on the current dataset for a vocabulary size of 200, 8 scales for feature extraction are suitable and segmented images used for vocabulary construction perform better than non-segmented images.

4.2.3 Computational complexity

To evaluate the efficiency of both the Delaunay and combinatorial triangulation schemes, experiments are performed to compare their classification accuracies and the computation time they take for a given set of images. The training and test sets are used to compare the classification accuracies of both the schemes. Rotation-invariant local features are densely extracted from the images at 8 scales. Experiments are performed 10 times for each triangulation scheme and the mean classification accuracies are reported in Table 4.1. In each experiment, a separate vocabulary of size 200 is constructed.

To find the time taken by each triangulation scheme for image representation, 20 images of 'Machaon' are selected due to its dense texture. Images are of standard size which is 640×480. A faster 'C' language implementation of the combinatorial triangulation is also evaluated which is named as *combinatorialF* triangulation. The experiments are run 10 times on a single

Table 4.1: Classification rates and time in seconds for each triangulation scheme

	Classification Accuracy	**Time**
Combinatorial triangulation	77.66	1860.30
CombinatorialF triangulation	-	201.61
Delaunay triangulation	**76.41**	**7.01**

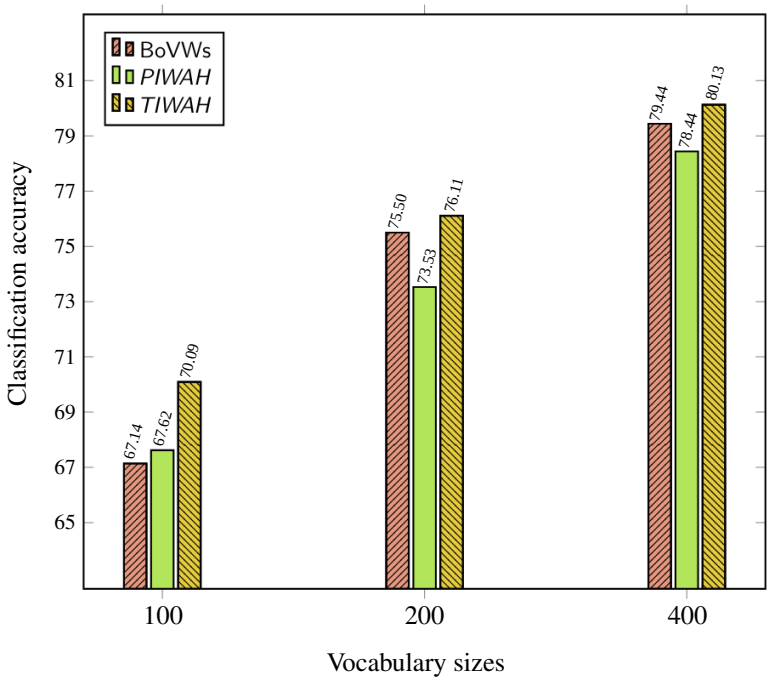


Figure 4.10: Performance of BoVWs, PIWAH and TIWAH on various vocabulary sizes

core and the mean time in seconds taken by each scheme is reported in Table 4.1. It can be concluded from Table 4.1 that the Delaunay triangulation performs much efficiently than the combinatorial triangulation. The Delaunay triangulation even performs efficiently than the 'C' implementation of the combinatorial triangulation. However, the image representation obtained using the combinatorial triangulation performs marginally better than the image representation obtained using the Delaunay triangulation. However, this difference can be compensated by using a more discriminating visual vocabulary achieved with specialized techniques such as the one proposed by [61].

4.2.4 Selection of size for vocabulary and training set

The sizes of vocabulary and training set are also evaluated on the current dataset. The sizes of the vocabulary are predefined as $\{100, 200, 400\}$. The results are shown for all the methods in Figure 4.10. Although *TIWAH* outperforms the other two methods yet it is more discriminating on smaller vocabulary sizes. The main reason behind this is the relationship of the similarity of the visual words and the vocabulary sizes. The vocabulary is constructed using the k-means clustering where the sizes of the clusters are inversely proportion to the number of clusters. Each cluster center represents a visual word. It means that with the increase in the number of clusters, less features are assigned to them. Since the local features assigned to the same cluster center or visual word are treated as identical, the similarity of the local features decrease with the increase in the size of the vocabulary. If the extreme case is considered such that each local feature is treated as a visual word, the similarity of the local features will be zero. It will cause

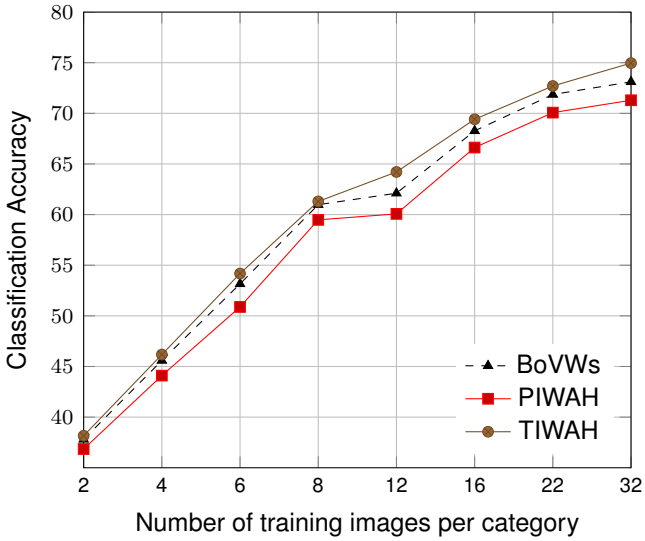


Figure 4.11: Performances of various methods with respect to training set sizes

over-fitting and the resultant angles histogram will be flat. Due to this reason, *TIWAH* is more discriminating on smaller vocabulary sizes. This phenomenon can be avoided by using the soft assignment where a given feature is assigned to a number of visual words in a weighted manner. The assignment in the k-means clustering is hard where a given feature is assigned to exactly one visual word. The soft assignment can perform better than the hard assignment as shown by a recent work [41].

An important parameter of the BoVWs model is the size of training set [68]. The performance of all the methods increase with the increase in the training set size as shown in Figure 4.11. However the superiority of *TIWAH* is more pronounced for bigger training set sizes.

4.2.5 Rotation-invariance

The proposed method is evaluated for rotation-invariance along with the BoVWs and *PIWAH*. Since the proposed method is an extension of *PIWAH* to achieve the rotation-invariance in the geometry, therefore the rotation-invariance of *PIWAH* is also evaluated and shown for comparison. However, the test set lacks the images to evaluate the mentioned methods for rotation-invariance. Due to this reason, as a first step, the rotated test set is synthetically generated. In order to emphasize on the evaluation of rotation-invariance, the segmentation masks are used to suppress the background in all the test images. These masks are used to extract the foregrounds i.e. the butterflies and paste them on a homogeneous backgrounds as shown in Figure 4.12. Doing so avoids the background clutter as well as the artifacts created due to the rotation of images. These artifacts are likely to create bias in the rotated test subsets. The images in the test dataset with homogeneous backgrounds are then rotated through predefined angles which are [30, 60, 90, 120, 150, 180]. Thus the test set consists of 7 test subsets. Exemplar rotated images

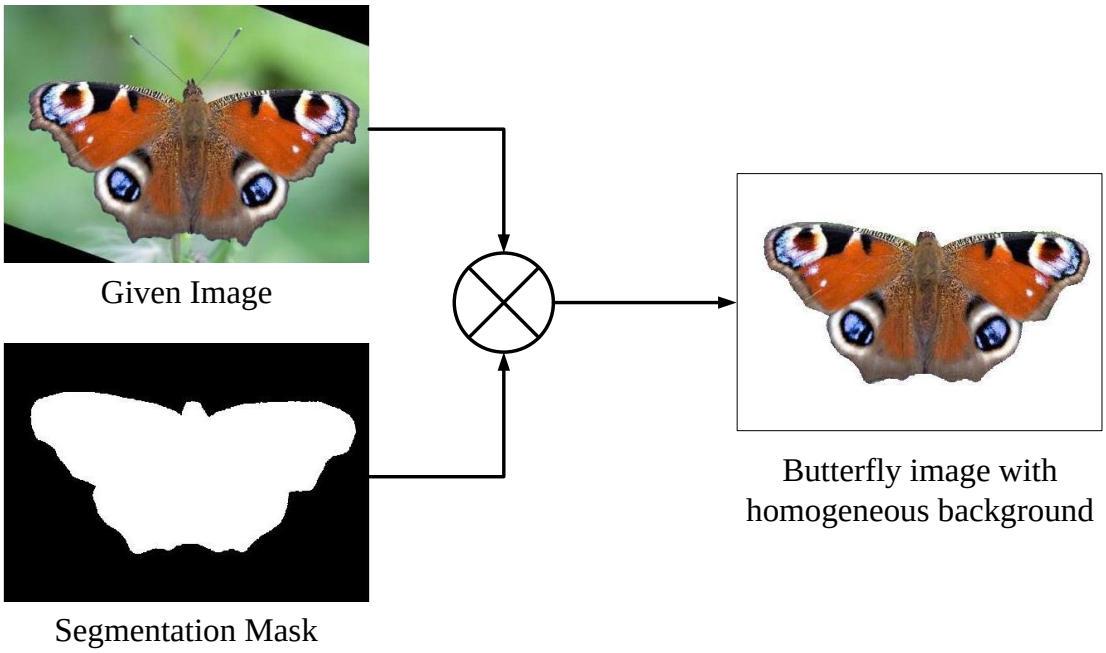


Figure 4.12: Background clutter and the artifacts in the image produced due to rotation are removed using the segmentation mask. The butterfly image is pasted on a homogeneous background.

with homogeneous background are shown in Figure 4.13. Once the rotated test sets are generated, rotation-invariant local features are extracted from these images at 8 scales thus achieving rotation and scale invariance locally.

The experiments for rotation-invariance are performed 10 times and the mean classification accuracy of each method is reported. In each experiment, the visual vocabulary of size 200 is generated using the training set. The images in the training set contain the usual background clutter. From these images, rotation-invariant local features are densely extracted at 8 scales and are used to construct the visual vocabulary as well as to train the model. Results shown in Figure 4.14 indicate that *TIWAH* is more robust to image rotations than the ordinary BoVWs image representation and the method proposed by Khan et al. [42]. This is due to the fact that *TIWAH* is based on the triangular relationships of identical visual words which prove to be more robust to the image rotations than the method of Khan et al. [42] which performs even worst than the simple BoVWs model on the rotated images.

4.3 Summary

The spatial information is induced into the BoVWs model by capturing the geometric co-occurrences of identical visual words. The image patches are represented by the local features such as SIFT, which are then assigned to the visual words of the visual vocabulary based on a

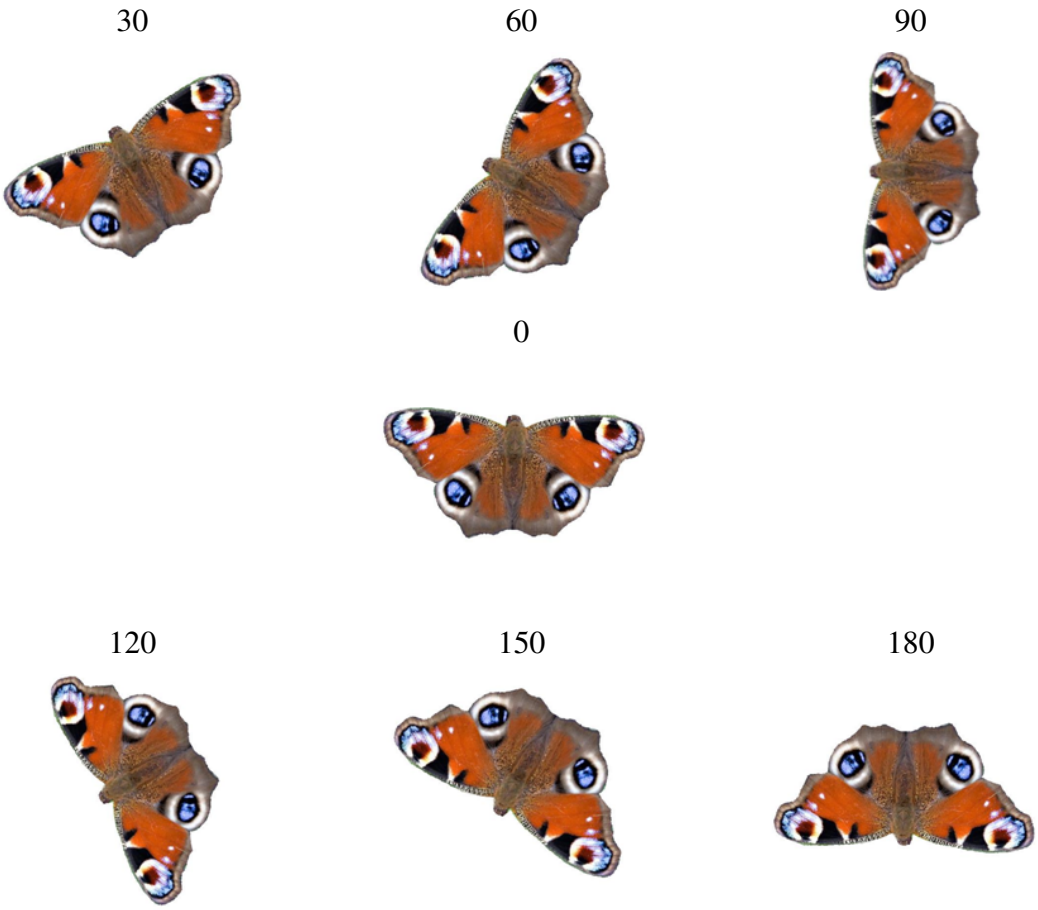


Figure 4.13: Synthetically rotated images with homogeneous backgrounds

similarity measure such as the Euclidean distance. Identical image patches are assigned identical words . The triangular geometric relationship of these visual words is achieved using angles of triangles made by the spatial locations of the visual words in the 2D image space. Since the angles and ratios of the sides of a triangle are invariant to changes in scale, position and rotations, this triangular geometric relationship among identical visual words is invariant to such image transformations. In a given image, identical visual words are triangulated and the angles are computed using two kinds of triangulation methods. The first method considers all the unique pairs of the instances of any given visual words which is called the *combinatorial triangulation*. The second triangulation method is the *Delaunay triangulation* which is a well known triangulation method from the computational geometry. The angles computed using any given triangulation method are then used to construct the angles histogram which is then used to represent the image.

Experiments are performed on the images of 15 different types of butterflies. Due to the fact that the butterflies are imaged in natural environment, the images contain severe background

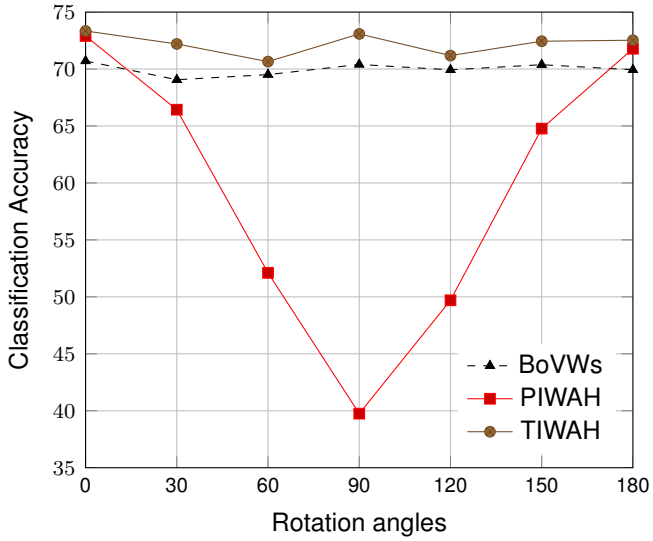


Figure 4.14: Rotation-invariance evaluation of various methods

clutter. The butterflies are imaged at various scales and with different rotations. Therefore, experiments are performed in a systematic way to optimize for the parameters of the proposed image representation to tackle with these issues. Following are the major findings of the experiments.

1. The use of segmentation masks at the stage of vocabulary construction results in higher classification rate. This is due to that fact that the discriminating power of the vocabulary increases when local features from the foreground are used to construct it.

2. On the current dataset, rotation-invariant local features densely extracted at 8 scales achieves the highest classification rate for the given dataset.

3. The Delaunay triangulation triangulates the positions of the identical visual words faster than the combinatorial triangulation at the cost of a marginal decrease in the classification rate.

4. On the current dataset, the proposed image representation is more discriminating than the BoVWs model at lower vocabulary sizes i.e. 200 words.

5. The proposed image representation is more robust to image rotations than the BoVWs model and the method presented by Khan et al. [40].

Following are the possible future directions of the current proposed image representation

1. The main purpose of using the segmentation masks is to use the features from the foreground to construct the visual vocabulary. Doing so enhanced the discriminating ability of the visual vocabulary. These segmentation masks are manually generated thus requiring

more human effort and time. Therefore a possible future direction of the current research would be to increase the discriminating nature of the visual vocabulary without the use such segmentation masks.

2. The number of scales are optimized on the current dataset and it needs to be done again in future if further classes and images are used. However, the method presented by [87] can be used which locally optimizes for the scale of the densely extracted features.

3. Instead of SIFT, other local features can be used such scale-invariant descriptor (SID) [45] and scale-less SIFT (SLS) [30] which are more robust to scale changes than SIFT.

CHAPTER 5

Encoding geometric co-occurrences of local features in image subspaces

The image representation presented in this chapter is based on the techniques presented in Chapter 3 and Chapter 4. As a first step, the object image region is segmented automatically as the background is homogeneous. The segmented image is then normalized and cropped to achieve invariance to changes in scale and object position. As a second step, the segmented image space is divided into sub-regions of circular shapes thus achieving invariance to image rotations. Finally, the rotation-invariant geometric relationships of local features are modeled in each sub-region to represent the image.

Section 5.1 presents the methodology to develop the proposed image representation which is invariant to changes in scale, position and image rotations. Since this image representation is based on the BoVWs model, circular tiling scheme and the geometric co-occurrences of the visual words, an overview of these terms is outlined. Results of the experiments performed to optimize for the values of various parameters such as the size of visual vocabulary and the number of partitions in the circular tiling are given in Section 5.3. The evaluation of the proposed image representation for rotation-invariance is also performed on rotated coin images. The contributions and future directions of the chapter are outlined in Section 5.4.

5.1 Methodology

In order to derive the image representation, the method presented in this chapter uses rotation-invariant geometric relationships of identical local features that are derived in rotation-invariant sub-regions of an image. The development of the proposed image representation requires minimal or no background clutter so that the object of interest can be automatically segmented. In addition, the proposed image representation is most suitable for circular objects because they can exhibit various in-plane rotations. Therefore the problem of classification of ancient coins based on reverse motif recognition is taken as a motivating example. The reverse motifs of the

Figure 5.1: Reverse motifs used for classification

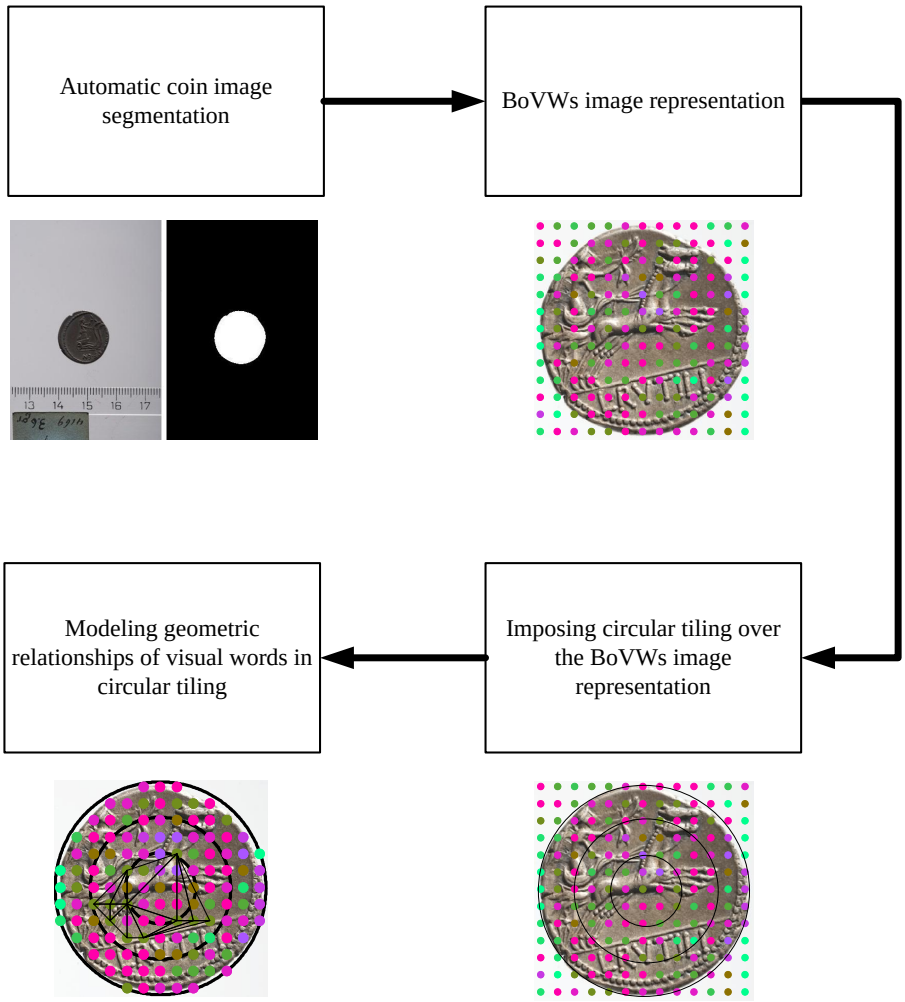


Figure 5.2: Steps for scale- translation- and rotation-invariant image representation used for coarse-grained classification of ancient coins

ancient coins used for experiments are shown in Figure 5.1. From the images of the ancient coins, several challenges can be observed. First, due to the flat and circular nature of the coins, they can exhibit rotations in the image plane. Such significant rotations can be observed in reverse motifs such as *Griffin* and *Gallop rider*. In addition to rotations, the images can come from different datasets and thus the coin region can be in different scales and at different image positions. Consequently, the method should be invariant to changes of the coin position, rotation and scale. As a first step, scale- and translation- invariance is achieved by using the automatic segmentation [104] of ancient coins.

Rotation-invariance is achieved by splitting the image space into circular sub-regions [4] and then modeling the rotation-invariant geometric relationships of identical local features in these sub-regions. Each step of the methodology depicted in Figure 5.2 is explained in this section.

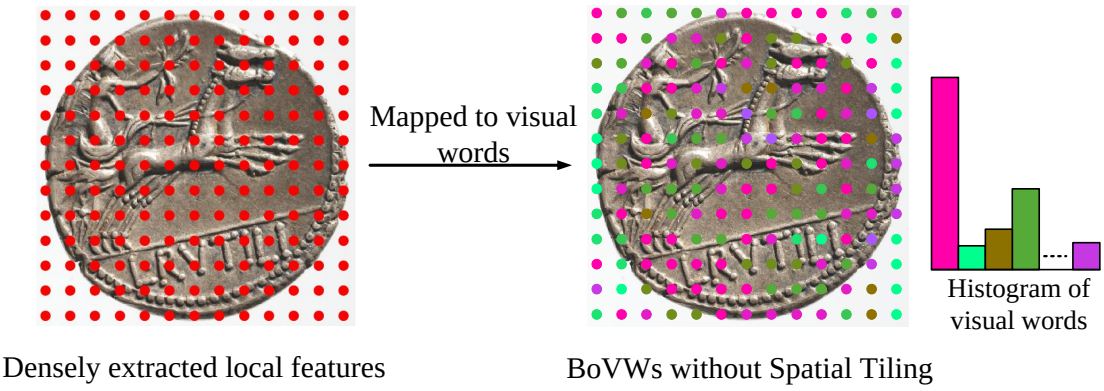


Figure 5.3: A given image is represented as a histogram of visual words

1. **Automatic coin image segmentation:** The automatic coin image segmentation proposed by Zambanini and Kampel [104] is used to achieve invariance to scale and translation. Once the coin image is segmented, it is cropped and normalized to a specified standard size of 480×480 to achieve the required invariance to object location and scale.

2. **Image representation using the BoVWs model:** Once the coin images are automatically segmented from the background, local rotation-invariant features such as SIFT [55] are densely extracted from them to achieve invariance to rotations locally. These features are then mapped to a visual vocabulary to represent the image as the histogram of visual words. Let the size of visual vocabulary $voc = \{v_1, v_2, v_3, \ldots, v_M\}$ be M. A given image is first represented as a set of descriptors

$$I = \{d_1, d_2, d_3, ..., d_N\} \tag{5.1}$$

where N is the total number of descriptors. A given descriptor d_k is then mapped to a visual word; v_i, using some similarity measures like the Euclidean distance as follows:

$$v\left(d_k\right) = \underset{v \in voc}{\arg\min}\ \mathrm{Dist}\left(v, d_k\right) \tag{5.2}$$

where d_k is the k^{th} descriptor in the image and $v\left(d_k\right)$ is the visual word assigned to this descriptor based on the distance $\mathrm{Dist}\left(v, d_k\right)$. In dense sampling based BoVWs, descriptors are collected from images using a grid with a specific pixels stride. Each descriptor is then assigned to a visual word from the vocabulary. An image is represented by the histogram of visual words where the number of bins of this histogram is equal to the size of the visual vocabulary M. The value of the bin b_i of this histogram gives the number of occurrences of a visual word v_i in an image. Figure 5.3 illustrates the whole process where local features are densely extracted from a given image on a regular grid. These local features are then mapped to the visual words of the visual vocabulary. Finally, the image is represented as the histogram of visual words.

3. **Imposing circular tiling over the BoVWs image representation:** The circular tiling proposed in Chapter 3 is imposed over the BoVWs image representation. In the circular tiling scheme, concentric circular sub-regions are imposed over the BoVWs image representation as shown in Figure 5.4. For a vocabulary size of M, a histogram of M visual words is generated for each circular tiling. These histograms are then concatenated in a single feature vector of length $M \cdot r$.

4. **Modeling rotation-invariant geometric relationships of visual words in circular tiling:** In each partition or tiling of the circular tiling scheme, the geometric relationships of identical visual words are modeled. These geometric relationships are established using the spatial locations of the identical visual words. The geometric relationships are achieved using the idea of Khan et al. [42]. They use the angles made by Pair-wise Identical Visual Words (*PIWs*) for the construction of the Pair-wise Identical Visual Words Angles Histogram (*PIWAH*) to represent the image. *PIWAH* is not invariant to image rotations because it is made from angles that are calculated with respect to x-axis. Inspired by the idea of Tao and Grosky [84], three identical words are considered to calculate the angles and denote them as *Triplets of Identical Visual Words* (*TIWs*) as shown in Figure 5.5. This will achieve a rotation-invariant triangular relationship among the words of a given triplet. Based on the angles calculated among members of each *TIWs*, the angles histogram is constructed in a similar manner as proposed by Khan et al. [42] and is denoted as *Triplets of Identical Visual Words Angles Histogram* (*TIWAH*).

   Let D_i be the set of all descriptors mapped to a visual word v_i, then the i^{th} bin of the histogram of visual words b_i, is the cardinality of the set D_i.

$$b_i = Card\,(D_i) \text{ where } D_i = \{d_k, k \in [1, \ldots, N] \mid v\,(d_k) = v_i\} \tag{5.3}$$

   From set D_i, all the distinct pairs of three descriptors are considered to calculate angles between the spatial positions of the descriptors as shown in Figure 5.5. As mentioned in Chapter 4 that the triangulation done for all distinct pairs of three descriptors is called *combinatorial triangulation*. The spatial position of a descriptor is given by its position

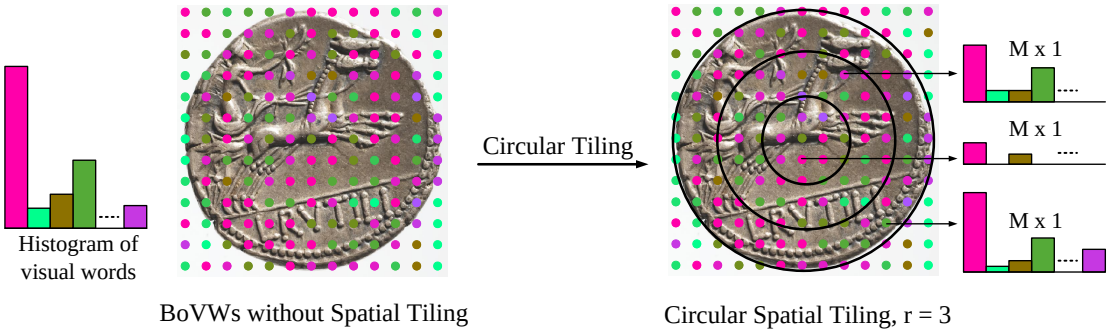


Figure 5.4: BoVWs image representation with and without circular tiling

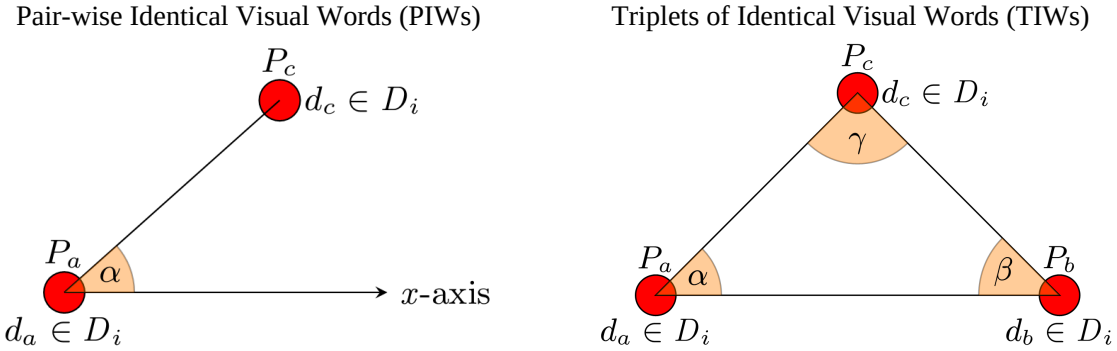


Figure 5.5: *PIWs* and *TIWs* for the descriptors d_a, d_b and d_c at image positions P_a, P_b and P_c respectively. All descriptors belong to the set of same visual words D_i.

on the dense sampling grid. The set of all *TIWs* related to a visual word v_i is defined as:

$$TIW_i = \left\{(P_a, P_b, P_c) \mid (d_a, d_b, d_c) \in D_i^3, d_a \neq d_b \neq d_c\right\} \tag{5.4}$$

where P_a, P_b and P_c are the spatial positions of the descriptors d_a, d_b and d_c respectively. The value of the i^{th} bin of the histogram shows the frequency of the visual word v_i. Therefore in case of *combinatorial triangulation*, the cardinality of TIW$_i$ is ${}^{b_i}C_3$ which is the number of all possible pairs of three distinct elements among the elements of D_i. The positions of the elements of each pair make a triangle. Calculating angles for such a huge number of triangles is time consuming. Delaunay triangulation is an efficient triangulation method from computational geometry. It has two main principles. First, it rejects the collinear points as they do not make a triangle. Second, the circumscribed circle defined by the three points should not contain any other point. These two principles significantly reduce the number of triangles. As an example, the triangulation among identical visual words using both the triangulation methods is depicted in Figure 5.6. It can be observed that the combinatorial triangulation produces a huge number of triangles while the Delaunay triangulation produces much smaller number of triangles. The angles of all the triangles are calculated using the law of cosines. The angles histogram is built from these angles for which the bins are empirically chosen between $0°$ and $180°$. The angles histogram for a specific word v_i is named as $TIWAH_i$. The i^{th} bin of the histogram of visual words associated with visual word v_i is replaced with $TIWAH_i$ in such a way that the spatial information is added without altering the frequency information of v_i. Finally $TIWAH_i$ of all the visual words are combined to represent a given image.

$$\begin{aligned} TIWAH &= (\psi_1 TIWAH_1, \psi_2 TIWAH_2, \ldots, \psi_M TIWAH_M) \\ \text{where} \quad \psi_i &= \frac{b_i}{\left\|TIWAH_i\right\|} \end{aligned} \tag{5.5}$$

where ψ_i is the normalization coefficient. For a visual vocabulary of size M, if the number of bins in angles histogram is θ, then the size of the *TIWAH* is $M\theta$. In each parti-

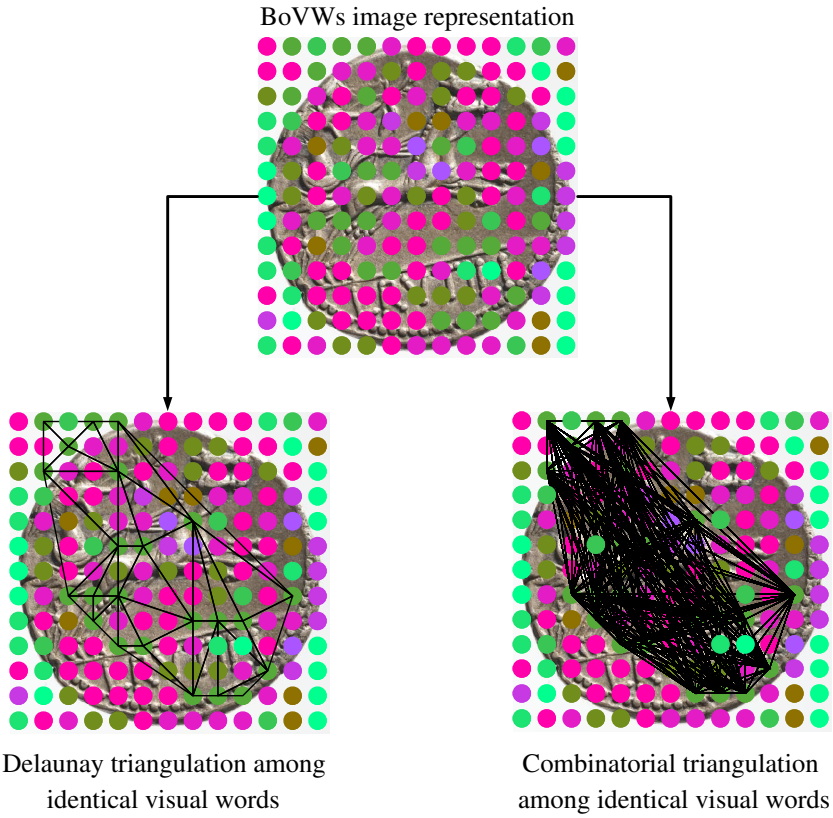


Figure 5.6: A comparison of the number of triangles produced by the Delaunay and combinatorial triangulation methods

tion or tiling of the circular tiling scheme, the geometric relationships of identical visual words are modeled using *TIWAH*. Figure 5.7 shows the combinatorial triangulation and Figure 5.8 shows the Delaunay triangulation among the instances of a given visual word. Now the triplets of identical visual words (TIW) are considered in each partition of the circular tiling. Due to this reason the number of triangles in case of combinatorial triangulation drops significantly while in case of Delaunay triangulation also reduces. Such modeling of the geometric relationship of identical visual words not only extracts more information from the circular tilings but also reduces the number of unique combinations of *TIWs*. Finally, the histogram of visual words and *TIWAH* of all the circular tilings are concatenated as described in Eq. 5.6. We call the final histogram *TIWAHCR*, which contains the information of *TIWs* and visual words other than *TIWs* for a given circular tiling as shown in Figure 5.9.

$$TIWAHCR = \left[TIWAH^1, TIWAH^2, \ldots, TIWAH^r, CR^1, CR^2, \ldots, CR^r\right] \tag{5.6}$$

Where for r number of circular tilings, $rPIWAH_j$ represents the angles histogram and CR_j represents the histogram of visual words for the j^{th} circular tiling. Therefore for a

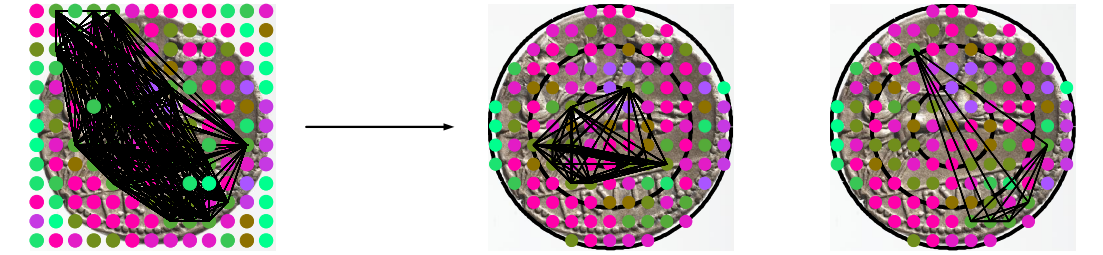


Figure 5.7: Combinatorial triangulation in circular tilings

vocabulary size of M, number of circular tilings r and number of bins in angles histogram θ, the length of *rPIWAHCR* is $(r \cdot M \cdot \theta) + (r \cdot M)$.

5. **Classification:** The images are represented as TIWAHCR histograms. These histogram act as feature vectors to train a multi-class support vector machine (SVM) in the one-vs-all mode. The Helinger kernel is used with SVM. While the additive kernels yield better results with SVM, they take more time for training than the linear SVM [39]. Vedaldi and Zisseman [92] propose to map the feature vectors explicitly using the approximation of additive kernels which are then used with linear SVM. Such approximations are not only efficient in terms of time but also are as accurate as the original kernels. The best value of the regularization parameter 'C' is calculated using the n-fold cross-validation.

5.2 Dataset

The dataset used for experiments belongs to the largest and the most diverse among the works that deal with image-based analysis of ancient coins [5, 6, 9, 10, 35, 38, 103, 105, 107]. It consists of 2224 images that belong to 29 different reverse motifs. Only the two recent works [43, 44]

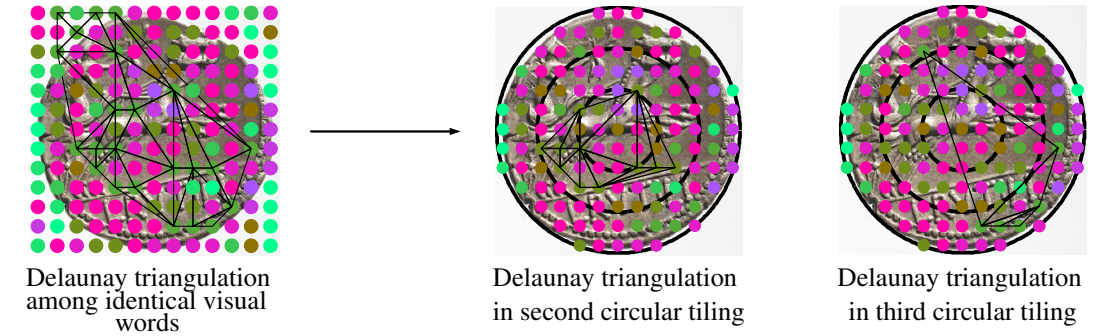


Figure 5.8: Delaunay triangulation in circular tilings

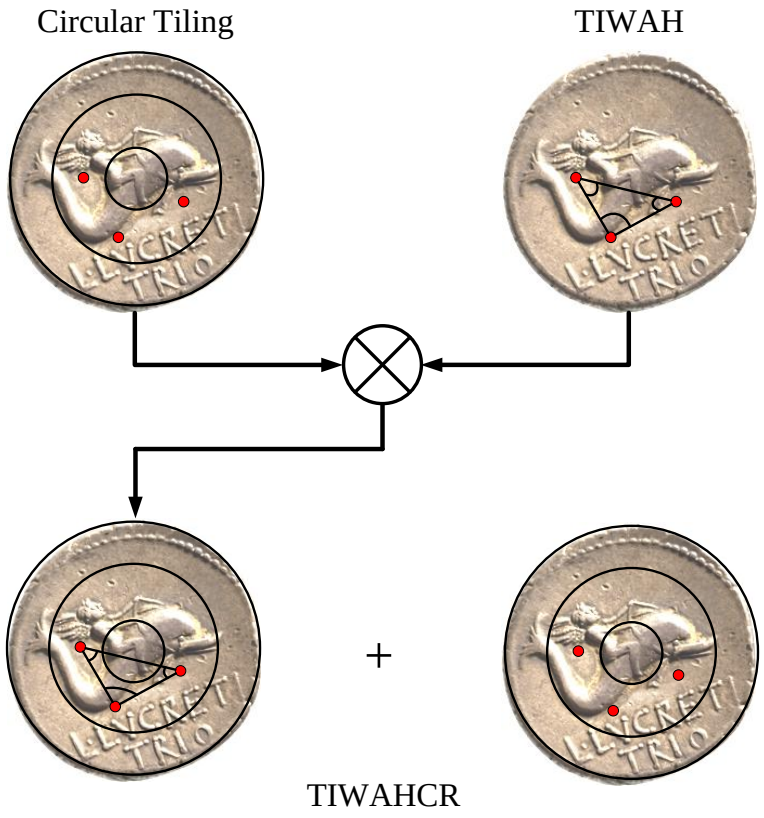


Figure 5.9: Circular tiling, *TIWAH* and *TIWAHCR*

use more images but with only 15 various obverse motifs. The images in the dataset have been collected from the following three sources.

1. **The Vienna Museum of Fine Arts (Kunsthistorishes Museum Wien)**: The museum has 6000 Roman Republican coins from which the images of 4200 coins were collected for the ILAC project [37]. In the dataset, 1014 coin images are from this museum.

2. **British Museum London**: The Department of Coins and Medals at the British Museum is home to one of the largest collections of ancient coins in the world. 857 of the images in the dataset are from the British Museum.

3. **acsearch**[1]: acsearch is an online auction website for ancient coins. It is also one of the largest online sources for images of the ancient coins. The dataset contains 353 coin images from this website.

The difference among the images of all the three sources can be observed from the sample images shown in Figure 5.10. At a first glance it can be observed that they differ from one

[1] http://www.acsearch.info (last accessed Oct 30, 2014)

Figure 5.10: Sample coin images from three sources i.e. The Vienna Museum of Fine Arts, The British Museum London and acsearch.info

another due to the imaging conditions. The images from the Vienna Museum are taken in a controlled environment due to which the changes in illumination are minimal as compared to the other two resources. However, they differ from one another due to rotations. Aberrations and missing coin parts are common among all the images. Another problem is that of the dirt appearing on the coins due to the preserving conditions. It is more pronounced on the coins of British Museum and acsearch. The coins of the Vienna Museum are comparatively cleaner than the other two sources. Due to the presence of dirt, the coin images of the same class vary from one another thus affecting the task of classification.

Table 5.1: Average time in seconds taken by each triangulation scheme for 29 images

Combinatorial triangulation	2860
CombinatorialF triangulation	41.6
Delaunay triangulation	**6.4**

5.3 Experiments and results

The experiments are performed in a sequential manner. Since the triangulation is an integral part of the image representation, as a first step, experiments are performed to select the triangulation method that not only performs better but is also efficient in terms of time. These experiments are performed in Section 5.3.1. The size of vocabulary is a basic parameter of the BoVWs model [68] and needs to be optimized for any give dataset [31]. Since the proposed image representation is an extension of the BoVWs model, experiments for the size of vocabulary are performed for each triangulation method in Section 5.3.2. The proposed method performs the triangulation of identical visual words in each partition of the circular tiling. Experiments are performed in Section 5.3.3 to optimize for the number of partitions in the circular tiling. Finally, the proposed image representation is evaluated for robustness to image rotations in Section 5.3.4 against the standard BoVWs model and the simple circular tiling. For the experiments, the dataset is divided into two disjoint training and test sets. The size of the training set is 1426 while the test set consists of the remaining 798 images.

5.3.1 Computational complexity of the triangulation methods

Since the triangulation is a part of the image representation, the desired triangulation method is the one that is efficient in terms of computation time. Both the combinatorial and the Delaunay triangulation methods are compared in terms of computational time. 29 images from the whole dataset are collected by selecting one image from each class at random. Images are of standard size which is 480×480. The *TIWAH* representation of these images is constructed using both the triangulation methods implemented in Matlab. A a faster 'C' implementation of the combinatorial triangulation is also used which is denoted by *combinatorialF* triangulation. The experiments are repeated 10 times on a single core and the mean calculation time taken by each method is reported in Table 5.1. It can be concluded from these results that the Delaunay triangulation is even faster than the *combinatorialF* triangulation.

5.3.2 Size of vocabulary

The size of vocabulary is an important parameter in the BoVWs model [68]. Optimization is done for the size of vocabulary on the current dataset by empirically selecting the values from the set $\{10, 50, 100, 150, 200, 400, 800\}$. For each value of the vocabulary size, classification runs are performed 10 times and the mean performances of the *TIWAH* constructed with both the types of triangulation methods are shown in Figure 5.11. *TIWAH* constructed with Delaunay triangulation is not only more efficient but also performs better than the one constructed with

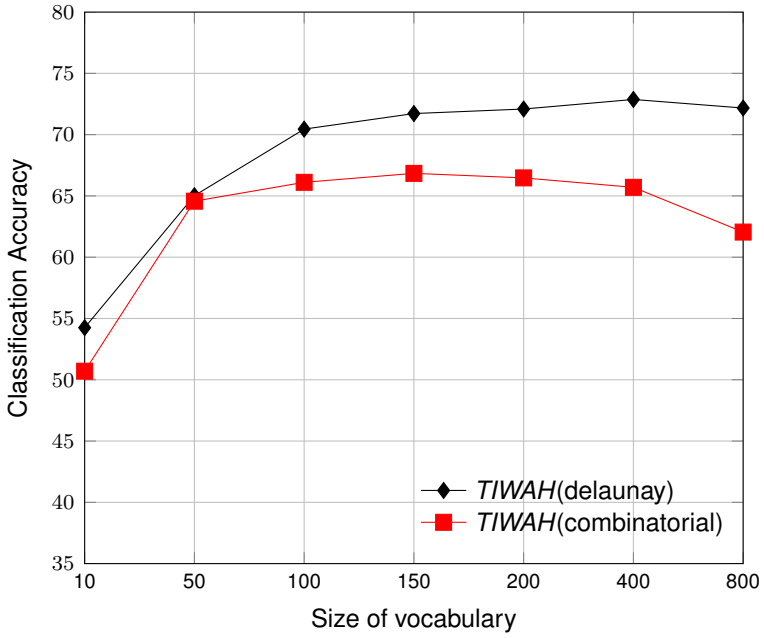


Figure 5.11: Performances of various methods with respect to the size of vocabulary

combinatorial triangulation. It can also observed that no significant performance improvement is gained from the vocabulary sizes larger than 150.

5.3.3 Number of tilings

Optimization is also done for the number of tilings in both the simple circular tiling and *TIWAHCR*. The values for the number of tilings are empirically selected from $\{2, 3, 4, 5\}$. Experiments are performed 20 times and the mean performances achieved by both the settings on all the empirically selected values are shown in Figure 5.12. The performance of *TIWAHCR* is superior to that of the simple circular tiling because it contains additional spatial information of the visual words in each tiling. On four tilings, *TIWAHCR* achieves the maximum performance. However with the increase in the number of tilings, its performance tends to converge towards that of the simple circular tiling. This is due to the fact that the width of the tilings decrease with the increase in their number thus decreasing the occurrences of identical words in any given tiling.

5.3.4 Rotation-invariance evaluation

In order to evaluate the proposed method for robustness to image rotations, we generate synthetically rotated coin images. The rotated coin images are produced by making one of them as reference image and then rotating this reference image in $90°$ steps. An example is shown in Figure 5.13.The synthetically rotated images allow to test stronger rotation differences as those already present in the dataset (generally lower than $90°$). Experiments are repeated 20 times

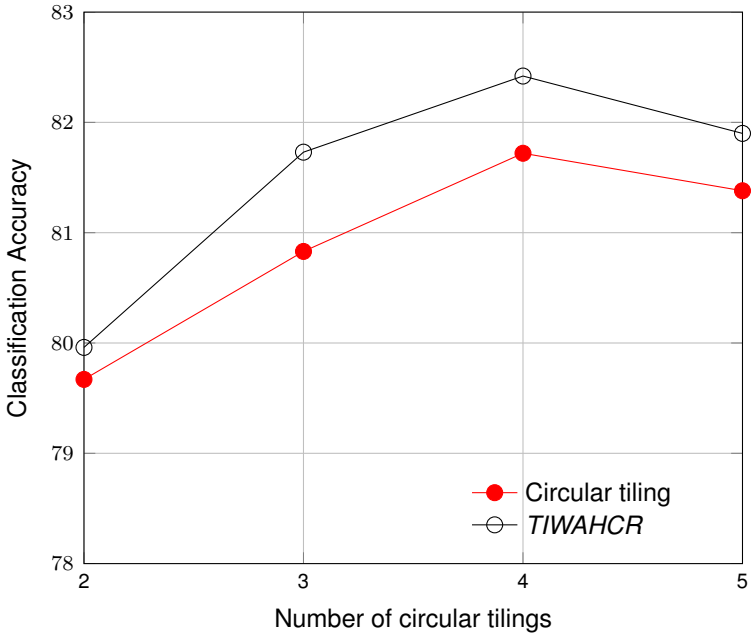


Figure 5.12: Performances of *TIWAHCR* and circular tiling with respect to the number of tilings

and the mean performances of the BoVWs, simple circular tiling and *TIWAHCR* are shown in Figure 5.14. The number of tilings in the circular tiling and *TIWAHCR* is 3. The size of vocabulary is 150 and at each iteration a new vocabulary is constructed. Local rotation-invariance is achieved by using rotation-invariant local features such as SIFT. The experimental results reflect the theoretical foundations of the proposed rotation-invariant image representation and show that rotations have no significant influence on the classification performance. The proposed method clearly outperforms the BoVWs model on rotated images while marginally performs better than the circular tiling. To summarize, calculating and combining *TIWAH* for each circular tiling achieves global spatial relationships of identical visual words in a rotation-invariant

Figure 5.13: Rotated coin images

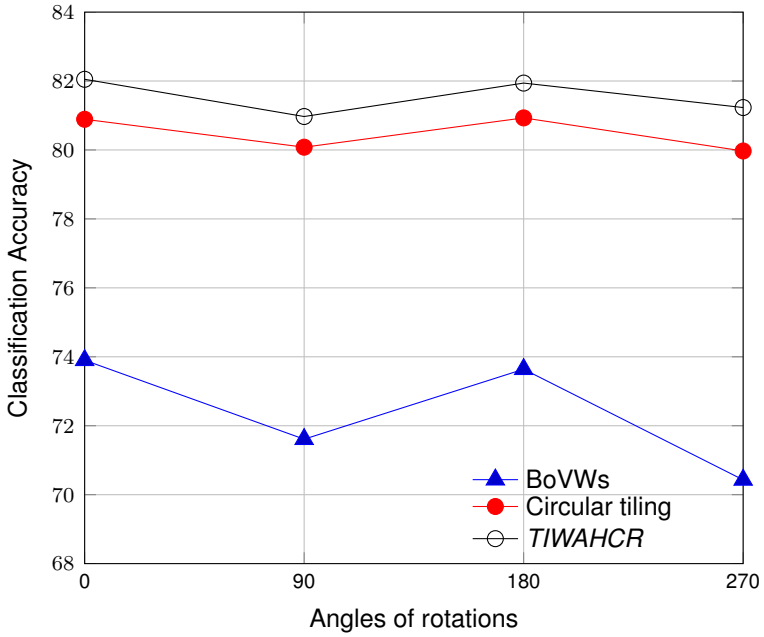


Figure 5.14: Performances of various methods on rotated images

manner. This leads to an increase in classification rate of ancient coins in the presence of severe rotations.

5.4 Summary

An image representation robust to changes in scale, translation and rotation is presented for ancient coin classification. As a first step, the coin image is automatically segmented from the homogeneous background. The segmented image is then cropped and normalized to achieve invariance to changes in scale and translation. Local rotation-invariant features are then densely extracted from the segmented image thus achieving rotation-invariance locally. These local features are then mapped to the visual words of a visual vocabulary thus representing the image as the Bag of visual words (BoVWs). The circular tiling scheme is then imposed over the BoVWs representation of the image. In circular tiling scheme, the image space is divided into a number of concentric circles. Afterward, the geometric relationships of the instances of identical visual words in each partition of the circular tiling scheme are modeled. These geometric relationships are achieved by establishing the rotation-invariant triangular relationships among the triplets of identical visual words. Two kinds of triangulation methods are used. In combinatorial triangulation, all the possible three combinations of the instances of visual words are considered. In the Delaunay triangulation, the number of triangles are far less than the combinatorial triangulation due to its basic principles. The angles of all the triangles are then calculated using the law of cosines. The angles histograms for each partition of the circular tiling scheme is built from these angles. In addition to the angles histogram, the histogram of visual words for each partition of

the circular tiling scheme is also calculated. The angles histogram is then concatenated with the histogram of visual words. Following are the main findings of the experiments:

1. The Delaunay triangulation is much efficient than the combinatorial triangulation in terms of run time.

2. The Delaunay triangulation also outperforms the combinatorial triangulation on the empirically defined sizes of vocabulary.

3. The combination of the angles histogram with circular tiling scheme not only outperforms the simple BoVWs model but is also more robust to image rotations.

Currently, the proposed image representation can only be used for automatically segmentable circular object like the ancient coins. It can be extended to other problems where the background is homogeneous and the object of interest can easily be segmented. In Chapter 3, the idea of the adapted-circular tiling is presented where the feature extraction grid is adapted to the circular tilings. Experimental results showed that the adapted-circular tiling performed better than the circular tiling. The proposed image representation can be extended to adapted-circular tiling in order to achieve better classification rates. Finally, the local features specifically designed for non-textured or almost flat objects can be used with the underlying BoVWs model. In this regard, the LIDRIC descriptor [108] is shown to outperform the most prevalent local descriptors such as SIFT.

CHAPTER 6

Conclusion

The conclusion of the thesis is given in this chapter by summarizing its contributions and implications. In addition to that, the shortcomings of the proposed methodologies in their current form are outlined in Section 6.1. The future work of the current work and other possible application areas where the proposed methodologies can be used are given in Section 6.2.

The most widely used paradigm for object category recognition-based image classification learns object models by training a classifier with the help of training images whose number varies from hundreds to thousands per object class [49]. The basic motivation behind the use of such a huge number of training images is to account for image variations that occur due to several factors such as changes in scale, translation and viewpoint. However, even with such an amount of training images, certain variations such as in-plane rotations can not be handled. The so-called local image descriptors have been developed and they have shown remarkable success by achieving invariance to variations like changes in scale, rotations and translation. However, these local descriptors achieve such invariance locally i.e. they are calculated on smaller image patches. In image classification problems such as object or scenes, the global geometry is more important as objects and scenes have specific geometric structures. Consequently, in this thesis, image representations are proposed that capture the global geometry in the shape and texture of the objects. Since, these features almost remain unchanged under several image variations such as changes in object scale, object position or in-plane rotations, the image representations based on these features will be invariant to such image variations leading to the invariant image classification. In the context of shape, the problem of image-based classification of ancient coins is considered as the objects minted on these coins can only be identified from their contours rather than the texture. On the other hand, the problem of image-based classification of butterflies is considered as any given specie of a butterfly can be identified from the patterns of their wings. Based on these properties of the objects i.e. shape in case of ancient coins and texture in case of butterflies, image representations are proposed that are invariant to changes in scale, translations and image rotations.

The method presented in Chapter 3 achieves image representation that is invariant to changes in scale, translation and rotations. The image-based classification of ancient coins is taken as a

motivating example. The ancient coins are imaged on homogeneous background and thus can be segmented automatically. The automatic coin segmentation proposed by [104] is used as this method is specifically proposed for ancient coins. Once the segmentation is performed, the image is cropped and normalized to achieve invariance to scale and translation. The segmented image is then divided into circular regions from which the local features are extracted thus achieving invariance to image rotations. The local features extracted from each circular region are then mapped to a visual vocabulary to build a histogram of visual words for each region. Finally, the histograms from all the circular regions are concatenated to represent the image. The results of the experiments performed on 1900 images of 21 different reverse motifs of the ancient coins indicated that the proposed image representation is invariant to changes in scale, translation and image rotations.

The purpose of the research presented in Chapter 4 is to obtain an image representation which is invariant to changes in scale, translation and image rotations by modeling the geometric relationships of identical texture patterns present on the objects. The image-based classification of butterflies is taken as an application where the insects have identical patterns on their wings. However, unlike ancient coins, they can not be automatically segmented due to the background clutter in the images caused by trees, flowers and branches etc. Due to this reason, for the construction of visual vocabulary, the segmentation masks are used to extract local features from the foreground thus making the visual vocabulary more discriminating. For image representation, local features are extracted from a given image and mapped to the visual vocabulary. Afterward, triangular relationships of identical features are modelled to represent the image. Since the image representation is based on triangular relationships, it is invariant to changes in scale, translation and image rotations and is also validated by the experimental results.

Finally, the ideas of both Chapter 3 and Chapter 4 are combined in Chapter 5 to propose a novel image representation which is invariant to changes in scale, translation and image rotations. The problem of reverse motif recognition-based ancient coin classification is used as a motivating example. However, it can be extended to other problems where the object of interest is imaged with homogeneous background and thus is automatically segmentable. The proposed image representation follows the method of Chapter 3 i.e. the automatic segmentation of the coin image and imposing the circular tiling on the segmented image. However, in each circular tiling, the triangular geometric relationships of identical local features are modelled to build the final image representation which is invariant to changes in scale, translation and image rotations. In addition, the experimental results demonstrated that the proposed method outperform the simple circular tiling on 2224 coin images of 29 different classes.

On the application side, the research presented in the thesis proposes a holistic system to classify ancient Roman Republican coins based on their reverse side motifs. The image classification is based on image representation that is enriched with spatial information to increase its discriminative power. This is achieved by combining a spatial pooling scheme with co-occurrence encoding of the local features. The research addresses the required geometric invariance properties of image-based ancient coin classification, as coins from different collections can be located at differing image locations, have various scales in the images and can undergo various in-plane rotations. Consequently, a given coin image is first automatically segmented and then normalized to achieve invariance to scale and translation. As a second step, the circular

tiling is imposed over the segmented image and local features from each circular region are collected and mapped to visual vocabulary. Then, in each circular tiling, the triangular geometric relationships of identical visual words are modeled to construct the final image representation.

Based on experimental results the system shows better performance than the bag-of-visual-words model which is the most prevalent paradigm used for image classification while still being invariant to the mentioned geometric transformations. For 29 motifs, the system achieves a classification rate of 82%. It is considered to act as a helpful tool for numismatists in the near future which facilitates and supports the traditional coin classification process by a faster presorting of coins.

6.1 Limitations

The proposed methods contribute towards the field of computer vision by introducing image representations that are robust to image variations caused by changes in scale, translation and image rotations. However, there are certain limitations which are outlined in the following:

Circular tiling

- The circular tiling is used to capture the geometry of the shape of the underlying object in a rotation-invariant manner. However, the circular tiling is only imposed once the coin image is automatically segmented from the background. As mentioned earlier, the coins are imaged on homogeneous background due to which they and can be segmented automatically. The circular tiling is not feasible in cases where the object of interest can not be segmented automatically.
- The circular tiling can also be used without automatic segmentation but then the object of interest has to be in the center of the image.

Geometric relationship of identical features

- The proposed method in Chapter 4 is based on the geometry between identical visual words based on which the image is represented in a scale- and rotation-invariant manner. However, such representation can be affected by occlusions. For instance, in case of butterflies, the geometry is achieved by using the spatial relationships of identical patterns on both the wings of the butterflies. If one of the wings is completely occluded due to some reason, then the representation will suffer.
- The notion of similarity is used and any two given local features are considered identical when they are assigned to the same visual word. However, the process of construction of vocabulary is performed with the k-means clustering where the same cluster can contain otherwise non-similar features. This negatively affects the discriminating nature of the visual vocabulary and the resulting image representation. The proposed method uses manually generated segmentation masks for the construction of the vocabulary to use local features from the foreground. However, in case of larger dataset, manual generation of the segmentation masks is not feasible.

Ancient Coin Classification System

- The ancient coin classification system proposed in the thesis uses images of the coins. It does not consider other information such as weight and size of the ancient coins. Due to this reason, it fails at places where the image data is not sufficient for instance due to very strong abrasions on the coin. Images of the forged coins have high similarity with the images of the original coins due to which forgery can not be detected by the proposed system as it solely depends on the image data.
- As mentioned before, the proposed system performs coarse-grained classification of the ancient coins. This means that several coin types share the same reverse motifs with minor changes such as in symbols or legends. Due to this reason, the classification that is solely based on the reverse motif is coarse-grained which should be refined by other classification cues such as legend recognition or dense feature matching [106].

6.2 Future work

Geometric relationship of identical features

The notion of 'similar' or 'identical' features needs to be refined. In the proposed method, features are considered identical only when they are assigned identical words from the visual vocabulary. However, for the construction of visual vocabulary, k-means is used which is hard clustering [39]. The features are assigned to a single cluster based on a similarity measure such as the Euclidean distance. As the distance of a given feature increases from the cluster center, its similarity with the centers decreases. This can affect the discriminating nature of the visual vocabulary in a negative manner. Soft clustering has been shown to perform better than the hard clustering for the image classification problems based on BoVWs where features are assigned to various clusters in a weighted manner [72]. Thus a possible future direction would be to use soft clustering for the construction of the visual vocabulary.

The segmentation masks are used to collect features from the foreground for vocabulary construction to increase its discriminating power. These masks are manually generated as there is no state the art segmentation algorithm that can effectively segment the butterfly images amidst the cluttered background. Manual segmentation is a tedious process that require adequate effort. Thus, in order to increase the discriminating nature of the visual vocabulary, specialized method such as the one proposed in [61] can be used.

Ancient coin classification system

Given the fundamental task of the project [37], a rate of 80% accurate results proves that it is possible to gain a real benefit from the developed algorithms. Of course, to push this rate to 98 or 99%, a multiple amount of reference images and probably more 'fine-tuning' (e.g., their proper selection based on the degree of preservation) would be required.

So ultimately, this approach is of the highest interest, both in respect of future automated coin classification and by considering all other uses of coin matching, be that monitoring

of (illegal) trade, retrieving scientific data from otherwise unspecified archives of images or counting known specimens per type. Potentially, the same procedure can be extended to count coin-dies, which would lead to answering the question of how many coins were produced, which is an often discussed issue of historic economies.

Bibliography

[1] Ankur Agarwal and Bill Triggs. Hyperfeatures - multilevel local coding for visual recognition. In *Proceedings of the 10th European Conference on Computer Vision (ECCV)*, volume 1, pages 30–43, 2006.

[2] Shivani Agarwal, Aatif Awan, and Dan Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(11):1475–1490, 2004.

[3] Timo Ahonen, Jiří Matas, Chu He, and Matti Pietikäinen. Rotation invariant image description with local binary pattern histogram fourier features. In *Scandinavian conference on image analysis (SCIA)*, pages 61–70, 2009.

[4] Hafeez Anwar, Sebastian Zambanini, and Martin Kampel. Supporting ancient coin classification by image-based reverse side symbol recognition. In *The International Conference on Computer Analysis on Images and Patterns (CAIP) (2)*, pages 17–25, 2013.

[5] Hafeez Anwar, Sebastian Zambanini, and Martin Kampel. Encoding spatial arrangements of visual words for rotation-invariant image classification. In *German Conference on Pattern Recognition (GCPR)*, pages 407–416, 2014.

[6] Hafeez Anwar, Sebastian Zambanini, and Martin Kampel. A rotation-invariant bag of visual words model for symbol-based ancient coin classification. In *The International Conference on Image Processing (ICIP)*, pages 5257–5261, 2014.

[7] Hafeez Anwar, Sebastian Zambanini, and Martin Kampel. Coarse-grained ancient coin classification using image-based reverse side motif recognition. *Machine Vision and Applications*, 26(2-3):295–304, 2015.

[8] Hafeez Anwar, Sebastian Zambanini, Martin Kampel, and Klaus Vondrovec. Ancient coin classification using reverse motif recognition: Image-based classification of roman republican coins. *IEEE Signal Processing Magazine*, 32(4):64–74, 2015.

[9] O. Arandjelović. Automatic attribution of ancient Roman imperial coins. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1728–1734, 2010.

[10] O. Arandjelović. Reading ancient coins: Automatically identifying denarii using obverse legend seeded retrieval. In *European Conference on Computer Vision (ECCV)*, pages 317–330, 2012.

[11] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002.

[12] Mark de Berg, Otfried Cheong, Marc van Kreveld, and Mark Overmars. *Computational Geometry: Algorithms and Applications*. Springer-Verlag TELOS, 2008.

[13] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011.

[14] Michael H. Crawford. *Roman Republican Coinage, 2 vols*. Cambridge University Press, 1974.

[15] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[16] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2Nd Edition)*. Wiley-Interscience, 2000.

[17] Richard Duncan-Jones. *Money and Government in the Roman Empire*. Cambridge, 1994.

[18] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes VOC Challenge. *Int. J. Comput. Vision*, 88(2):303–338, 2010.

[19] Bin Fan, Fuchao Wu, and Zhanyi Hu. Rotationally invariant descriptors using intensity order pooling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):2031–2045, 2012.

[20] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007.

[21] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[22] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from googleß image search. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1816–1823, 2005.

[23] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 264–271, 2003.

[24] Bob Fisher, Toby P. Breckon, Kenneth Dawson-Howe, Andrew Fitzgibbon, Craig Roberston, E Trucco, and Christopher Williams. *Dictionary of Computer Vision and Image Processing, 2nd Edition*. Wiley, 2 edition, 2013.

[25] Wei Gao and Haizhou Ai. Face gender classification on consumer images in a multiethnic environment. In *The International Conference on Advances in Biometrics (ICB)*, pages 169–178, 2009.

[26] Kristen Grauman and Trevor Darrell. Efficient image matching with distributions of local invariant features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 627–634, 2005.

[27] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.*, 8:725–760, 2007.

[28] Kristen Grauman and Bastian Leibe. *Visual Object Recognition*. Morgan and Claypool Publishers, 2010.

[29] Philip Grierson. *Numismatics*. Oxford University Press, 1975.

[30] Tal Hassner, Viki Mayzels, and Lihi Zelnik-Manor. On sifts and their scales. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, pages 1522–1528, 2012.

[31] Jian Hou, Jianxin Kang, and Naiming Qi. On vocabulary size in bag-of-visual-words representation. In *Proceedings of the 11th Pacific Rim Conference on Advances in Multimedia Information Processing: Part I*, pages 414–424, 2010.

[32] Thorsten Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)*, pages 137–142, 1998.

[33] Thorsten Joachims. Training linear svms in linear time. In *Proceedings of the ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226, 2006.

[34] Frederic Jurie and Bill Triggs. Creating efficient codebooks for visual recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 604–610, 2005.

[35] M. Kampel and M. Zaharieva. Recognizing ancient coins based on local features. In *International Symposium on Visual Computing*, pages 11–22, 2008.

[36] A. Kavelar, S. Zambanini, and M. Kampel. Word detection applied to images of ancient roman coins. In *The International Conference on Virtual Systems and Multimedia (VSMM)*, pages 577–580, 2012.

[37] A. Kavelar, S. Zambanini, M. Kampel, K. Vondrovec, and K. Siegl. The ILAC-project: Supporting ancient coin classification by means of image analysis. *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XL-5/W2:373–378, 2013.

[38] Albert Kavelar, Sebastian Zambanini, and Martin Kampel. Reading the legends of roman republican coins. *J. Comput. Cult. Herit.*, 7(1):5:1–5:20, 2014.

[39] Andrea Vedaldi Ken Chatfield, Victor Lempitsky and Andrew Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proceedings of the British Machine Vision Conference*, pages 76.1–76.12, 2011.

[40] R. Khan, C. Barat, D. Muselet, and C. Ducottet. Spatial orientation of visual word pairs to improve bag-of-visual-words model. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11, 2012.

[41] R. Khan, C. Barat, D. Muselet, and C. Ducottet. Spatial orientation of visual word pairs to improve bag-of-visual-words model. *Comput. Vis. Image Underst.*, 2015.

[42] Rahat Khan, Cecile Barat, Damien Muselet, and CHristophe Ducottet. Spatial orientation of visual word pairs to improve bag-of-visual-words model. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11, 2012.

[43] Jongpil Kim and Vladimir Pavlovic. Ancient coin recognition based on spatial coding. In *The International Conference on Pattern Recognition (ICPR). in print*, 2014.

[44] Jongpil Kim and Vladimir Pavlovic. Improving ancient roman coin recognition with alignment and spatial encoding. In *ECCV Workshop on Computer Vision for Art Analysis*, 2014.

[45] I. Kokkinos and A. Yuille. Scale Invariance without Scale Selection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, 2008.

[46] Josip Krapac, Jakob J. Verbeek, and Frédéric Jurie. Modeling spatial layout with fisher vectors for image categorization. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1487–1494, 2011.

[47] Neeraj Kumar, Alexander C. Berg, Peter N. Belhumeur, and Shree K. Nayar. Describable visual attributes for face verification and image search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(10):1962–1977, October 2011.

[48] Young H. Kwon and Niels da Vitoria Lobo. Age classification from facial images. *Comput. Vis. Image Underst.*, 74(1):1–21, 1999.

[49] CH. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 951–958, 2009.

[50] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2169–2178, 2006.

[51] Ales Leonardis and Horst Bischof. Dealing with occlusions in the eigenspace approach. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 453–458, 1996.

[52] Ales Leonardis and Horst Bischof. Robust recognition using eigenimages. *Comput. Vis. Image Underst.*, 78(1):99–118, 2000.

[53] Haibin Ling and Stefano Soatto. Proximity distribution kernels for geometric context in category recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[54] David Liu, Gang Hua, Paul Viola, and Tsuhan Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[55] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60:91–110, 2004.

[56] Michael J. Lyons, Julien Budynek, and Shigeru Akamatsu. Automatic classification of single facial images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(12):1357–1362, 1999.

[57] Jiri Matas and Stepán Obdrzálek. Object recognition methods based on transformation covariant features. In *European Signal Processing Conference (EUSPIC0)*, pages 1721–1728, 2004.

[58] Sancho McCann and David G Lowe. Spatially local coding for object recognition. In *Asian Conference on Computer Vision (ACCV)*, pages 204–217, 2012.

[59] G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley Series in Probability and Statistics, 2000.

[60] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005.

[61] Andrej Mikulík, Michal Perdoch, Ondrej Chum, and Jiri Matas. Learning a fine vocabulary. In *Proceedings of the 11th European Conference on Computer Vision (ECCV), Part III*, pages 1–14, 2010.

[62] Andrej Mikulík, Michal Perdoch, Ondrej Chum, and Jiri Matas. Learning vocabularies over a fine quantization. *Int. J. Comput. Vision*, 103(1):163–175, 2013.

[63] Madhav Moganti, Fikret Ercal, Cihan H. Dagli, and Shou Tsunekawa. Automatic pcb inspection algorithms: A survey. *Comput. Vis. Image Underst.*, 63:287–313, 1996.

[64] Baback Moghaddam and Alex Pentland. Probabilistic visual learning for object detection. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 786–793, 1995.

[65] Joseph L. Mundy and Andrew Zisserman, editors. *Geometric Invariance in Computer Vision*. MIT Press, Cambridge, MA, USA, 1992.

[66] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2161–2168, 2006.

[67] M. Nölle, H. Penz, M. Rubik, K. Mayer, I. Holländer, and R. Granec. Dagobert - a new coin recognition and sorting system. In *The International Conference on Digital Image Computing: Techniques and Applications*, pages 329–338, 2003.

[68] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *Proceedings of the 9th European Conference on Computer Vision (ECCV) - Volume Part IV*, pages 490–503, 2006.

[69] Stephen O'Hara and Bruce A. Draper. Introduction to the bag of features paradigm for image classification and retrieval. *CoRR*, abs/1101.3354, 2011.

[70] Otavio Augusto Bizetto Penatti, Fernanda B. Silva, Eduardo Valle, Valerie Gouet-Brunet, and Ricardo da Silva Torres. Visual word spatial arrangement for image retrieval and classification. *Pattern Recognition*, 47(2):705–720, 2014.

[71] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, pages 143–156, 2010.

[72] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[73] Rainer Planinc and Martin Kampel. Robust Fall Detection by Combining 3D Data and Fuzzy Logic. In *ACCV Workshop on Color Depth Fusion in Computer Vision*, pages 121–132, 2012.

[74] J. Rabin, J. Delon, and Y. Gousseau. Circular earth mover's distance for the comparison of local features. In *The International Conference on Pattern Recognition (ICPR)*, pages 1–4, 2008.

[75] M. Reisert, O. Ronneberger, and H. Burkhardt. An efficient gradient based registration technique for coin recognition. In *MUSCLE CIS Coin Competition Workshop*, pages 19–31, 2006.

[76] J. C. Russ. *The Image Processing Handbook*. CRC Press, 5th edition, 2006.

[77] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlatons. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2033–2040, 2006.

[78] Bernt Schiele and James L. Crowley. Object recognition using multidimensional receptive field histograms. In *European Conference on Computer Vision (ECCV)*, pages 610–619, 1996.

[79] Bernt Schiele and James L. Crowley. Recognition without correspondence using multidimensional receptive field histograms. *Int. J. Comput. Vision*, 36:31–50, 2000.

[80] David Sear. *Roman Coins and Their Values*. Spink & Son, 5th revised edition edition, 2003.

[81] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1470–1477, 2003.

[82] Michael J. Swain and Dana H. Ballard. *Indexing via color histograms*. PhD thesis, 1990.

[83] Michael J. Swain and Dana H. Ballard. Color indexing. *Int. J. Comput. Vision*, 7:11–32, 1991.

[84] Yi Tao and William I. Grosky. Spatial color indexing using rotation, translation, and scale invariant anglograms. *Multimedia Tools Appl.*, 15(3):247–268, 2001.

[85] Marco Alexander Treiber. *An Introduction to Object Recognition: Selected Algorithms for a Wide Variety of Applications*. Springer Publishing Company, Incorporated, 2010.

[86] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

[87] Tinne Tuytelaars. Dense interest points. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2281–2288, 2010.

[88] Tinne Tuytelaars and Cordelia Schmid. Vector quantizing feature space with a regular lattice. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[89] A. Vailaya, M. A. T. Figueiredo, A. K. Jain, and H. J. Zhang. Image classification for content-based indexing. *IEEE Trans. on Image Processing*, 10(1):117–130, 2001.

[90] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[91] Manik Varma and Andrew Zisserman. A statistical approach to texture classification from single images. *Int. J. Comput. Vision*, 62(1-2):61–81, 2005.

[92] A. Vedaldi and A. Zisserman. Sparse kernel approximations for efficient classification and detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[93] Olga Veksler. Star shape prior for graph-cut image segmentation. In *Proceedings of the 10th European Conference on Computer Vision (ECCV): Part III*, pages 454–467, 2008.

[94] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3360–3367, 2010.

[95] Josiah Wang, Katja Markert, and Mark Everingham. Learning models for object recognition from natural language descriptions. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 2.1–2.11, 2009.

[96] Zhenhua Wang, Bin Fan, and Fuchao Wu. Local intensity order pattern for feature description. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 603–610, 2011.

[97] Isaac Weiss. Geometric invariants and object recognition. *Int. J. Comput. Vision*, 10(3):207–231, 1993.

[98] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1800–1807, 2005.

[99] John Wright, Allen Y. Yang, Arvind Ganesh, S. Shankar Sastry, and Yi Ma. Robust face recognition via sparse representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):210–227, 2009.

[100] Lingxi Xie, Qi Tian, and Bo Zhang. Spatial pooling of heterogeneous features for image applications. In *Proceedings of the ACM international conference on Multimedia (MM)*, pages 539–548, 2012.

[101] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *GIS*, pages 270–279, 2010.

[102] Yi Yang and Shawn Newsam. Spatial pyramid co-occurrence for image classification. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1465–1472, 2011.

[103] M. Zaharieva, M. Kampel, and S. Zambanini. Image based recognition of ancient coins. In *The International Conference on Computer Analysis of Images and Patterns (CAIP)*, pages 547–554, 2007.

[104] S. Zambanini and M. Kampel. Robust automatic segmentation of ancient coins. In *The International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 273–276, 2009.

[105] Sebastian Zambanini and Martin Kampel. Robust automatic segmentation of ancient coins. In *The International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 273–276, 2009.

[106] Sebastian Zambanini and Martin Kampel. Coarse-to-fine correspondence search for classifying ancient coins. In *Asian Conference on Computer Vision Workshops (2)*, pages 25–36, 2012.

[107] Sebastian Zambanini and Martin Kampel. Coarse-to-fine correspondence search for classifying ancient coins. In *eHeritage Workshop, ACCV*, pages 25–36, 2012.

[108] Sebastian Zambanini and Martin Kampel. A local image descriptor robust to illumination changes. In *Scandinavian conference on image analysis (SCIA)*, pages 11–21, 2013.

[109] Sebastian Zambanini and Martin Kampel. Classifying ancient coins by local feature matching and pairwise geometric consistency evaluation. In *The International Conference on Pattern Recognition (ICPR)*, pages 25–36, 2014.

[110] Sebastian Zambanini, Albert Kavelar, and Martin Kampel. Classifying ancient coins by local feature matching and pairwise geometric consistency evaluation. In *The International Conference on Pattern Recognition (ICPR)*. in print, 2014.

[111] E. Zhang and M. Mayo. Enhanced spatial pyramid matching using log-polar-based image subdivision and representation. In *The International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 208–213, 2010.

[112] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vision*, 73(2):213–238, 2007.

[113] Shiliang Zhang, Qi Tian, Gang Hua, Qingming Huang, and Wen Gao. Generating descriptive visual words and visual phrases for large-scale image applications. *Trans. Img. Proc.*, 20(9):2664–2677, 2011.

[114] Xi Zhou, Kai Yu, Tong Zhang, and Thomas S. Huang. Image classification using supervector coding of local image descriptors. In *Proceedings of the 12th European Conference on Computer Vision (ECCV)*, pages 141–154, 2010.