TECHNISCHE UNIVERSITÄT MÜNCHEN
Lehrstuhl für Mensch-Maschine-Kommunikation

Conception and Assessment of Semantic Feature Descriptors for Earth Observation Images

Reza Bahmanyar

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Prof. Dr. sc. techn. Gerhard Kramer
Prüfer der Dissertation:	1. Prof. Dr.-Ing. habil. Gerhard Rigoll
	2. Prof. Dr. rer. nat. Daniel Cremers
	3. Prof. Dr.-Ing. habil. Mihai Datcu

Die Dissertation wurde am 30.11.2015 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 30.06.2016 angenommen.

Abstract

The volume of civil high resolution Earth Observation (EO) images has steeply increased during the past decade due to numerous advances in airborne and spaceborne imaging technologies and has already leveraged a number of new applications. On the other hand, the large quantity of available images has extremely increased the challenge of exploring and understanding the full content of the image (i.e., their semantics). Therefore, the development of new image mining systems providing satisfactory results with reasonable computational effort, became highly demanded.

The existing EO image mining systems are usually based on extracted image features provided by various feature descriptors which can represent either pixel level patterns or the higher level semantics of images. Thus, developing feature descriptors which are able to represent the content of images relevant to the users' requirements helps to improve the accuracy and efficiency of image mining systems.

As a consequence, this dissertation introduces new approaches based on *Latent Dirichlet Allocation (LDA)*, a topic model for low and high level image feature descriptions. Moreover, the dissertation proposes novel methods based on LDA and information theory for evaluating various image feature descriptors independent of their application case. Since users usually evaluate image mining results based on their semantics, we conducted user studies for assessing the issues such as the *sensory* and the *semantic* gaps which affect the user acceptance of the results. Furthermore, this dissertation shows the importance of prior knowledge about the semantic structure of images in shortening the semantic gap between users and computers.

All corresponding experiments are conducted on multispectral and SAR (airborne and spaceborne) images; the results are validated by employing standard classification and clustering methods (e.g., SVM and k-means) in order to be comparable to previously obtained results in our discipline. The results demonstrate that by using higher level feature descriptors increases, the user acceptance of image mining results increases because the images are described by their semantic content. Furthermore, the results show that a evaluation of the feature descriptors regardless of their application allows us to generalize the evaluation outcomes to various ap-

plications. In addition, our studies and experiments indicate that the sensory and the semantic gaps should not be overlooked due to their high impact upon the user acceptance of image mining results. Finally, our analyses show that exploring the space of image features leverages an understanding of the image semantics.

Zusammenfassung

Die Menge an hochaufgelösten zivilen Erdbeobachtungsbildern hat sich in den letzten 10 Jahren aufgrund zahlreicher Fortschritte bei den Bildaufnahmetechnologien in der Luft- und Raumfahrt stark erhöht und hat bereits zu einer Reihe von neuen Anwendungen geführt. Andererseits hat die groe Menge an verfügbaren Bildern auch die Herausforderungen beim Durchsuchen und Verstehen ihres gesamten Inhalts (d.h. ihrer Semantik) extrem gesteigert. Daraus entstand der Bedarf nach neu zu entwickelnden Image-Mining-Systemen mit zufriedenstellender Güte und vertretbaren Rechenzeiten.

Die momentan vorhandenen Image-Mining-Systeme der Erdbeobachtung basieren in der Regel auf extrahierten Bildmerkmalen, die durch verschiedene Merkmalsdeskriptoren erzeugt werden, die entweder pixelbezogene Muster oder die Semantik von Bildern auf höherer Ebene darstellen. Daher hilft die Entwicklung von neuen Merkmalsdeskriptoren, die den Inhalt von Bildern, wie von den Nutzern erwartet, beschreiben können bei der Verbesserung der Genauigkeit und der Effizienz von Image-Mining-Systemen.

Daher stellt diese Dissertation neue Ansätze mit *Latent Dirichlet Allocation (LDA)* vor, einer Themenmodellierung für Merkmalsdeskriptoren auf niedriger und hoher Ebene. Weiterhin werden in der Dissertation neuartige auf LDA und auf der Informationstheorie basierende Methoden vorgeschlagen, um verschiedene Bildmerkmalsdeskriptoren unabhängig vom jeweiligen Anwendungsfall zu beurteilen. Da Nutzer ihre erhaltenen Image-Mining-Ergebnisse normalerweise aufgrund ihrer Semantik beurteilen, haben wir Nutzerstudien zur Beurteilung von Fragen wie der sensorischen oder semantischen Lücke durchgeführt, die die Nutzerakzeptanz der Ergebnisse beeinflussen. Weiterhin zeigt diese Dissertation, wie wichtig Vorwissen über die semantische Struktur von Bildern ist, um die semantische Lücke zwischen Nutzern und Rechnern zu verkleinern.

Alle zugehörigen Experimente wurden mithilfe von multispektralen und SAR-Bildern aus der Luft- und Raumfahrt durchgeführt; die Ergebnisse wurden mit Standardverfahren zur Klassifizierung und zum Clustering (z.B. mit SVM und k-

Means) validiert, um kompatibel mit bereits früher erhaltenen Resultaten in unserem Fachgebiet zu sein. Die Ergebnisse demonstrieren, dass die Nutzerakzeptanz von Image-Mining-Resultaten durch die Verwendung von höheren Merkmalsdeskriptoren steigt, da die Bilder dann durch ihren semantischen Inhalt beschrieben werden. Weiterhin zeigen die Ergebnisse, dass eine vom Anwendungsfall unabhängige Bewertung der Merkmalsdeskriptoren es erlaubt, die Bewertungsergebnisse für unterschiedliche Anwendungsfälle zu verallgemeinern. Darüber hinaus weisen unsere Untersuchungen und Experimente nach, dass die sensorische als auch die semantische Lücke wegen ihrer hohen Auswirkungen auf die Nutzerakzeptanz von Image-Mining-Resultaten nicht übersehen werden sollten. Schlielich zeigen unsere Untersuchungen auch, dass eine Untersuchung des Bildmerkmalsraums das Verstehen der Bildsemantik unterstützt.

Acknowledgments

I would like to show my sincere gratitude to Prof. Dr. Mihai Datcu, my supervisor at the German Aerospace Center (DLR), for his continuous scientific guidance, support, and motivation. I am also grateful to Prof. Dr. Gerhard Rigoll for giving me an opportunity to carry out this research work under the doctoral program of Technical University of Munich. I am so thankful to Prof. Dr. Peter Reinartz for providing me a very encouraging working environment at the Remote Sensing Technology Institute (IMF) of DLR along with motivation and supports over all these years. Additionally, I would like to mention that this dissertation would not have been possible without financial and administrative supports of Munich Aerospace.

I would like to thank Mr. Gottfried Schwarz for the favor of proofreading this dissertation. He gave me many helpful comments about language issues and the structure of my dissertation. I would like to thank Ambar Murillo Montes de Oca for her continuous motivations and helps in thinking out of the box and making this dissertation a multidisciplinary research by including psychology and human cognition disciplines. She also helped me to smooth out the language in this dissertation. I would like to thank Dr. Shiyong Cui and Mohammadreza Babaee for the valuable technical discussions we had, which allowed me to know and learn from their experiences. I would like to thank all my other friends in DLR: Dr. Daniela Espinoza Molina, Dr. Corneliu Octavian Dumitru, Wei Yao, Kevin Alonso Gonzalez, Nazli Deniz Cagatay, Dr. Jagmal Singh, and Peter Schwind. Many thanks to Ms. Theresia Hantel for her kind help in solving various daily difficulties in life during my doctoral research study in DLR.

Most importantly, none of this would have been possible without the love and patience of my family. My deepest gratitude to my beloved wife Tannaz, and my parents who supported me all possible ways to keep me motivated during these years of research.

Contents

1

Introduction

1.1 Motivation

Recent advances in airborne and spaceborne Earth Observation (EO) imaging technologies for high resolution optical (multispectral) and Synthetic Aperture Radar (SAR) instruments have led to a steep growth in the volume of civil EO images within the 1 m resolution category. Typical examples are the multispectral imagers WorldView-2 and WorldView-3 as well as the SAR satellites TerraSAR-X and COSMOSkyMed. The specific characteristics of these high resolution EO images, as well as their availability in remote sensing archives, have leveraged new high resolution EO applications allowing, for instance, highly accurate small-scale object identification and classification [1]. However, the large quantity of available images has tremendously increased the challenge of exploring the full amount of the image content in order to provide focused information in a simple and intuitive human understandable form. Dealing with this challenge, the development of new image mining systems, such as classification and retrieval systems providing results satisfactory to users with reasonable computational effort, were in high demand. Existing EO image mining systems are usually based on descriptions of extracted image features provided by various feature descriptors [2, 3, 4]. These descriptors can be either pixel level or higher level descriptors; the latter ones are built upon the former ones. While the pixel level descriptors provide primitive image features, the higher level descriptors represent images with their semantic content. Thus, developing feature descriptors which are able to represent image content relevant to the users' requirements improves the accuracy and efficiency of image mining systems.

Taking all the above into consideration, the focus of this dissertation is to introduce new approaches for low and high level image feature descriptions. Moreover, it proposes novel methods for evaluating various image feature descriptors, independent of the application case. Since users usually evaluate image mining results based on their semantics, this dissertation studies the issues which affect the user

acceptance of the results, such as the *sensory gap* and the *semantic gap*, which will be discussed below.

In order to validate the proposed methods and to be comparable to the previous researches in the field, in this dissertation, standard classification and clustering methods (e.g., SVM and k-means) are employed. Moreover, we conduct our experiments on multispectral images acquired from the WoldView-2 satellite and SAR images obtained from the TerraSAR-X satellite. In addition, in order to be comparable to exiting research in the field, the proposed methods are tested on a widely used multispectral remote sensing dataset, namely the UC Merced Land Use Dataset [5]. Experimental results demonstrate that using higher level feature descriptors increases the user acceptance of image mining results because they describe images by their semantic content such as objects or object parts. Furthermore, the results show that evaluating the feature descriptors, regardless of their application, makes it possible to generalize the evaluation outcomes to various applications. Furthermore, our studies and experiments indicate that the sensory and the semantic gaps (to be discussed below) should not be overlooked due to their high impact upon the user acceptance of image mining results.

1.1.1 Semantics-driven Image Content Descriptors

High resolution EO satellite images have opened up the opportunity for detailed local object discrimination and identification (especially within urban areas), which was not possible with previously available low and medium resolution EO images [6]. Therefore, neither the conventionally used feature descriptors, nor the corresponding image analysis techniques are sufficient to deal with the semantic contents of high resolution EO images. The existing feature descriptors usually focus on certain properties of the images and represent them as a vector, the so-called *feature vector*, where the vector elements contain the components of these image properties. However, due to the variety of the object features in high resolution images, it is essential to consider multiple image properties. This is essential for developing highly discriminative feature descriptors.

In order to combine several image properties, various feature fusion methods have been proposed in the literature such as feature vector concatenation [7]. Since feature fusion usually increases the dimensionality of the fused feature vector and the vector elements may contain redundant information, several compact feature selection and dimensionality reduction methods have been suggested [8]; however, because these methods will usually be evaluated based on the statistical analysis of the fused feature vectors applied in feature fusion scenarios, they may miss some valuable properties of real images. Moreover, in order to semantically describe the objects in a given high resolution EO image, the local image context also has to be considered. This is not possible when only using the conventional pixel level feature descriptors and image analysis techniques. Therefore, we have to rely on

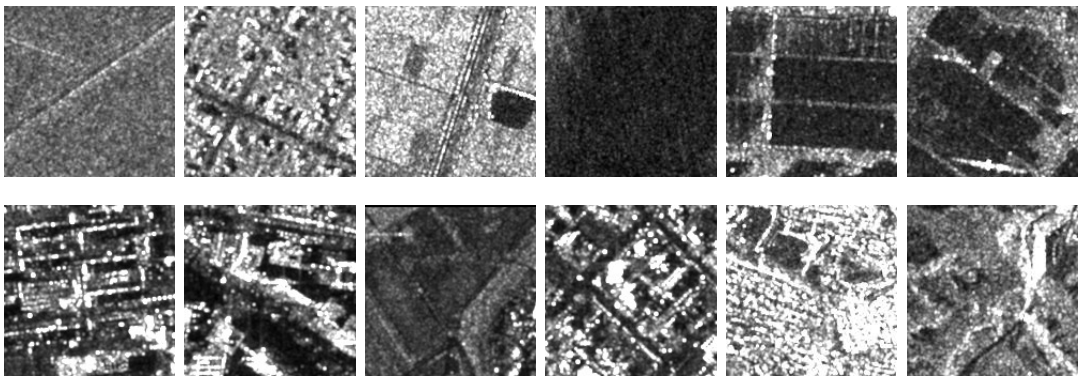


Figure 1.1: Typical highly diverse SAR image patches.

feature descriptors which are able to describe objects as semantics by exploiting their various characteristics and local spatial contexts. Furthermore, in order to use semantics for object discrimination and identification, semantics-driven image analysis techniques should be used. This issue is illustrated by Figure 1.1 showing highly diverse and typical example patches of a high resolution SAR image, acquired by the TerraSAR-X satellite.

Since image description lies at the foundation of any image mining system, selecting a feature descriptor or a combination of feature descriptors which are able to represent various image properties, affects the retrieval results to a large degree. A comprehensive representation of images exploiting their diverse properties while avoiding an arbitrary increase in the feature vector size, calls for a selection of the most representative feature descriptors which share only a few redundant image properties.

In order to select optimized feature descriptors, their abilities to discriminate different object categories should be evaluated for each application. However, due to the different requirements of various remote sensing applications, an evaluation of feature descriptors, by applying them to a single task and measuring their case-specific performance, cannot provide a generally valid assessment of their performance. Thus, it is very important to look for generally valid feature descriptor evaluation methods which can be applied to every image mining task.

As a consequence, this dissertation proposes various feature descriptor evaluation methods based on statistics and information theory. The proposed methods not only help to select optimized feature descriptors for feature fusion tasks, but also allow a generally valid feature descriptor evaluation. In order to be comparable to the previous research in the domain, a number of experiments were conducted, For each dataset, the mostly applied feature descriptors are re-considered and further analyzed. The computed feature vectors form a Euclidean space, the so-called *feature space*. Then standard classification and clustering algorithms such as SVM and k-

means are applied to this feature space. Their performance is further evaluated by conventionally used metrics such as classification accuracy.

1.1.2 The Sensory and the Semantic Gaps

The already mentioned image mining systems usually rely on human supervision resulting in semantic data annotations, either having been entered during training, verified during validation, or attached during routine operations. However, since the "gold standard" is set by user created references or user acceptance, user subjective biases are included in this standard. Due to the different image interpretations of various users being prone to the sensory and the semantic gaps (see below), user-created standards may be incorrect or inconsistent. As a result, established image mining systems based on individual user image interpretation can still perform unsatisfactorily to other users.

1.1.2.1 The Sensory Gap

The sensory gap refers to the differences between an object in reality and its representation based on the signals recorded by sensors. Causes behind the sensory gap can lie in the scene (e.g., perspective, occlusion, typical scene variance) or on sensor level (e.g., resolution, field of view, recorded spectral bands, detector noise) [9]. In EO images, the sensory gap for human interpretation is rather wide due to the various applied sensors (e.g., radar, multi- and hyper-spectral instruments) which record their information very differently from the human visual system. For example, SAR is a form of radar imaging which creates images by transmitting, receiving and processing radar pulses. SAR illuminates a target area from a moving carrier by transmitting repeated pulses of radio waves. The pulse echoes are then received and the image pixel amplitudes together with their phase information become available as data products after focusing [10]. In contrast, the human visual system reacts exclusively to visible light to create images of objects. As another example, multispectral sensors acquire images at selected optical frequency bands across the electromagnetic spectrum, including frequencies beyond the visible light range such as data in the ultraviolet or infrared. Although these optical sensors capture images in a passive way (as the human visual system does), the acquired images may contain information about objects which the human visual system is not able to perceive by its receptors tuned to red-green-blue (RGB) stimuli. Figure 1.2 shows typical remote sensing data acquired by the WorldView-2 multispectral satellite, which captures images in five visible spectral bands, ranging from 400 nm to 690 nm, including the well-known RGB bands. In addition, WorldView-2 provides data from three bands of (near-)infrared wavelengths within the 705 nm to 1040 nm wavelength range (even better band properties are expected from the upcoming WorldView-3 instrument). The first row of Figure 1.2 depicts the RGB band images,

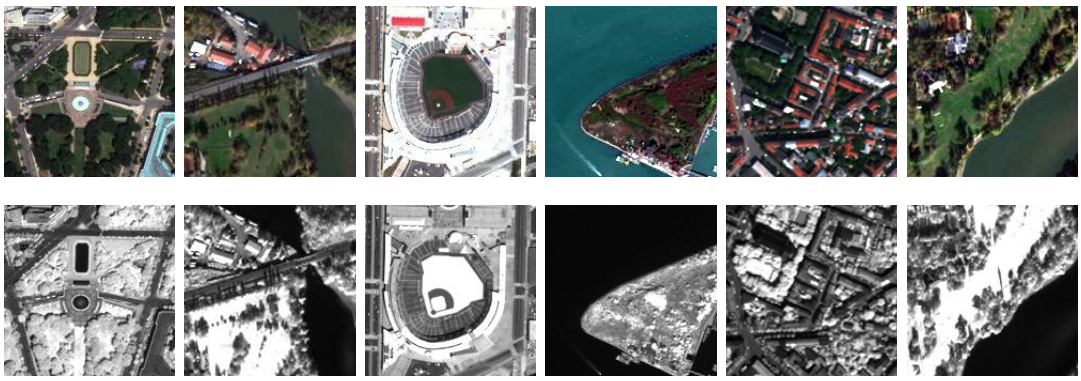


Figure 1.2: Top row: multispectral RGB patches, Bottom row: infrared patches.

while the second row shows the three infrared band images. The images illustrate that vegetated regions are characterized by strong infrared signals providing valuable information about plants with high chlorophyll content, while the human visual system only sees a strong green component of the vegetation.

In addition to the type of sensor, the complexities of EO images such as their resolution, perspective, or scale of the visual information, also affect the sensory gap [11]. The perspective of satellite images, for example, is a particular challenge in the interpretation of EO images. They represent objects from a bird's eye view which human users are not accustomed to. The sensory gap also includes effects of digital image processing. In order to interpret an image, processing routines discriminate the objects within an image using the descriptions of their various feature types (e.g., color, texture, and shape), using different feature descriptors. The discrimination of the objects is further used for object identification. Therefore, in addition to the sensor characteristics and image complexities, feature descriptors strongly affect the sensory gap due to image processing routines. Taking all the above into account, object discrimination is a fundamental step in both human and computer image interpretation which is affected mainly by the sensory gap. Object naming is then performed based on the interpreted image.

1.1.2.2 The Semantic Gap

The semantic gap has been defined by most researchers as the difference between the user's understanding of objects in an image, and the computer's interpretation of those objects [9, 12, 13, 14]. When users are presented with an image, they first discriminate the objects in it and then name the objects. Digital image processing follows a similar procedure; however, different methods are used for each step (e.g., object discrimination based on feature descriptors, object naming according to machine learning algorithms), which causes the semantic gap. In addition to the semantic gap between users and computers, each individual user will interpret images

slightly differently, and will use different terms to label the objects within them; thus, a semantic gap also arises between users which we call the linguistic semantic gap. While previous research addressed this gap as a vocabulary problem [15, 16], showing that it is unlikely for two people to assign the same label to a given object, this problem has not yet been considered in the context of the well-known semantic gap. However, since the existing image mining systems have usually been verified either by comparing their results to reference data or by measuring the degree of user acceptance in interactive systems, neglecting the linguistic semantic gap may make the image retrieval results for a specific user and search goal unsatisfactory to other users.

1.1.3 Quantifying and Analyzing the Gaps

In this dissertation, a set of user experiments are conducted to study different factors in user perception and identification of objects within EO images. The results are then quantified and analyzed by statistical methods such as the *Kullback-Leibler divergence* method [17]. In order to measure the semantic gap between users and computers, the results of the user experiments are compared to the results of existing machine learning methods such as *Latent Dirichlet Allocation (LDA)* [18] and k-means clustering. In our user experiments, an RGB image from the WorldView-2 satellite is used. Since the sensor characteristics of this image (i.e., the passive signal recording and the RGB spectral bands) are close to those of the human visual system, the analysis of the effects of the image complexities (which are shared by all EO image types) on the sensory and the semantic gaps is not affected by the sensor-specific characteristics.

The results highlight EO image complexities as causes of the sensory gap, which strongly affect the object identification by users. Moreover, our experimental results show that subjective biases exist when the results of image mining are validated neglecting the linguistic semantic gap.

In order to overcome this problem, our proposal is to take into account the linguistic semantic gap, and to increase the diversity of our data sets being used in each domain (e.g., using various EO datasets for the foreseen EO applications), which will include different user perspectives and compensate for the individual subjective biases. Moreover, user-trained models could be stored and further used by other systems, including image interpretations made by several users.

Furthermore, the sensory gap discussed above has a significant influence on the semantic gap together with other issues such as the background knowledge of users. In particular, feature descriptors cause a key influence on the sensory gap due to digital image processing and, as a consequence, on the semantic gap. Therefore, using highly descriptive feature descriptors which extract more relevant information according to the user requirements helps to reduce the semantic gap.

1.2 Image Feature Descriptors

In this section, we summarize the image description methods which have been used during past years in EO image analysis. We first review the commonly used primitive feature extraction methods in the EO domain. Then we show how the image descriptors and image analysis techniques shifted over the years from pixel level to higher semantic levels in the EO domain. Following this, we introduce the methods which are usually used for evaluating and comparing feature descriptors.

1.2.1 Primitive Feature Extraction Methods

In this section, we briefly review primitive feature extraction methods which have been applied to EO images (e.g., multispectral and SAR images) over the years. In this review, we mainly focus on texture feature descriptors, since they have been applied to EO images in many previous publications. Texture refers to spatial intensity variations in an image which cause repetitive patterns, the so-called *textons* [19]. As Figure 1.1 and Figure 1.2 show, texture is the most common feature occurring within multispectral and SAR images. In addition, we introduce primitive feature extraction methods which mainly extract the geometry-based image features such as edges and corners.

In the following, we show that a variety of methods have been employed to extract EO image features. The decision as to which method is more powerful and accurate usually depends on the specifications of the EO images selected for a specific task [20, 21, 22].

1.2.1.1 Patterns of Pixel Values

The simplest approach to extract the texture features of an image is vectorizing the pixel values within a certain neighborhood of every pixel. Clearly the neighborhood must be large enough to encompass the dominant texture variations. However, by enlarging the neighborhood, the dimensionality of feature vectors increases, which poses a larger computational burden to image mining systems. In order to make the dimensionality of the feature vectors independent of the neighborhood size, one approach is to use histograms of pixel values within a neighborhood. This approach quantizes the pixel values to a certain number of bins. The *RGB Color Histogram (rgbHist)* has been used in many previous researches for color image analysis [23]. In the EO area, Yang and Newsam [5] employed rgbHist features, extracted from RGB color channels of a multispectral image dataset, in a classification task. The results indicated that rgbHist features can give superior results for the object classes which are homogeneous in color. In SAR images, however, due to the rather large range of the pixels' brightness values, quantization leads to either a high quantization error (when a small number of bins is used) or very high dimensional feature vectors

(when a large number of bins is used). In order to extract texture information of SAR images using their raw pixel values, Cui *et al.* [24] introduced *Mean-Variance-Ratio (MVR)* texture descriptors. Statistical moments such as mean and variance have been used for a compact representation of texture features of various image types such as multimedia and medical images [25, 26]. For SAR images, in regions with fully developed speckle, the relation between mean and variance is $L = \frac{m^2}{\sigma}$ (where L is the number of looks, m is the mean, and σ is the variance). However, this relation does not hold in the regions with strong structures. Therefore, MVR uses mean ratios in different directions in addition to the local mean and variance which results in a superior discrimination ability [24]. As another approach, Liu and Fieguth [27] introduced a *Random Projection (RP)* method for projecting high dimensional vectors into a lower dimensional vector space without the loss of salient information. Using RP, the pixel values within an arbitrary large neighborhood can be vectorized; and then the dimensionality of the resulting vector is reduced to a reasonable size. In the EO area, this method has recently been successfully applied to SAR image segmentation by Hou *et al.* [28].

As another method to represent local image textures, Chen *et al.* [3] proposed the *Weber Local Descriptor (WLD)* inspired by Weber's law, saying that humans notice the change in a stimulus as a valid signal if its ratio to the original intensity of the stimulus is above a certain constant value. The WLD feature vector for each given pixel is generated by computing its differential excitation (i.e., the ratio between the intensity difference between the current pixel and its neighbors) and the gradient orientation of the current pixel. WLD has been successfully applied to various multimedia image mining tasks such as face recognition [29]. In EO image scenarios, Cui *et al.* [24] studied the performance of WLD on SAR images. The authors showed that the presence of multiplicative speckle noise limits the discriminability of WLD in SAR images. In order to remedy this limitation, Cui *et al.* [24] proposed to replace the gradient orientation mechanism of WLD with the *Ratio of Mean Differences (RMD)* in vertical and horizontal directions. They named this method *Adapted Weber Local Descriptor (AWLD)*. The idea of using RMD has been initiated by the *ratio edge detector* which has been developed by Touzi *et al.* [30] in order to compensate the effects of the SAR speckle noise.

A simple but powerful method for representing the local image patterns is *Local Binary Patterns (LBP)* proposed by Ojala *et al.* [31]. LBP considers a circular local neighborhood around each image pixel. Then it assigns binary values "0" or "1" to the neighboring pixels if their values are smaller or greater than the value of the central pixel, respectively. LBP is designed to be gray-scale and rotation invariant [31]. To the best of our knowledge, LBP was first applied to EO images by Lucieer *et al.* [32] for segmenting CASI (Digital Compact Airborne Spectrographic Imager) and LiDAR (a type of imagery in which a laser pulse is transmitted and its reflection is used to measure the distance) images. It was later used by a number of works for EO image classification [33, 34, 35]. However, to the best of our

knowledge, LBP has not been applied to SAR images due to the specific properties of SAR images such as the existence of speckle noise. In order to use the local patters of SAR images, Dai *et al.* [36] proposed a theoretically and computationally simple feature, the so-called *Multilevel Local Pattern Histogram (MLPH)*. This method quantizes SAR images into various channels at particular contrasts and then generates a histogram of the local image patterns. The authors compared the classification accuracy obtained by using MLPH to other texture descriptors and showed that it is superior.

1.2.1.2 Pixel Value Co-occurrence Feature Extraction

Based on the idea that the spatial relationships between the pixel values over an image found the texture information of the image, Haralick *et al.* [2] proposed the use of statistics of the *Gray Level Co-occurrence Matrix (GLCM)* as texture descriptors. The co-occurrence matrix is generated by measuring the occurrence of a specific gray value in a given position within the image. The feature vector is then obtained by computing the statistics of the co-occurrence matrix such as contrast, homogeneity, correlation, and entropy. The features extracted by GLCM has been used for classification of EO multispectral images by a number of previous works [5, 37]. GLCM has been successfully applied to SAR images for various tasks such as mapping sea ice patterns [38], water surface extraction [39], and agricultural crop classification [40]. In order to reduce the computational burden of co-occurrence matrix generation, Kandaswamy *et al.* [41] proposed an approximate texture representation based on the notion of patch re-occurrence. The authors showed that in this scenario, GLCM leads to highly accurate unsupervised SAR image classification.

1.2.1.3 Model-based Feature Extraction

The idea that an image is the result of a stochastic random process has led to the development of a number of model-based image texture description methods such as *Gaussian Markov Random Fields (GMRFs)*. In these methods, the feature vectors are derived based on the parameters of the stochastic process. As an example of previous early works using GMRFs in the EO domain, we can mention the work by Chellappa and Chatterjee [42], who employed features extracted based on GMRFs in a classification scenario, and the work by Tsai and Tseng [43] who used GMRF-based features for the segmentation of multispectral EO images. Furthermore, various GMRFs have been proposed and successfully applied to SAR image texture representation in many image mining tasks [4, 44, 45, 46]. Recently, an intensive study of GMRFs for spatial content understanding of SAR images has been conducted by Singh [47].

1.2.1.4 Feature Extraction in Transform Domains and Using Filter Banks

Prior research has shown that the extraction of SAR image features is easier in the frequency domain such as the features of moving objects. Since for moving objects the echo signal is non-stationary, accumulating the signal's energy and extracting the object features using spatial domain feature descriptors is difficult [48]. In order to ease this difficulty, a series of previous works [49, 48, 50] proposed to use a *Matched Fourier Transform (MFT)* due to its important properties introduced by Wang *et al.* [51]. As another method dealing with moving objects, Sun *et al.* [52] proposed to use a *Fractional Fourier Transform (FrFT)*, a generalized form of a Fourier transform. This transformation has been introduced to the signal processing domain by Almeida in [53]. It owes its strength to considering both Doppler frequency and Doppler modulation rate of non-stationary signals. Moreover, the linear operator used in FrFT, makes it robust in the presence of multiple moving objects [52]. In the context of stationary objects, Singh and Datcu [54, 55, 56] proposed a number of methods based on FrFT for SAR image classification. The authors showed that the ability of FrFT to use the phase information of complex-valued SAR images makes it superior to spatial domain feature descriptors.

Describing SAR image features by a set of parameters computed based on a *Short Time Fourier Transform (STFT)* has been proposed by Popescu *et al.* [57]. Based on this method, Li and Ogihara [58] introduced a non-linear STFT for the retrieval of music information. It has been shown in [20] that a non-linear STFT also leads to superior results in SAR image classification.

Another method for analyzing an image in the frequency domain is using *Quadrature Mirror Filter (QMF)* banks [59, 60, 61]. These filter banks split a discrete-time signal into different sub-bands, where each sub-band can then be processed independently. One of the first applications of QMF banks in a content-based image retrieval for various EO image types was made by Li and Castelli [62]. Recently, QMF banks have been further successfully applied to SAR image classification by Dumitru and Datcu [20].

Due to the importance of scale in texture analysis, various wavelet-based texture feature descriptors have been proposed for SAR image analysis in the literature. Fukuda and Hirosawa [63] used a dyadic decomposition in combination with a bank of low and high pass filters for image wavelet transformation. They showed that the proposed feature descriptor provides promising classification of polarimetric SAR images. In order to represent the multiple scales and orientations of images, Gabor filter banks have been widely used in previous works. Even though applying Gabor filtering results in a wavelet decomposition of an image, due to the non-orthogonality of the decomposition, it suffers from redundancy [64]. Therefore, a bank of selected Gabor filters with various scales and rotations is usually applied to the given images. Furthermore, Gabor filtering is used as a linear filter in image processing standards

such as MPEG-7 [65]. Gabor filter banks have been used in various image analysis applications in the EO domain such as image retrieval [62, 66]. To the best of our knowledge, in SAR image analysis, Gabor filter banks have been used by Du [67] in a segmentation task. This method has been applied later to various SAR image segmentation and classification tasks [68, 69]. In a number of recent works, Singh and Datcu [54, 56] introduced new variations of using Gabor filter banks for SAR images. In these methods, log-cumulants of Gabor filtered images with enhanced performance over linear moments have been used. The authors showed the superior performance of these methods in SAR image classification.

1.2.1.5 Interest Point Detectors and Feature Descriptors

The *Scale-Invariant Feature Transform (SIFT)* proposed by Lowe [70] is an interest point detector and feature descriptor based on the geometry of local image features such as edges and corners. SIFT has been widely used by many previous researches due to its invariance against translation, rotation, and image scale. In the EO domain, SIFT has been mostly adapted and employed for image registration. For example, Hasan *et al.* [71] adapted the SIFT descriptors for the registration of multi-modal images (i.e., images acquired by different sensors). Recently, Sedaghat and Ebadi [72] introduced a new variant of SIFT so-called *Adaptive Binning Scale-Invariant Feature Transform (AB-SIFT)* for EO image registration. They showed that AB-SIFT is robust against local geometrical distortions. In addition to image registration, Yuan and Hu [73] employed SIFT descriptors in an EO image classification scenario for extracting clouds. In a comparative study, Yang and Newsam [74], compared the performance of SIFT descriptors and Gabor filter banks for the classification of panchromatic images acquired from the IKONOS satellite. They used two classification techniques, namely a *Maximum A Posteriori (MAP)* classifier [74] and a *Support Vector Machine (SVM)* classifier [75]. The results showed that when SIFT descriptors are used, MAP achieve a higher classification accuracy than when the Gabor descriptors are used; however, for an SVM classifier, Gabor feature descriptors are more discriminative than SIFT descriptors. Recently, Dellinger *et al.* [76] proposed the *Synthetic Aperture Radar Scale-Invariant Feature Transform (SAR-SIFT)*, a new adaptation of SIFT to SAR images. They introduced a new method for computing the image gradient which makes SAR-SIFT robust against speckle noise.

Being inspired by SIFT, Bay *et al.* [77] proposed *Speeded Up Robust Features (SURF)* which is much faster and more robust than SIFT against various image transformations due to using a combination of two-dimensional Haar wavelet responses, scale space techniques, and the integral images. Therefore, this technique has been preferred by a number of authors in the EO domain for detecting interest points and describing image features [78, 79, 80]. However, other previous works such as [5] preferred SIFT for SURF, because the authors believe that SIFT pro-

vides a higher degree of invariance. SURF has been applied to SAR images for image registration by [79, 80].

1.2.2 From Pixel Level to Semantic Level Descriptors

In previous research, EO images usually have been considered as a whole and have been analyzed by conventional techniques such as segmentation and pixel-based classification [33, 36, 37]. These methods perform quite well for low resolution EO images, where the texture over each image region is relatively homogeneous (e.g., agricultural fields, water surfaces). However, the recently available high resolution satellite images provide a much higher number of object categories and the applied object discrimination techniques have to include higher level descriptions of the local image contexts. This has shifted interest toward patch-based image analysis in recent years. To the best of our knowledge, a patch-based approach of remote sensing images has been first considered for analysis in [81]. This method was later applied to SAR images by a number of researches [24, 82]. In a patch-based method, an image is usually split into non-overlapping patches, where the patch size is selected so that each patch holds enough information on the local image context (objects or object parts). Each image patch is then assigned to an object category as containing a particular object.

In order to analyze image patches using the higher level descriptions of their local image contexts, *Bag-of-Words (BoW)*, a simplifying method developed for natural language processing, has been successfully modified and applied to remote sensing images of various types [5, 83, 84, 24]. In this model, feature vectors are computed for all the pixels on a given grid of each image patch using a sliding window of a given size. The local feature vectors of the entire image patch collection then generate a feature space, where each feature is considered as a point, a so-called *feature point*. In a next step, the structure behind the distribution of the feature points in the feature space is modeled by a *codebook* which is composed of a set of basis points, the so-called *codewords*. The codewords are usually defined by applying a clustering method (e.g., k-means) on randomly selected samples of the feature points. In a final step, using the codebook, the local feature vectors of each image patch are coded to integrate all important features of the image patch into a single vector. To this end, a code matrix is generated, where each row contains the response values of a particular feature point to different codewords. The response values can be obtained using various coding schemes such as voting-, reconstruction-, and salient-based methods [85]. The response values of all the feature points to each codeword are then integrated to form a single code value using a pooling technique (e.g., sum, average, maximum pooling) [86]. The output of the pooling step is a vector with a dimensionality equal to the codebook size. In contrast to global feature vectors, since the entire patch collection is contributing to the codebook generation, the codewords in BoW are able to provide a higher level description of local image contexts.

1.2.2.1 Feature Coding Methods

A variety of possibilities to generate codebooks as well as various methods to compute the response values make feature coding a hot topic. The classical feature coding method is *Hard Voting (HV)* [87]. For each codeword, HV counts the number of its nearest neighboring feature points as the code value. Using a kernel function, *Soft Voting (SV)* [88] is proposed to not only consider the distances between the feature points and the codewords, but also to allow each feature point to be described by more than one codeword. In order to improve the feature space structure description of the codewords, reconstruction-based methods have been applied to coding scenarios. These methods reconstruct each feature point by a group of codewords constrained by the number of the contributing codewords such as *Sparse Coding (SC)* [89], and the locality of the codewords such as in *Local Coordinate Coding (LCC)* [90] and *Local-constraint Linear Coding (LLC)* [91]. Considering the locality of the codewords in combination with the maximum pooling in LLC leads to a salient representation of the feature points. More precisely, if a certain number of nearest codewords are used to code a feature point, codewords closer to the feature point will receive a stronger response than the others. In order to represent the salient characteristics of the feature points and to avoid the computation cost of the LLC method, *Salient Coding (SaC)* [92] and its variants such as *Group Salient Coding (GSC)* [93] have been proposed. All methods code each feature point independently of the other points. However, Roweis and Saul in [94] showed that considering the relationships between the neighboring points helps to discover the global structure of the data. Therefore, in this dissertation, we propose a coding method which considers each point in relationship with its neighboring points as a local structure of the feature space, instead of treating each point individually. We name this method *Locally Linear Salient Coding (LLSaC)* [95] which is a new variant of SaC.

1.2.2.2 Semantic Level Feature Descriptors

Representing image patches using BoW models does not provide a good estimate of image semantics due to disregarding the statistical relations between the visual words. However, it has been shown in previous works ([96, 97, 98, 99]) that these relations can result in the discovery of objects and their parts in images. These works have used generative models such as *probabilistic Latent Semantic Analysis (pLSA)* [100] and LDA [18] for the unsupervised discovery of object parts, the so-called *topics*. The image contexts are then represented by mixtures of the discovered topics. The authors of [96] investigate pLSA and LDA for object categorization and localization. They demonstrated the possibility of recognizing and localizing object categories by learning from unlabeled image collections. In [97], the authors used the topics obtained by pLSA in combination with a nearest neighbor classi-

fier for scene classification. They showed that the statistical model discovered by pLSA is appropriate for the classification of datasets with multiple object categories in each image. In [98], it has been shown that pLSA-based image representation improves the retrieval performance on large-scale datasets due to the compact description of the image contents. Inspired by [101], Hoerster *et al.* in [99] verified that the topics discovered by LDA outperform the ones obtained by pLSA in large-scale retrieval tasks owing to the completely generative probabilistic model provided by LDA. Later, various extensions of LDA have been introduced for scene classification and segmentation [102, 103]. In this dissertation, we study the discrimination of EO image patches using their semantic level representations, which we name *Bag-of-Topics (BoT)* [104], obtained from the BoW models of image patches.

1.2.3 Feature Descriptor Evaluation

There are many previous efforts to compare different feature descriptors based on the performance of their image-related tasks such as classification, retrieval, and matching. In a comprehensive study of local features [105], the authors evaluated different descriptors for classification tasks. They extracted various image features using different methods. Then they compared the classification accuracies using different descriptors. In another survey [106], the authors evaluated the accuracy of feature descriptors for image retrieval tasks. They measured the relevance of the images described by different feature descriptors to a given query. In a recent comparative evaluation article [107], the authors compared feature descriptors in image matching tasks. They used five different metrics to compare the performance of different descriptors in finding the matched points on the images. These comparative evaluations usually require reference images with well understood content from various aspects such as color, texture, and shape. However, the content of the visual data is not fully known due to limitations such as the sensory gap and the semantic gap [9].

This problem is even more serious in EO scenarios (especially for SAR images) because the sensors record signals which are very different from what the human visual system perceives [11]. There are a number of studies regarding SAR images which compare various feature descriptors [20, 21, 22]. However, they also limited their comparisons to a specific task such as classification, or feature descriptor performances measured by comparing the labeling results to a reference annotation. Dealing with these limitations, in this dissertation, we propose various feature descriptor evaluation methods based on statistics and information theory [108, 109, 110]. These methods are designed to provide a generally valid evaluation of feature descriptors for a given dataset.

1.3 The Sensory and the Semantic Gaps

In spite of the large efforts in developing efficient image mining systems, the results of the developed systems, particularly in EO, are not always satisfactory for users conducting content based searches [12]. While image mining systems usually perform based on user supervisions in the form of annotated data (either for training or validation), the content of visual data still is not fully understood by users due to the issues caused by the *sensory* and the *semantic* gaps [9].

1.3.1 The Sensory Gap

The sensory gap refers to the difference between an object in reality and its interpretation based on the signals recorded by sensors [9]. It is caused either by the scene parameters (e.g., clutter, occlusion, illumination) or by the sensor parameters (e.g., perspective, perceptual spectra). In multimedia imaging, the signals recorded by the cameras as well as the image perspectives are quite familiar to the users. Therefore, the sensory gap is narrow and can be attenuated by training image mining systems on multiple interpretation of objects [9]. In EO, however, the sensory gap is rather large due to the wide variety of sensors (e.g., radar, multispectral, hyperspectral) which record signals very differently from the human visual system [11]. Moreover, the complexities of the EO images (e.g., resolution, perspective, or scale of the visual information) have a large impact on the sensory gap [11].

The perspective of the images is a particular challenge in EO, since they present a bird's eye view. As described in the *recognition by components* theory [111], objects can be segmented into their *geometric components (geons)*, and we recognize them based on the identification of their geons and their structural relationships, which we then match to mental representations. Object recognition should be perspective invariant, so long as the structural relationship between geons can be identified from the different perspective. This is not the case when objects are viewed from above, since major object components can be occluded, making it harder to match the object to the stored mental description. Therefore, from this perspective, object identification is more difficult [111].

Despite the important role of the sensory gap in user image understanding and developing efficient image mining systems, to the best of our knowledge, it has not been studied yet in multimedia or in the EO domain. Therefore, in this dissertation, we evaluate the sensory gap in EO images using human perception and a computational method based on an LDA topic model [112].

1.3.2 The Semantic Gap

Most of the previous research has defined the semantic gap as the difference between the object understanding of the users in an image, and the computer's interpreta-

tion of those objects [9, 12, 13, 14, 113]. While users seek semantics (objects and thereof parts), image mining systems process images based on their primitive visual properties (primitive features) such as texture [113]. Since usually combinations of primitive properties build semantics, image mining systems which conduct their object discriminations only based on the primitive features are not able to correctly detect the semantics within an image.

In addition, users interpret image semantics differently, and use different terms to label them. We call this difference between the users' semantic interpretations of images the *linguistic semantic gap*. Previous research addressed this as a *vocabulary problem* [15, 16], showing that it is unlikely for two people to assign the same label to a given object. The demand for developing more efficient data mining systems has been met with methods usually performing based on human supervision in the form of annotated data, either for training or validation. Thus, different manually annotated datasets have been created; and are used for various purposes. However, due to disregarding the linguistic semantic gap, Torralba and Efros [114] showed that in spite of efforts devoted to creating general and unbiased datasets, due to subjective and objective reasons (e.g., the purpose of the datasets), they suffer from strong built-in biases. As a result, the verified systems based on reference datasets still do not provide results satisfying the user requirements [14]. This has also been confirmed in [114] by training a model on a dataset and then testing it on another one. The results showed that the agreement is low even between datasets which appear to be similar. In EO scenarios, the linguistic semantic gap is even more challenging due to the rather wide sensory gap. In spite of its large impact on the user acceptance of image mining systems' results, the linguistic semantic gap has been rarely considered in the context of the well-known semantic gap. Theodosiou *et al.* [115] conducted a user experiment to study the linguistic semantic gap as a cause of the existing built-in biases in manually labeled image datasets. They investigated the influences of users' ages and genders on the linguistic semantic gap, in an image labeling task. They showed that these parameters (age and gender) affect the way how users interpret semantics, and consequently how they label images.

Research on the semantic gap has considered differences between user and computer interpretations of an image, and proposed methods to bridge it, such as introducing various machine learning algorithms [99], using correlations among multiple data modalities (e.g., image, text, meta-data) [116, 117], discovering semantic rules between users and computers [12], and using interactive models [13]. The proposed methods have been verified either by comparing results to reference data, or by measuring the degree of user acceptance in interactive systems. However, since the *gold standard* is set by user created references or user acceptance, user subjective biases are included in this standard. Thus, although these methods result in a narrower semantic gap between computers and users, the linguistic semantic gap remains; therefore, the resulting model for a specific user and search goal may still not be satisfactory to other users. Moreover, in this way, all existing methods proposed

for bringing a system closer to a reference dataset or to a user decision, in principle shorten the semantic gap, although only some authors directly pointed this out in their publications [12, 13, 99, 116, 117]. Moreover, only part of the high improvement achieved by bridging the gap is generalized, the bigger part is subjective and specific to the reference data or to the particular user.

While bridging or shortening the semantic gap, as the difference between the users' image interpretations and that of a computer, has been the focus of many researchers during recent years [113, 118, 119], quantifying the semantic gap has been attempted less frequently [120]. However, designing and developing an efficient image mining system with a narrow semantic gap initially requires quantifying and analyzing the gap. Liu and Song in [120] proposed a method based on information theory for measuring the semantic gap. The authors introduced various aspects of the semantic gap (e.g., the linguistic semantic gap) and listed the possible challenges in developing new methods for measuring the semantic gap.

In this dissertation, we conduct a set of user experiments to study the linguistic semantic gap and its influence on biasing the image mining systems' results [121]. Moreover, we proposed a new method based on an LDA topic model for measuring the semantic gap, as the gap between the users' and a computer's interpretation of images [108].

1.4 Mathematical Models and Theories

In this section, we explain the statistical and information theory-based approaches which are used in this dissertation for analyzing EO image patches. These methods can be generally divided into two main categories, namely *Machine Learning (ML)* approaches (e.g., *Support Vector Machine (SVM)* classification [75], k-means clustering [122], *Latent Dirichlet Allocation (LDA)* topics modeling [18]) and *data coding* approaches based on information theory such as *Huffman Coding (HC)* [123].

1.4.1 Machine Learning Approaches

Machine Learning (ML) approaches are algorithms which can analyze and make predictions based on learning from a given dataset [124]. Various ML approaches have been proposed (e.g., SVM models, clustering models, Bayesian network models) and used for a wide range of data analysis applications such as natural language processing, search engines, and computer vision. According to the degree of human supervision being used during the learning process, the ML approach can be categorized into unsupervised, semi-supervised [125], or supervised methods.

In the context of EO images, ML approaches have been applied to a wide range of image analysis tasks such as change detection, object detection, and land use classification [126]. An overview of the applications of various ML approaches in

EO image analysis can be found in [126]. In this section we briefly introduce and explain the ML approaches which have been used in this dissertation for dealing with EO images. The k-means clustering method [122] is a basic ML approach which is commonly used in EO image analysis [127]. It can cluster objects with similar properties based on their extracted features without human supervision. The unsupervised nature of k-means makes it useful in providing an overview of the structure of the image content when there is no former knowledge of the image content available [127].

SVM models [75] are one of the most frequently used supervised ML approaches to deal with EO images [126]. In an image classification scenario, the goal of SVM is to find a decision plane in the multidimensional space formed by the features extracted from images in such a way to separate the set of objects belonging to different classes. SVM has been used in various scenarios dealing with EO images such as change detection [128], SAR image content understanding [129], and interactive SAR image annotation [82, 130].

Probabilistic topic models are statistical ML approaches which have been originally developed in natural language processing for abstract representations of large collections of text documents by discovering a set of latent topics. The topics are learned completely unsupervised, which allows the indexing of the documents in the collection without any prior knowledge. During the last years, topic models have been adapted to image analysis. Among the various topic models, it has been shown by a previous publication [99] that due to its fully probabilistic behavior, *Latent Dirichlet Allocation (LAD)* [18] provides more descriptive topics than the other topic models such as *probabilistic Latent Semantic Analysis (pLSA)* [100], which are valid even for unseen images. LDA has been applied to various EO image analysis tasks such as change detection [131], image retrieval [132], and semantic image annotation [84, 133, 134].

In the following, we will explain k-means clustering, SVM classification, and LDA topic modeling in more detail.

1.4.1.1 k-means Clustering

One of the most widely used clustering methods for exploratory data analysis is k-means due to its simplicity and fast convergence in practice. It aims to divide a given set of m-dimensional points $X := \{x_1, x_2, ..., x_N\}$ into k ($\leq N$) number of clusters $S := \{s_1, s_2, ..., s_k\}$ in such a way that similar points are grouped into the same cluster. This is done by assigning each data point to the cluster with the nearest center and minimizing the squared distances between the points and the cluster centers:

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{x_j \in s_i} \|x_j - \boldsymbol{c}_i\|^2 , \tag{1.1}$$

where $c_i \in \mathbb{R}^m$ is the center of gravity of the cluster s_i:

$$c_i = \frac{1}{|s_i|} \sum_{x_j \in s_i} x_j. \tag{1.2}$$

The optimization is usually solved by iterative refinement technique which alternates between an *assignment* step, assigning each point to the cluster center with the smallest squared Euclidean distance, and an *update* step, computing new cluster centers using Equation (1.2). It converges when no more change occurs in the assignments.

1.4.1.2 The Support Vector Machine Classifier

Support Vector Machines (SVMs) have been widely used for data classification during the past decade [135]. SVM is usually applied to a supervised or a semi-supervised task, in which a given dataset is split into training and testing sets. Each instance is then depicted as a point, in a vector space. In order to use SVM in our work, each image patch is represented by a vector of its important features using a feature coding method (e.g., BoW). Each point is then assigned a label defining the class to which its corresponding image patch belongs to. The main idea behind this technique is to separate any two data classes by finding a decision boundary between them, which has the maximum distance from all the support vector points.

Since SVMs are binary classifiers, the class labels can only take two values (± 1). However, in many real-world problems the datasets contain more than two classes. In our experiments, we use the software *LIBSVM* developed by Chang *et al.* [136]. In this software, a binary approach has been used which extends the *one-versus-all* method for multi-label classification. To this end, it builds a set of binary classifiers, and each is trained to discriminate one class from the rest. The classifiers are then combined to make a multi-class classifier. In the following, we explain a binary classification of a linearly separable dataset using SVMs which is mostly based on [135] .

1.4.1.2.1 Binary Classification Using SVMs: Suppose that $\mathbb{D} = \{(\vec{x}_i, y_i)\}$ is a set of training data points where $x_i \in \mathbb{R}^m$, $i = 1, ..., N$; is an input feature vector which belongs to either the positive or negative class. The points are then labeled by a vector $y \in \mathbb{R}^N$, where $y_i = \{1, -1\}$. Representing the instances as points in a vector space, SVM then has to find a linear discriminating hyperplane in this vector space. The hyperplane is defined by a normal vector ω called *weight vector* which is perpendicular to the hyperplane. Because many perpendicular hyperplanes to the normal vector can be defined, the interception term b is used to specify the hyperplane. In order to find the linear hyperplane efficiently, a kernel approximation function $\phi(.)$ is used to map the data points to a higher dimensional space. In

our experiments we use a χ^2 kernel [136]. Therefore, every point x_i lying on the hyperplane satisfies the relationship

$$\omega^T \phi(x_i) = -b. \tag{1.3}$$

Consequently, the decision function is as follows which is also known as the *primal* of the classification function:

$$f(x_i) = sign(\omega^T \phi(x_i) + b), \tag{1.4}$$

where *sign* determines the positive or negative label of the point x_i.

SVMs choose ω and b in such a way that the *geometric margin*, a band that separates the support vectors of the two classes, is maximized. When a unit normal vector $\omega/|\omega|$ is used, the geometric margin can be defined by two parallel hyperplanes to the decision boundary as follows:

$$\begin{aligned} \omega^T \phi(x_i) - b &= 1, \\ \omega^T \phi(x_i) - b &= -1. \end{aligned} \tag{1.5}$$

The distance between the two hyperplanes is equal to $\frac{2}{||\omega||}$. Thus, the weight vector should be minimized in order to maximize the hyperplane distance. The square root in $||\omega||$ makes the optimization problem infeasible. Therefore, $||\omega||$ is replaced by $\frac{1}{2}||\omega||^2$ which turns the problem into a quadratic optimization:

$$arg \min_{\omega,b} \ \frac{1}{2}||\omega||^2, \tag{1.6}$$

subject to:

$$y_i(\omega^T \phi(x_i) - b) \geq 1, \tag{1.7}$$

which guarantees that no point lies in the area between the margins and equality is achieved for the support vector points. This constrained problem can be represented by introducing *Lagrange multipliers* λ:

$$arg \min_{\omega,b} \max_{\lambda \geq 0} \{\frac{1}{2}||\omega||^2 - \sum_i \lambda_i [y_i(\omega^T \phi(x_i) - b) - 1]\}. \tag{1.8}$$

In order to represent the classifier as a function of support vectors, the α_i corresponding to the non-support vectors are set to zero. As the derivative of Equation (1.8) with respect to ω and b vanishes, the following constraints are derived:

$$\omega = \sum_i \lambda_i y_i \phi(x_i), \tag{1.9}$$

$$\sum_i \lambda_i y_i = 0. \tag{1.10}$$

Substituting these constraints in Equation (1.8), one can solve the *dual* formulation of Equation (1.8) instead:

$$\max_{\lambda_i}\{\sum_i \lambda_i - \frac{1}{2}\sum_i\sum_j \lambda_i\lambda_j y_i y_j \overline{\phi}(x_i, x_j)\}, \tag{1.11}$$

where $\overline{\phi}(x_i, x_j)$ is a kernel function. In our experiments, we use a *Radial Basis Function* as the kernel:

$$\overline{\phi}(x_i, x_j) = \exp(-\gamma||\phi(x_i) - \phi(x_j)||^2), \quad \gamma > 0, \tag{1.12}$$

where γ is the kernel parameter.

1.4.1.2.2 Soft Margin: Real world datasets may not be linearly separable. To handle these cases, a *soft margin* method has been introduced to find a hyperplane that discriminates points as much as possible by allowing misclassification and incurring a cost depending on how far the misclassified instances are on the wrong side. In the soft margin method, a *slack variable* ξ_i measures the degree of misclassification of x_i. The optimization problem of this method is the trade-off between a large margin and less cost to pay which is formulated as follows:

$$\min_{\omega,\xi,b}\{\frac{1}{2}||\omega||^2 + C\sum_i \xi_i\}, \tag{1.13}$$

subject to:

$$y_i(\omega^T\phi(x_i) - b) \geq 1 - \xi_i, \tag{1.14}$$

where C is the regularization parameter. A smaller C allows a larger margin, whereas a large C means the misclassification highly affects the function. The dual form of the soft margin problem is:

$$\max_{\lambda_i}\{\sum_i \lambda_i - \frac{1}{2}\sum_i\sum_j \lambda_i\lambda_j y_i y_j \phi(x_i)^T\phi(x_j)\}, \ 0 \leq \lambda_i \leq C. \tag{1.15}$$

In the experiments, the best parameters C and γ are determined through cross-validation.

1.4.1.3 The Latent Dirichlet Allocation Topic Model

Latent Dirichlet Allocation (LDA), proposed by Blei *et al.* [18], is a generative statistical model. LDA is considered as a step beyond *probabilistic Latent Semantic Indexing (pLSI)*, presented by Hofmann [137] for probabilistic modeling of text data. Suppose each document $\mathbf{w}_d = \{w_{d1}, w_{d2}, ..., w_{dN_d}\}$ is contained in a text corpus D, where every word-token w_{dn} is drawn from a fixed dictionary of N_V words,

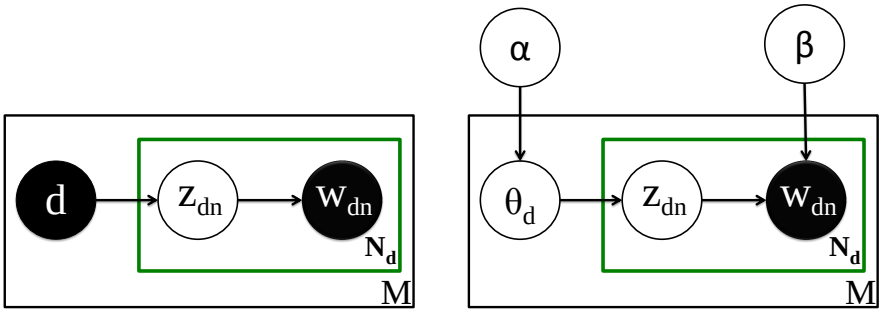
$V = \{v_1, v_2, ..., v_{N_V}\}$. In contrast, pLSI assumes that $\mathbf{w}_d$ is composed of a mixture of K topics, where each topic is defined as a distribution over the words in V. Figure 1.3 (a) shows pLSI as a directed graphical model. In this model, conditional dependencies between variables are represented by arrows and the boxes indicate the repetition of sampling steps, where the number of repetitions is indicated in the corner of each box. This model shows the conditional independence of the observable word-token w_{dn} and the document index d given an unobserved topic z_{dn}. pLSI models the observations as the co-occurrences of the words and documents, where the probabilities of the co-occurrences are obtained by:

$$p(w_{dn}, d) = p(d) \sum_{z_{dn}} p(w_{dn}|z_n)p(z_{dn}|d). \tag{1.16}$$

In this model, $p(z_{dn}|d)$ provides the topic mixing proportions and d is a multinomial random variable, indexing to the documents in the corpus. Therefore, the number of possible values for d is equal to the number of documents and makes pLSI to learn as many mixing proportions as there are training documents. This not only causes a growth in the number of the model parameters by increasing the number of training documents, but also violates the generalizability of the learned model to unseen documents. In order to overcome these shortcomings, LDA creates a probabilistic model in the level of documents by considering the topic mixing proportions as latent variables. Using LDA, the number of model parameters is independent of the training document collection size, and the learned model can be easily applied to unseen documents.

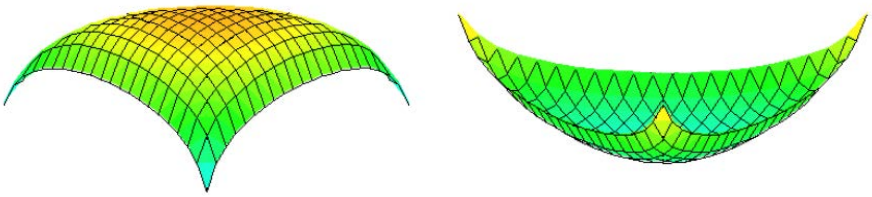
Recently, LDA has been adapted and been used successfully in image analysis by assuming images as mixtures of visual patterns (topics) recurring through the entire corpus [138]. Therefore, we use the generative property of LDA for image representation and image structure learning in this thesis. Figure 1.3. (b) shows a graphical model of LDA. Comparing the two models in Figure 1.3, the prior parameter α which allows LDA to estimate the latent topic mixing weights θ for each document is the main advantage of LDA over pLSA. In the following, we explain the generative process, posterior inference, and parameter estimation of LDA in more details based on the terminology used in our work.

1.4.1.3.1 Generative Process: Using a BoW model, we have a collection of M images, $D = \{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_M\}$, where each image is a sequence of N_d visual words, $\mathbf{w}_d = \{w_{d1}, w_{d2}, ..., w_{dN_d}\}$, and every w_{di} is drawn from a fixed dictionary of N_V visual words. For a short description of an image while preserving the essential statistical relationships of the words, LDA assumes each image as a composition of K topics. To generate each $\mathbf{w}_d$, LDA chooses a K-dimensional *Dirichlet random variable* θ_d (a multinomial distribution with $\theta_{dj} \geq 0$ and $\sum_{j=1}^{K} \theta_{dj} = 1$) from a *Dirichlet distribution* $Dir(\alpha)$, which is a multivariate probability distribution parameterized



(a) (b)

Figure 1.3: (a) Graphical model of pLSI. (b) Graphical model of LDA.



(a) (b)

Figure 1.4: Symmetric Dirichlet distribution for three topics ($K = 3$). (a) $\alpha > 1$. (b) $\alpha < 1$.

by a K-dimensional vector α. Since a Dirichlet distribution is used as a prior and there is no knowledge about the distribution components, a symmetric Dirichlet distribution is considered in which $\alpha_1 = \alpha_2 = ... = \alpha_K = \alpha$. The components of θ_d determine the occurrence probabilities of the K topics in $\mathbf{w}_d$. The Probability Density Function (PDF) of θ_d for a symmetric Dirichlet distribution is:

$$p(\theta_d|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{j=1}^{K} \theta_{dj}^{\alpha-1}, \tag{1.17}$$

where $\Gamma(.)$ is a Gamma function. Figure 1.4 shows the influence of α in a symmetric Dirichlet distribution for three topics ($K = 3$). It illustrates that with $\alpha > 1$ the probability mass is distributed over all the topics which cause LDA to include more number of topics for generating an image model; however, with $\alpha < 1$ the probability mass is concentrated to the corners of the simplex which triggers LDA to consider only a few topics for an image.

In a next step, for generating each visual word-token w_{dn} of the image, LDA chooses a topic-token z_{dn} from the topic mixture θ_d. Then w_n is picked from

$p(w_{dn}|z_{dn}, \beta)$, a multinomial probability distribution conditioned on the selected topic, where $\beta_{N_V \times K}$ is a matrix parameterizing the word probabilities within the topics. Given the parameters α and β, the marginal distribution of $\mathbf{w}_d$ is obtained as:

$$p(\mathbf{w}_d|\alpha, \beta) = \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d)\ p(w_{dn}|z_{dn}, \beta) \right) d\theta_d, \tag{1.18}$$

where $p(z_{dn}|\theta_d) = \theta_{dj}$ when the corresponding topic to θ_{dj} is assigned to the topic-token z_{dn}. The marginal distribution of an image can then be written according to the model parameters as:

$$p(\mathbf{w}_d|\alpha, \beta) = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \int \left(\prod_{j=1}^{K} \theta_{dj}^{\alpha-1} \right) \left(\prod_{n=1}^{N_d} \sum_{j=1}^{K} \prod_{l=1}^{V} (\theta_{dj}\beta_{jl})^{w_{dn}^l} \right) d\theta_d, \tag{1.19}$$

where w_{dn}^l is a vector of size N_V indexing the dictionary V. This vector is set to "1" for the element indexing the selected word v_l and set to "0" for all the other elements.

1.4.1.3.2 Posterior Inference and Parameter Estimation: In the inference step, the posterior distribution of the hidden variables (i.e., the conditional probability of topics) for an observed $\mathbf{w}_d$ is computed as:

$$p(\theta_d, \mathbf{z}_d|\mathbf{w}_d, \alpha, \beta) = \frac{p(\theta_d, \mathbf{z}_d, \mathbf{w}_d|\alpha, \beta)}{p(\mathbf{w}_d|\alpha, \beta)}, \tag{1.20}$$

where $\mathbf{z}_d$ is the set of N_d topic-tokens generating $\mathbf{w}_d$. In the parameter estimation step, the model parameters α and β are computed by maximizing the log likelihood of the images:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} log\ p(\mathbf{w}_d|\alpha, \beta). \tag{1.21}$$

Since both the posterior and the log likelihood are hard to compute, various methods have been introduced to approximate them. For example, Gibbs sampling, Laplace approximation, and variational-based methods for posterior estimation. In our work, we use the LDA implementation of Blei *et al.*[1] which uses a variational *Expectation Maximization (EM)* procedure for posterior and parameter approximation. While in the E-step a variational inference is performed to approximate the posterior, in the M-step, by fixing the approximated posterior, the parameters α and β are estimated by maximizing the lower bound for the log likelihood of the images. For a full description of the EM procedure, we refer the readers to [18].

[1] https://www.cs.princeton.edu/∼blei/topicmodeling.html

1.4.2 Data Coding Based on Information Theory

Information theory, developed by Claude E. Shannon [139], has been widely used in various fields such as natural language processing, cryptography, and data compression as a way to quantify information. In this theory, the principal measure is *entropy* which represents the number of required bits for storing or communicating a symbol of a message. In order to store or transmit information (e.g., text, images), it is usually represented as a sequence of symbols which is sometimes shorter than the original representation, which is called source coding or data compression in digital signal processing [140]. Data compression is categorized into lossy and lossless techniques. While a lossy compression technique identifies unnecessary information and removes it (resulting in an information loss), a lossless compression technique detects redundant information and eliminates it (i.e., it causes no information loss). In the EO domain, data compression techniques such as *Lempel-Ziv-Welch (LZW)* [141] have been used in a number of previous works as an image feature descriptor [142, 143, 144, 145]. Moreover, a compression based similarity measure, the so-called *Fast Compression Distance (FCD)* [146], has been used by previous works for EO image retrieval [143, 145, 147].

Being inspired by the previous applications of information theory and compression techniques in EO image analysis, in this dissertation, we employ *Huffman Coding (HC)* [123] (a source coding technique which is used for the lossless compression of information) in order to code the information provided by local image features. The quantified information is then used for a region-wise comparison of different image patches.

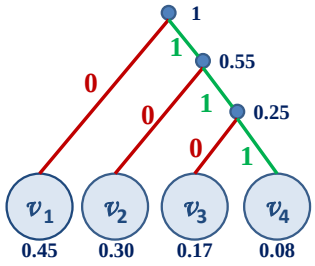
1.4.2.1 Huffman Coding

Huffman Coding (HC), introduced by Huffman [123], is a lossless data compression algorithm based on the entropy of the data symbols (e.g., letters in a text document). This algorithm builds a variable-length code table containing strings of binary code words (i.e., strings of "0" and "1"), where the lengths of the code words correspond to the occurrence probability of the symbols. More precisely, the most common symbols are expressed by short binary strings whereas the less common symbols are assigned longer strings.

In our work, we adapt and use Huffman coding for image analysis using a BoW model of images. Assume the BoW model of image $\mathbf{w}_d$ is denoted as a sequence of N_d visual words, $\mathbf{w}_d = \{w_1, w_2, ..., w_{N_d}\}$, where every w_{dn} is drawn from a fixed dictionary of N_V visual words, $V = \{v_1, v_2, ..., v_{N_V}\}$. Thus $\mathbf{w}_d$ can be represented as the occurrence probabilities of every v_j, $\mathbf{w}_d = \{p(v_1), p(v_2), ..., p(v_{N_V})\}$. In this image representation, each v_j is assumed as a visual symbol. The goal of Huffman coding is then to build a table of code words $Q = \{q_1, q_2, ..., q_{N_V}\}$ (code word q_j corresponds to the visual symbol v_j) in such a way that the weighted average code

v_j	$p(v_j)$	q_j
v_1	0.45	0
v_2	0.30	10
v_3	0.17	110
v_4	0.08	111

(a)



(b)

Figure 1.5: (a) Table representing visual words, their occurrence probabilities, and the code word assigned to them by Huffman coding. (b) Huffman binary tree created based on (a).

length $L(Q)$ is minimized:

$$L(Q) = \sum_{j=1}^{N_V} p(v_j).length(q_j). \tag{1.22}$$

Figure 1.5 shows an example of creating a Huffman tree and its code table. In order to generate the code words, the Huffman algorithm creates a binary tree in which initially all the symbols are positioned in the leaf nodes weighted by their occurrence probabilities. The tree is created from right to left by taking the two least probable leaf nodes and putting them together to form a parent node with a probability that equals the sum of the two child nodes. Then the left and the right branches are assigned "0" and "1", respectively. The same procedure is repeated for the leaf nodes and the new parent nodes until $N_V - 1$ parent nodes are generated. Finally, the sequence of the binary values in the path from the root node to each leaf node is assigned as a binary code to the symbol thus generating the code table. Every visual word-token w_{dn} is then assigned a code word q_j instead of v_j. This results in a binary representation of the image.

1.5 Contributions

The main contributions of this dissertation can be summarized in five points:

- **Image feature coding:** We propose an efficient feature coding strategy to provide a compact but descriptive BoW representation of image features.
- **Semantic level image description:** We introduced a new approach based on *Latent Dirichlet Allocation (LDA)*, a topic model, to represent *Earth Observation (EO)* image patches by semantically meaningful features.

- **Feature descriptor evaluation:** We propose an approach based on information theory using a *communication channel* model for measuring the information quantity that various feature descriptors extract from a given image collection and deliver to image mining systems, regardless of user labeling. Furthermore, we propose a novel approach based on *Huffman Coding (HC)*, a lossless compression technique, to measure the overlaps between the information obtained by different feature descriptors.

- **The sensory and semantic gaps:** We conduct user studies for assessing the issues limiting user understanding of EO image semantics, namely the *sensory* and *semantic* gaps. In addition, we conduct computational evaluations in order to study the difference between user image understanding and the interpretation of the images by a computer. Moreover, we propose a method for measuring the semantic gap based on an LDA model.

- **Feature space evaluation and exploration:** We propose a clustering-based approach to evaluate the semantics of EO image patches based on the descriptions of their extracted features. Additionally, we show the importance of the visualization and exploration of images based on their extracted features for better understanding of their semantics.

1.6 Thesis Overview

Chapter 2 describes our proposed efficient feature coding strategy. Chapter 3 introduces a new LDA-based approach for modeling EO image patches by semantically meaningful features. Chapter 4 proposes an approach based on communication channel model for measuring the quantity of the extracted information from images by various feature descriptors. Chapter 4 further proposes a novel HC-based approach for quantifying the overlaps between the information obtained by different feature descriptors. Chapter 5 assesses the causes of the sensory gap in EO images by means of a human perceptual evaluation and a computational evaluation. Moreover, this chapter proposes methods to quantify and study the two main types of the semantic gaps, the gap across users; and the gap between users and computers. Chapter 6 introduces a clustering-based approach for evaluating the semantics of EO image patches based on their extracted features. Moreover, this chapter discusses the importance of the visualization and exploration of images based on their extracted features for better understanding of image semantics. Chapter 7 summarizes and concludes this dissertation and discuss directions for future research.

2

Efficient Feature Coding

For many years, the *Bag-of-Words (BoW)* approach has been used to describe image content. This approach extract local primitive features (i.e., feature vectors) from candidate images, and code them via a codebook to integrate each image's local features into a single feature vector. The coded feature vectors are then categorized and annotated by semantic labels. This basic approach and a number of alternatives have been described by many publications [24, 83, 5, 84]; their authors tacitly assumed that all extracted features have to be coded correctly and efficiently. During the past years, developing efficient coding strategies have been addressed by many previous researches [87, 88, 89, 90, 91, 92, 85]. However, there is still place to improve the developed coding strategies which as will be shown below, can lead to promising classification results.

In this chapter, we propose a new *Locally Linear Salient Coding (LLSaC)* feature coding strategy. It considers each feature point in relation with its neighboring points and codes the local interrelationships of the feature space instead of coding every feature point individually. This allows LLSaC to discover the global structure of the feature space even when applying small codebooks. Therefore, LLSaC helps image mining systems to avoid the drawbacks of using large codebooks such as excessive storage requirements, the curse of dimensionality (which increases the computational effort), and their limited degrees of freedom.

2.1 Locally Linear Salient Coding

Recently, the BoW approach has shown promising results for the modeling of EO images [24, 83, 5, 84]. In this technique, feature coding plays the key role and has a significant impact on both accuracy and speed of image mining systems [85]. A variety of possibilities for codebook generation and different ways how to compute the code values have led to the development of several feature coding strategies in recent years.

The existing feature coding methods such as *Hard Voting (HV)* [87], *Soft Voting (SV)* [88], *Sparce Coding (SC)* [89], *Local-constraint Linear Coding (LLC)* [91], and *Salient Coding (SaC)* [92] perform poorly for small codebooks [92]. Since the existing methods code each feature point independently from the other points, a small number of the codewords relative to the number of the feature points limits them in modeling the structure of the entire feature space. Increasing the codebook size significantly improves their performance in modeling the feature space structure; however, this introduces problems such as considerable storage requirements, the curse of dimensionality (which increases the computational effort), and the limited degrees of freedom [148]. Considering these issues, any coding strategies which achieve high accuracies with small codebooks support the scalability of learning systems and, therefore, are highly desired. In order to solve the problem, we propose an LLSaC concept which uses the relationships between the feature points as additional information thus assisting the feature space representation by a small number of codewords. As a result, LLSaC codes the local structures of the feature space instead of coding each feature point individually, where each local data structure is defined by a set of linear coefficients reconstructing a feature point from its neighboring points. The coefficients are then used to update the response of the original feature point to the codewords. In order to compute the feature point responses, LLSaC employs the saliency-based method introduced by Huang *et al.* [92]. The authors of [92, 85] showed that using the saliency information of the feature space provides promising classification accuracies.

When we compare LLSaC to SaC and other coding strategies such as HV, SV, and LLC on Fifteen Natural Scenes multimedia dataset (for details, please refer to Section B.4), it turns out that LLSaC significantly outperforms the other techniques even for small codebooks. In addition to this multimedia dataset, we could show that LLSaC also ranks first when applied to EO images, which verifies the generalizability of our proposed method.

2.1.1 Salient Coding

Huang *et al.* [92] proposed SaC based on the idea that saliency is a fundamental characteristic of a feature space and codebook-based coding strategies. SaC has been developed initially to code the saliency information of feature points by considering the relative distances of each feature point to its $\hat{K}$ nearest codewords. In this strategy, the codeword which is closer to a feature point compared to the other codewords can strongly describe the feature point independently from the other codewords. The salient response η_{ik} of the feature point x_i to the codeword v_j is then obtained by:

$$\eta_{ij} = \begin{cases} \Psi(x_i) & if\ j = \underset{j}{\operatorname{argmin}}(\|x_i - v_j\|_2) \\ 0 & else \end{cases}, \tag{2.1}$$

$$\Psi(x_i) = \Phi(\frac{\sum_{v_t}(\|x_i - v_t\|_2 - \|x_i - v_j\|_2)}{\sum_{v_t}\|x_i - v_t\|_2}),\ v_t \in N(x_i)\ \wedge v_t \neq v_j, \tag{2.2}$$

where $\Phi(.)$ is a monotonically decreasing function and $N(x_i)$ is a set of the $\hat{K}$ nearest codewords to the feature point x_i. According to [92], we use $\Phi(x) = 1 - x$ in our experiments for the sake of normalization simplicity. Since according to Equation 2.1 each feature point only responds to its nearest codeword, SaC is considered as a hard assignment strategy [93]. The final feature vector is then generated through maximum pooling [86].

2.1.2 Linear Representation of Non-linear Structures

Using linear coefficients to represent the non-linear structure of data has been introduced first by Roweis and Saul [94] in their proposed neighborhood preserving dimensionality reduction method, the so-called *Locally Linear Embedding (LLE)*. The idea is that every original data point $x_i \in \mathbb{R}^m$ can be reconstructed by a linear combination of its neighboring points $x_l \in \mathbb{R}^m$, given a set of weights $u_{il} \in U$. To compute the weights that best reconstruct the data points, the following cost function is minimized,

$$E(U) = \sum_i |x_i - \sum_l u_{il} x_l|^2, \tag{2.3}$$

where u_{il} determines the contribution of x_l in reconstructing x_i. Thus, each row of the matrix U should sum to one, $\sum_l u_{il} = 1$. Moreover, in order to allow only the contributions of the immediate neighbors, for every non-neighboring point $u_{il} = 0$. The optimal weights are obtained in closed form by solving a least squares problem. For more details about computing optimal weights, the reader is referred to [149].

2.1.3 Methodology

The main idea behind LLSaC is that if a feature point is reconstructed from its neighbors, the original feature point response to codewords could also be reconstructed from the neighboring point responses. Figure 2.1 shows the framework of our proposed method. In feature point space, the reconstruction weights u_{il} ($i = 1$ and $l \in [2, 4]$ in Figure 2.1) are computed based on the linear reconstruction of the point x_1 from its neighbors. The reconstruction weights are then passed to the code space. There, the salient response η_{ij} of x_1 to the codeword v_j (computed in the same way as for SaC) is updated by the weighted average of the neighboring points'

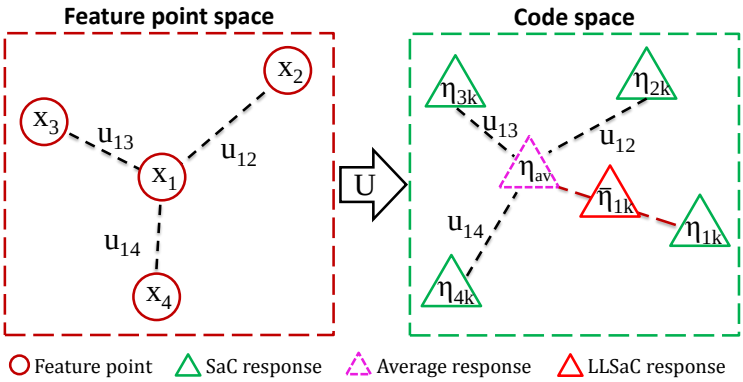


Figure 2.1: Framework of the LLSaC method. For every feature point, the reconstruction weights are computed in the feature point space. The weights are further used to update the feature points' salient responses in the code space.

salient responses η_{av}. In this weighted averaging, the responses are weighted by their corresponding reconstruction weights obtained from feature space. η_{av} is computed as follows:

$$\eta_{av} = \sum_{l} u_{il}\eta_{lj}, \;\; x_l \in NP(x_i), \tag{2.4}$$

where $NP(x_i)$ is the set of $\bar{K}$ nearest feature points to x_i. The salient response η_{ij} of x_i is then updated by:

$$\bar{\eta}_{ij} = \frac{1}{2}(\eta_{ij} + \eta_{av}), \tag{2.5}$$

where $\bar{\eta}_{ij}$ is the updated salient response. After updating all the salient responses, they are integrated to form the final image feature descriptor using maximum pooling.

2.1.4 Results and Discussion

In our experiments, we compared LLSaC to SaC and other coding strategies such as HV, SV, and LLC. In order to be consistent and comparable with previous feature coding articles (e.g., [92], [85], [150]), the coding toolkit developed by Chatfield *et al.* [150] was used. Moreover, in order to compare LLSaC with SaC more precisely, the experiments are run on the Fifteen Natural Scenes dataset and the results are compared to the results reported in the original SaC article [92]. In addition, in order to show the generalizability of our proposed method to EO scenarios, the performance of LLSaC is compared to SaC on the UC Merced LandUse dataset.

For consistency with previous works, the 128-dimensional *Scale-Invariant Feature Transform (SIFT)* descriptors [70] are extracted densely for every 4 pixels. In order to provide a richer description of the image features, SIFT is extracted for three scales: 16×16, 24×24, and 32×32 pixels. Then, k-means clustering is

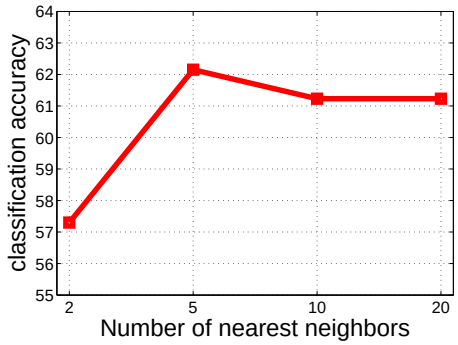


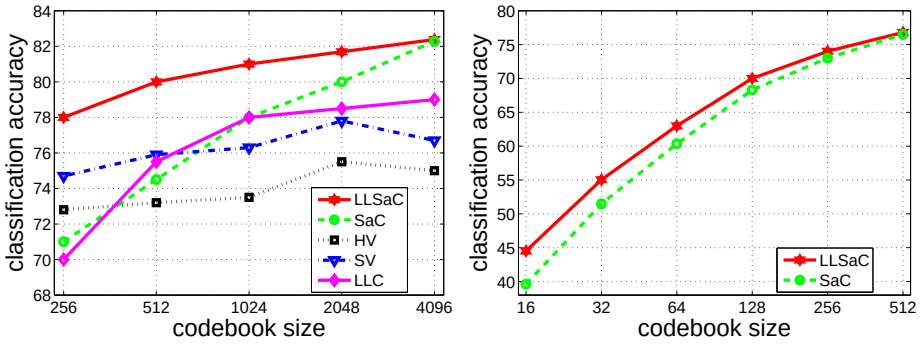
Figure 2.2: Performance of LLSaC on the Fifteen Natural Scenes dataset for different numbers of nearest neighbors $\bar{K}$.

applied to the samples of the local feature points to generate codebooks of various sizes. In the coding step, according to [92], the number of nearest codewords $\hat{K}$ is set to 5 for LLSaC, SaC, and LLC. In contrast, for LLSaC, the number of neighbors which reconstruct the feature points is fixed to $\bar{K} = 5$ based on an empirical study. We study the influence of $\bar{K}$ on the performance of our proposed method in the classification of the Fifteen Natural Scenes dataset. Figure 2.2 shows the performances for $\bar{K} \in \{2,\ 5,\ 10,\ 20\}$ with a codebook size of 16. The figure indicates that the small number of neighbors cannot provide enough information about the structure of the data. However, using too many neighbors affects the locality of the reconstruction weights.

In order to compare the performances of the coding strategies, they are used to classify our test images using a *Support Vector Machine (SVM)* classifier [136]. The setup parameters of SVM such as cost and gamma are set to 1 according to [92]. Then the classification accuracies are reported for various codebook sizes. For each codebook size, the experiments are run 10 times and the average result is presented.

Figure 2.3. (a) shows the classification accuracies for LLSaC, SaC, HV, SV, and LLC. As the graph shows, LLSaC outperforms all the other methods under various codebook sizes. Moreover, as the dictionary size increases, the performance of LLSaC converges to that of SaC. Since LLSaC provides the codewords with the responses from local structures of the feature space, even a small number of codewords can discover the global structure of the feature points. Therefore, LLSaC is more robust when the codebook size is modified. However, as the number of codewords increases significantly (to 80% of the number of local feature points in each image), they can represent the structure of the data with no need for additional information from local structures. Consequently, LLSaC and SaC perform similarly for a large number of codewords.

Figure 2.3. (b) indicates that LLSaC also outperforms SaC on the UC Merced Land Use dataset. Since the feature points of various image types (e.g., multimedia, EO) have different topologies in feature space, the higher performance of LLSaC for



(a) (b)

Figure 2.3: (a) Performance comparison of LLSaC and other coding strategies for different codebook sizes on the Fifteen Natural Scenes dataset. The results for SaC, HV, SV, and LLC are taken from [92]. (b) Performances of LLSaC and SaC for different codebook sizes on the UC Merced Land Use dataset.

both multimedia and EO images verifies its generalizability.

2.1.5 Summary

In this section, we proposed LLSaC, a new variant of SaC. This method remedies the limitation of SaC in representing the structure of feature spaces by small codebooks. Since small codebooks increases the scalability of the applied learning systems, the development of methods which perform well when using small codebooks, is highly demanded. LLSaC discovers the global structure of the feature space by exploiting local linear reconstructions of the feature points. This knowledge is then used to update the salient responses which are computed as in SaC. Our experimental results indicate that LLSaC significantly outperforms other coding strategies in describing the feature space structure even with a low number of codewords for both multimedia and EO datasets.

3

Semantically Meaningful Image Descriptors

Most image descriptors being used for image classification and understanding (such as extracted feature vectors) consist of a sequence of numerical coefficients that often have no directly visible relationship with the semantical meaning of a given image patch to be annotated. To remedy this situation, we propose a new approach to model EO image patches by semantically meaningful features. Since users usually understand visual data and categorize them based on their local contexts (detected objects, parts thereof, and their spatial and temporal neighborhoods), a semantically meaningful image description brings image mining systems closer to the image understanding of human users, which makes their results more satisfactory. In our approach, we apply *Latent Dirichlet Allocation (LDA)* [18] to *Bag-of-Words (BoW)* models of the image patches in order to analyze the statistical relationships between their visual words. The relationships are further used to discover the semantic structure behind the entire image patch dataset as a set of topics (contexts). The image patches are then modeled by mixtures of the topics, which we call a *Bag-of-Topics (BoT)* model.

3.1 The Bag-of-Topics Model

In recent years, BoW has been successfully adapted and applied to EO image understanding by analyzing adjacent image patches [83, 5, 84]. In principle, BoW assumes each image patch as a combination of visual words regardless of the statistical relationships between them. However, it has been shown in previous works [96, 97, 98, 99] that by considering the statistical relationships between the visual words, the image descriptions become more semantically related and meaningful.

In this section, we thus propose a model which uses the statistical relationships of the visual words in order to describe EO image patches by semantically meaningful

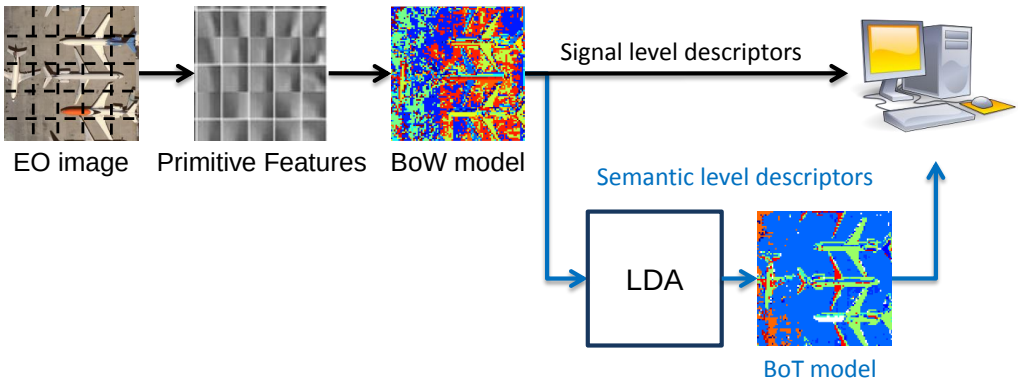


Figure 3.1: Example of using signal and semantic level descriptors in an image mining system.

features (i.e., describing their spatial and temporal contexts). Figure 3.1 shows an overview of our proposed model. In this model, primitive features extracted from EO image patches are represented using local primitive feature descriptors such as *Mean/Variance (MV)* and selected Gabor transform coefficient. Then each image patch is modeled as a BoW. In a next step, LDA is applied to the BoW model of the image patches in order to discover the existing contexts (topics) of the entire dataset. Finally, each image patch is represented as a combination of the topics (BoT model). Figure 3.2 shows the BoT and BoW models of some multispectral and SAR image patches. The samples show that BoT describes image patches with a few but semantically understandable features; however, the visual words in the BoW model hardly reflect any human understandable semantics. For example, in Figure 3.2. (c), BoT describes the baseball diamond clearly by meaningful topics such as grass and sand, while the meaning of the visual words in the BoW model are hard to understand. The illustrations in Figure 3.2. (j to r) show that understanding the semantics of visual word is even more difficult for SAR data. In Figure 3.2. (j), for example, the BoW hardly represents any structure; however, a highway and its neighboring areas can be recognized in the BoT model. Since an image description using BoT is closer to user image understandings, an image mining system which employs BoT image models provides more relevant and satisfactory results to the user queries.

In order to evaluate the BoT model approach on EO images, a set of experiments were conducted on a multispectral image patch dataset and a SAR image patch dataset. In these experiments, a classification method (e.g., a *Support Vector Machine (SVM)* [136]) is applied to the BoT and BoW representations of images. The accuracies and the run-times of the classifications are then compared for the two models. Experimental results demonstrate that BoT provides a compact representation of the images; in addition, it either causes no significant reduction, or in many cases, even increases the classification performance. While a compact rep-

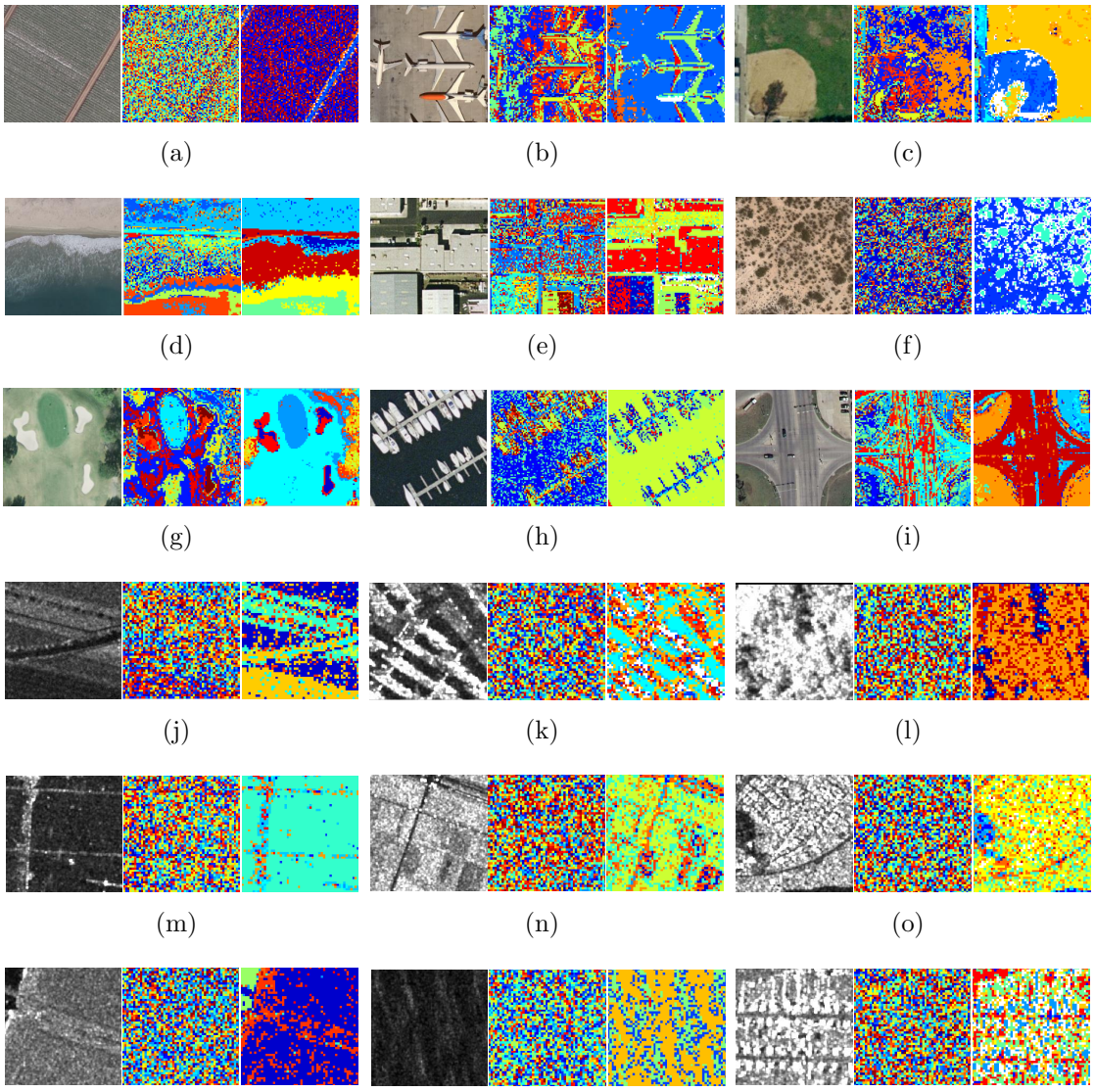


Figure 3.2: BoW and BoT representations of samples of multispectral (a to i) and SAR (j to r) EO images. For each sample, the first image is the original image. The second and the third images are BoW and BoT representations of the image, respectively. The various colors depict the visual words (in BoW) or the topics (in BoT). Dictionaries of 200 visual words generated from MV features are used. The BoT models of the images are made for 20 topics. (a) Agricultural field. (b) Airplanes. (c) Baseball diamond. (d) Beach. (e) Buildings. (f) Chaparral. (g) Golf course. (h) Harbor. (i) Intersection. (j) Highway. (k) Urban area. (l) Forests. (m) Flooded areas. (n) Agricultural field. (o) Agricultural field. (p) Urban area. (q) Water surface. (r) Urban area.

resentation improves the scalability of the image mining systems by decreasing the computational effort; further, topics are more discriminative than visual words.

3.1.1 Methodology

When using a BoW model, each image patch $\mathbf{w}_d$ is defined as a sequence of N_d visual words, $\mathbf{w}_d = \{w_{d1}, w_{d2}, ..., w_{dN_d}\}$. In $\mathbf{w}_d$ every w_i is drawn from a fixed dictionary of N_V visual words, $V = \{v_1, v_2, ..., v_{N_V}\}$, generated by applying k-means clustering to the sequence of the primitive image feature vectors (extracted by primitive feature descriptors such as MV and Gabor) for N_V clusters. LDA then discovers the latent structure of the entire image patch dataset as a set of K topics by approximating the posterior $p(\theta_d, \mathbf{z}_d|\mathbf{w}_d, \alpha, \beta)$ in an inference step and learning the β in a parameter estimation step (for more details please refer to Section 1.4.1.3). While $\beta_{N_V \times K}$ is a matrix representing the probability of each v_l in every topic, the posterior provides each $\mathbf{w}_d$ with a topic mixing weight θ_d (a multinomial distribution with K possible outcomes). In addition, we manually set the Dirichlet parameter α in our experiments. By using topics to describe each image patch, BoT considers each topic mixture as a vector in a K-dimensional Euclidean space (feature space). Since the number of topics is usually much smaller relative to the number of visual words ($K \ll N_V$), the feature space created by topics has a lower dimensionality than the one created by visual words. In other words, BoT provides a more compact description of image patches which helps to increase the scalability of image mining systems and reduces their computing time.

3.1.2 Results and Discussion

In our experiments, we assessed the semantic-level descriptions of a multispectral and of a SAR image dataset, namely the UC Merced Land Use and the Fifteen Class TerraSAR-X Image Patches datasets defined in Section B.3 and Section B.2, respectively. As a first step, we extract local primitive features using MV and Gabor feature descriptors from each image patch. MV descriptors are two-dimensional vectors representing the mean and the variance of every local image point. The local feature descriptors are computed using a non-overlapping sliding window of 3×3 pixels. Furthermore, we compute Gabor feature descriptors from sliding widow of 32×32 pixels with 50% overlap. In this descriptor, we set the scale parameter $S = 3$ and the rotation parameter $R = 6$ which results in feature vectors of 36 elements. Moreover, for the multispectral images, the feature vectors are computed for each individual color channel and concatenated later to form the final feature vector. After extracting the local primitive features, each image patch is represented by a BoW model for various dictionary (codebook) sizes (50, 100, 200, and 300 codewords). To generate the dictionaries, k-means clustering is applied to 10% of the feature vectors, selected randomly, where the cluster centers are considered as

visual words. The final model is derived as a histogram by assigning the feature vectors to their nearest visual words using a *hard voting (HV)* coding strategy.

In the next step, LDA is applied to the BoW models in order to discover the latent structure behind the entire dataset as a set of topics. Each image patch is then described as a mixture of the topics (BoT model). The semantic level of the topics is usually correlated to the number of topics discovered by LDA. More precisely, a small number of topics leads to general concepts (e.g., forest, urban area), while a larger number of topics provides more detailed contents (e.g., trees, buildings). Evaluating various numbers of topics allows us to assess the effects of different semantic levels in discriminating image classes. Since the resulting topics are not unique, we run LDA three times for each experiment and average over the final results obtained by the three sets of topics.

In order to assess the value added by BoT to the BoW model, the performance of SVM in the classification of both representations of the EO datasets is measured. In order to generalize the task, we select randomly from every class 70 samples for training, 20 samples for parameter optimization, and the remainder for testing. The results are cross-validated by running 10 experiments and repeating each experiment 10 times. Finally, the accuracy and run-time are averaged over the experiments.

Figures 3.3 and 3.4 show the classification accuracies and the run-times. In these figures, the plotted horizontal lines depict the classification results obtained using the BoW models of MV and Gabor feature descriptors. In addition, the plotted curves illustrate the classification results using BoT for different numbers of topics. Moreover, the columns in the figures represent the results for various dictionary sizes. Since the number of topics is usually smaller than the number of visual words, using BoT allows a compact representation of the data; it either causes no significant reduction in the performance or increases the classification accuracy by a sufficient number of topics. The BoT model can increase the discriminability of the descriptors, because contexts represented by topics are usually more descriptive than the contents described by visual words. For example, in Figure 3.3 (a), for 50 topics and for MV feature descriptors, BoT outperforms BoW; with a similar size of the BoW and BoT feature vectors, topics provide a more discriminable representation of the data. Moreover, BoT speeds up the classification by compacting the data representation. In Figure 3.3 (d), for example, the BoT (with 60 topics) built using the BoW (with 300 visual words) of MV feature descriptors obtains a higher accuracy than the BoW model and it is 15 times faster during classification. According to Figure 3.4, for SAR data, BoT performs similarly to BoW; however, it is much faster and, therefore, more efficient than BoW.

Furthermore, comparing the two primitive feature descriptors indicates that the discriminability of the topics depends on the informativeness of the BoW model built upon primitive feature descriptors. For example, in Figures 3.3 and 3.4, the discovered topics from MV are more discriminable than the Gabor topics, because the BoW model of MV features are more discriminable than the BoW model of the

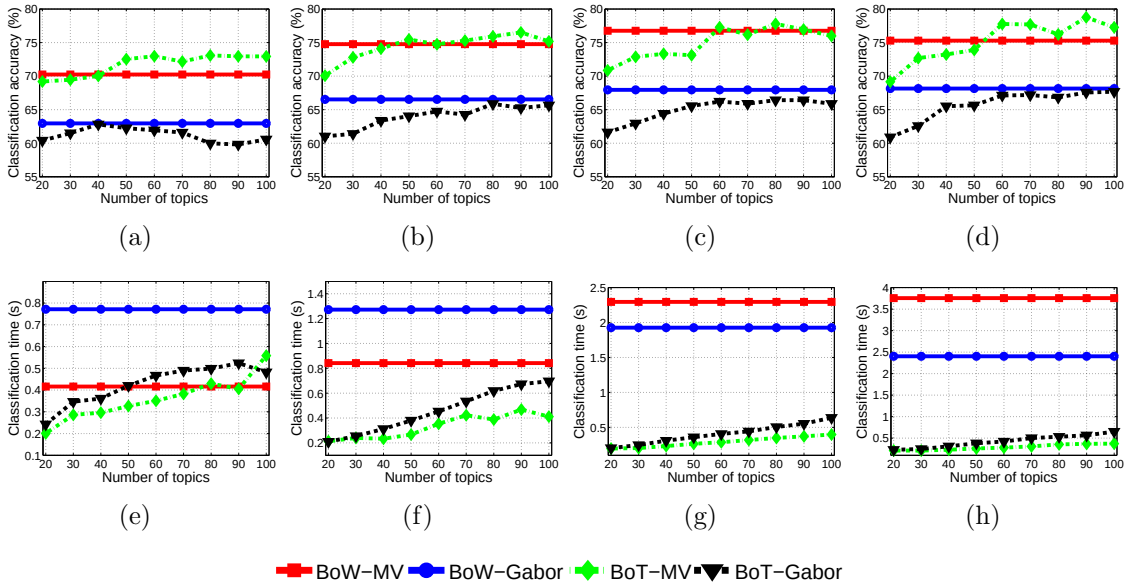


Figure 3.3: Accuracy and run-time of the classification using BoW and BoT models for various dictionary sizes and different numbers of topics. In these experiments, SVM is applied to the UC Merced Land Use dataset. (a) Dictionary size = 50. (b) Dictionary size = 100. (c) Dictionary size = 200. (d) Dictionary size = 300. (e) Dictionary size = 50. (f) Dictionary size = 100. (g) Dictionary size = 200. (h) Dictionary size = 300.

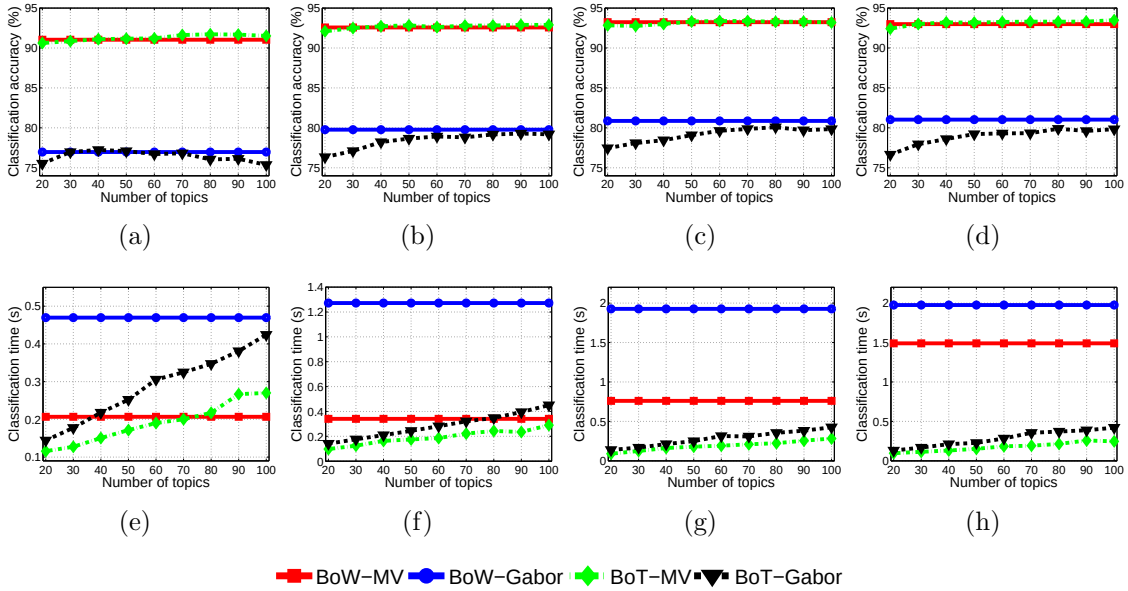


Figure 3.4: Accuracy and run-time of the classification using BoW and BoT models for various dictionary sizes and different numbers of topics. In these experiments, SVM is applied to the Fifteen Class TerraSAR-X Image Patches dataset. (a) Dictionary size = 50. (b) Dictionary size = 100. (c) Dictionary size = 200. (d) Dictionary size = 300. (e) Dictionary size = 50. (f) Dictionary size = 100. (g) Dictionary size = 200. (h) Dictionary size = 300.

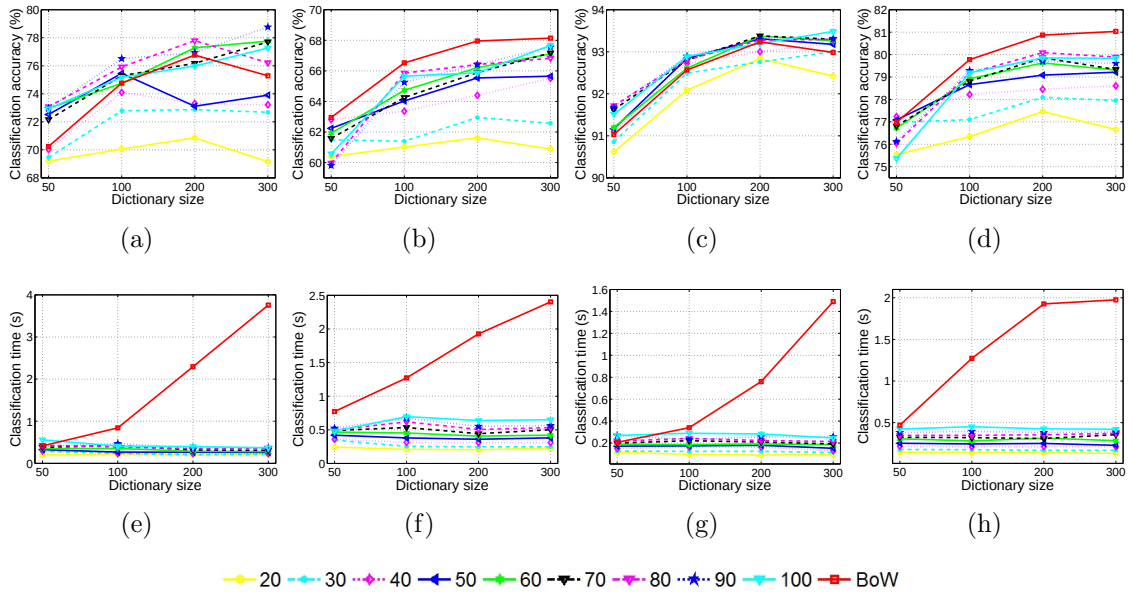


Figure 3.5: Assessing the dictionary size in topic discriminability by comparing the classification accuracies and run-times for the UC Merced Land Use and the Fifteen Class TerraSAR-X Image Patches datasets. (a) MV and multi-spectral. (b) Gabor and multi-spectral. (c) MV and SAR. (d) Gabor and SAR. (e) MV and multi-spectral. (f) Gabor and multi-spectral. (g) MV and SAR. (h) Gabor and SAR.

Gabor features, which is indicated by the higher classification accuracies of the MV descriptors comparing to those of the Gabor descriptors. The outperformance of the MV descriptors relative to Gabor descriptors in our experiments could be due to using a smaller neighborhood during extraction of the MV primitive features. The high resolution of the EO images used in our experiments provides details of the objects; therefore, to discriminate the object classes, feature descriptors with more locality perform better.

Figure 3.5 shows how the dictionary size affects the visual words and the topics generated from the two datasets. It shows the classification accuracies and run-times versus dictionary size for the BoW model (the red solid curve) and the BoT models for various numbers of topics. As the results show, the performances usually improve sharply for small dictionary sizes, but they decrease for larger sizes. Furthermore, increasing the dictionary size brings about a higher dimensionality of the BoW descriptors which causes the run-time to increase dramatically. Since a larger number of visual words helps LDA to discover more descriptive topics, this leads to more discriminable descriptors; increasing the dictionary size usually speeds up the classification using BoT.

3.1.3 Summary

In this section, we applied LDA to the BoW model of two EO image patch datasets to discover their context as a set of topics. Then, the image patches are described as a mixture of the topics (BoT model). The BoT approach can be used in various learning scenarios such as image classification and retrieval. In our work, BoT is evaluated in image classification by applying an SVM classification to the BoT models of image patches. The results are then compared to the accuracies achieved by the BoW model. Experimental results demonstrate that the BoT model can provide comparable results to those of the BoW model; however, the description of data is much more compact in the BoT model. Consequently, BoT not only increases the scalability of image mining systems, but also discriminates various image classes to a higher degree. As a result, we show the effects of different numbers of BoT topics on the classification performance. However, the selection of an optimized number of topics still deserves more detailed investigations. Moreover, since the BoT model builds upon the BoW model, the improvements in the efficiency of the BoW model such as using LLSaC strategy (which we proposed in Section 2.1) scales up the discriminablity of the topics in the BoT model.

4

Evaluation and Comparison of Feature Descriptors

In this chapter, based on information theory, we propose a method for evaluating the descriptiveness of the extracted features from EO image patches using various feature extraction methods, and a method for quantifying the degree of similarity of the provided information by pair-wise comparison of feature descriptors.

The first method evaluates and compares the informativeness of feature descriptors for image mining systems. The method models an image mining system (e.g., a system based on *Latent Dirichlet Allocation (LDA)* topic model [18]) as a communication channel. In this model, image patches represent the source, the topics discovered by LDA stand for the receiver, and the feature descriptors are considered as the information carriers. The channel mutual information is then computed as the informativeness measure. We show that the channel mutual information computed for each feature descriptor correlates with the discriminability of the image patches represented by that descriptor.

The second proposed method uses *Huffman Coding (HC)* [123] to measure the similarities between the information obtained by different feature descriptors, which we call *information overlap*. The analysis of information overlap, especially for feature descriptor fusion tasks, allows a more compact but still comprehensive image representation by selecting a lower number of more distinct feature descriptors. Moreover, this method measures the information overlap for each image patch individually which makes it independent of user image labeling.

4.1 Feature Evaluation Based on a Communication Channel Model

In this section, we propose an approach to quantify the amount of information a feature descriptor provides to image mining systems. To this end, we proposed

a new formulation of *Latent Dirichlet Allocation (LDA)* topic model, proposed by Blei *et al.* [18], as a *communication channel* model, introduced by Shannon [139]. In our proposed communication channel model, the inputs are images, the outputs are the provided results, and the carriers are feature descriptors. The mutual information carried by the feature descriptors (from input to output) is then computed as the information quantity. The idea behind using LDA is that it can automatically discover the latent semantic structure of an image dataset due to its probabilistic generative behavior. Thus, computing the channel's mutual information allows us to assess the ability of a primitive feature descriptor in providing semantic information of images.

In our experiments, LDA is applied to the UC Merced Land Use image patch dataset, which discovers the latent semantic structure of the dataset a set of topics. In addition, in order to represent the image patches, we use the following primitive feature descriptors: *RGB Color Histogram (rgbHist)*, *Weber Local Descriptor (WLD)* [3], *Scale-Invariant Feature Transform (SIFT)* [70], *WLD-Color*, and *SIFT-Color* as defined in Appendix A. These descriptors allow us to study the content of an image dataset from various aspects, namely color, texture, shape, and the combinations of color with texture and shape. Figure 4.1 shows a graphical model of the LDA components and their interconnections as a communication channel. In this model, the discovered topics are considered as the channel outputs. Since each feature descriptor represents a particular image property, the computation of the mutual information for each descriptor yields the quantity of that property within the image patches. We then compare the computed mutual information to the accuracy of a supervised classification method such as *Support Vector Machine (SVM)*. The results indicate that the mutual information quantity provided by a feature descriptor is proportional to its performance in image mining tasks such as classification.

4.1.1 Modeling LDA as a Communication Channel

Information theory developed by Claude E. Shannon [139] provides mathematical models to quantify information. It models the relation between variables as a communication channel, and computes the amount of information transferred between input and output variables via the channel. In this model, the two basic measures are *entropy* and *mutual information*. Entropy measures the amount of uncertainty of the variables and mutual information measures the amount of shared information between the input and output variables. In this section, we use this model to quantify the amount of information received by an image mining system from an image collection. In our method, LDA is used to emulate an image mining system. Using information theory, the authors of [120] quantify the amount of information provided by an image collection to an image retrieval system, by computing the entropy of the images. In this method, although entropy quantifies the amount of information

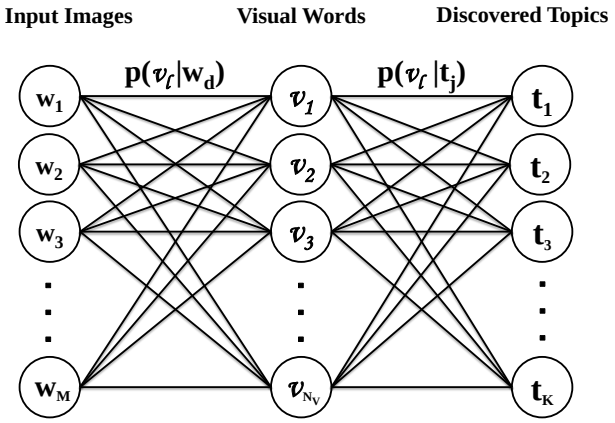


Figure 4.1: LDA is modeled as a communication channel. Input is the BoW model of images, output is the discovered topics, and the information is carried by feature descriptors.

of the images, the impacts of the transmission process on the transmitted information from the images to the retrieval system are neglected. In order to explain this shortcoming, we model the structure of LDA as a communication channel, as shown in Figure 4.1. In this model, we consider the given images as input, the topics discovered by LDA as output, and the feature descriptors as information carriers. Then we compute the channel's mutual information which defines the information quantity received by the output from the input, considering the transmission process. We suppose that the discovery of topics by LDA is highly dependent on the mutual information of the channel, which will be shown in the following sections by the experimental results.

In our proposed communication channel model, the input image patches are coded based on their properties. In order to analyze various image properties, we use different primitive feature extraction methods (e.g., rgbHist, WLD, SIFT, WLD-Color, SIFT-Color). Then by applying k-means clustering to the extracted feature vectors for N_V clusters, we generate a codebook of N_V visual words, $V = \{v_1, v_2, ..., v_{N_V}\}$. In a next step, each image patch $\mathbf{w}_d$ is represented by a BoW model, where each visual word w_{di} is drawn from the visual word codebook V. Then the probability of every visual word v_l in each image patch $p(v_l|\mathbf{w}_d)$ is obtained by counting the occurrences of the visual words. Thus, each image patch is represented as a probability mass function over the visual words. According to Figure 4.1, the visual word representations (the BoW models) are transmitted through the channel to LDA, and are further used to discover K latent topics of the image patch dataset as the channel outputs. Every topic is defined as a probability mass function over the visual words v_l ($l \in [1, N_V]$), parametrized by the matrix $\beta_{N_V \times K}$ (obtained during the LDA parameter estimation phase). In Figure 4.1, the set of topics is denoted

as $T = \{t_1, t_2, ..., t_K\}$, and the probability of each visual word within a topic t_j is $p(v_l|t_j) = \beta_{lj}$.

The mutual information between the input visual words V and the output topics T is then computed as the difference between the input entropy $H(V)$ and a conditional entropy, $H(V|T)$, in which the input is conditioned by the output. The conditional entropy represents the amount of uncertainty about the input when the output is known. The mathematical notation of the mutual information is as follows:

$$I(V;T) = H(V) - H(V|T), \tag{4.1}$$

where $H(V)$ is obtained by:

$$H(V) = -\sum_{l=1}^{N_V} p(v_l) \log p(v_l). \tag{4.2}$$

In computing the input entropy, we compute the probability of every visual word v_l in the entire dataset by marginalizing the probability of the visual words across all M image patches as follows:

$$p(v_l) = \sum_{d=1}^{M} p(v_l|\mathbf{w}_d)p(\mathbf{w}_d). \tag{4.3}$$

In this equation, we assume that the image patches are equally probable, $p(\mathbf{w}_1) = p(\mathbf{w}_2) = ... = p(\mathbf{w}_M) = \frac{1}{M}$. In Equation 4.1, the conditional entropy $H(V|T)$ is computed as follows:

$$H(V|T) = -\sum_{j=1}^{K} p(t_j) \sum_{l=1}^{N_V} p(v_l|t_j) \log p(v_l|t_j), \tag{4.4}$$

where the topic marginal distributions are obtained by:

$$p(t_j) = \sum_{d=1}^{M} p(t_j|\theta_d)p(\theta_d|\alpha), \tag{4.5}$$

where $p(t_j|\theta_d)$ is the probability of every topic in the image patch $\mathbf{w}_d$; it is equal to θ_{dj} (for details, please refer to Section 1.4.1.3). In addition, $p(\theta_d|\alpha)$ is obtained according to Equation 1.17.

4.1.2 Results and Discussion

In this section, we applied our proposed communication channel model to the UC Merced Land Use dataset in order to assess five different feature descriptors. We use

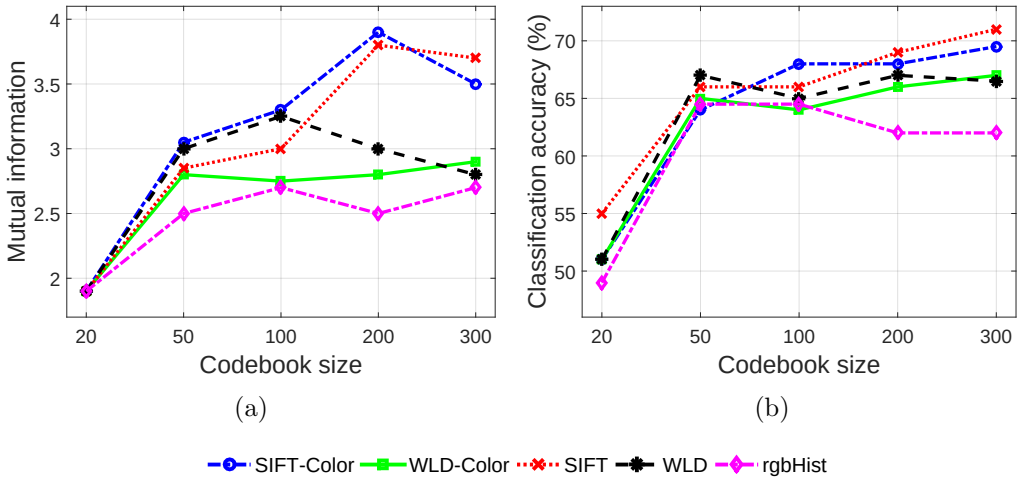


Figure 4.2: (a) Mutual information of the the communication channel computed for five kinds of feature descriptors and various number of visual words. (b) SVM classification accuracies for the five feature descriptors versus the number of visual words.

this dataset because it contains various classes in the sense of spatial patterns. There are classes homogeneous in color, classes homogeneous in texture, classes homogeneous in shape, and classes containing images which have no shared features. This variety of spatial patterns enables us to study three different properties of the images (color, texture, and shape) and their combination by the rgbHist, WLD, SIFT, WLD-Color, and SIFT-Color feature descriptors. These descriptors are computed locally in a dense way from a 32×32 sliding window with 50% overlap. In our experiments, LDA discovers the latent structure of the dataset as a set of 21 topics. Then we compute the channel's mutual information for different feature descriptors to quantify the amount of information delivered to the topics through each descriptor. In order to apply LDA to the images, they are represented by a BoW model. Since the visual words are the basic image elements contributing to the input information generation, we perform our experiments for different dictionary sizes. This allows us to study the effects of the number of visual words (the amount of input information) on the amount of delivered information. Figure 4.2. (a) shows the computed mutual information for five feature descriptors. According to the figure, increasing the number of visual words (increasing the source information) generally increases the amount of the transmitted mutual information; however, beyond a certain number of visual words the changes become insignificant. Moreover, comparing the feature descriptors, rgbHist and SIFT carry the smallest and the largest quantity of mutual information, respectively.

In order to show the generalizability of the predicted behaviors of the feature descriptors also for other image mining systems, we perform a supervised multi-class classification using SVM on the same data. Comparing the results of our

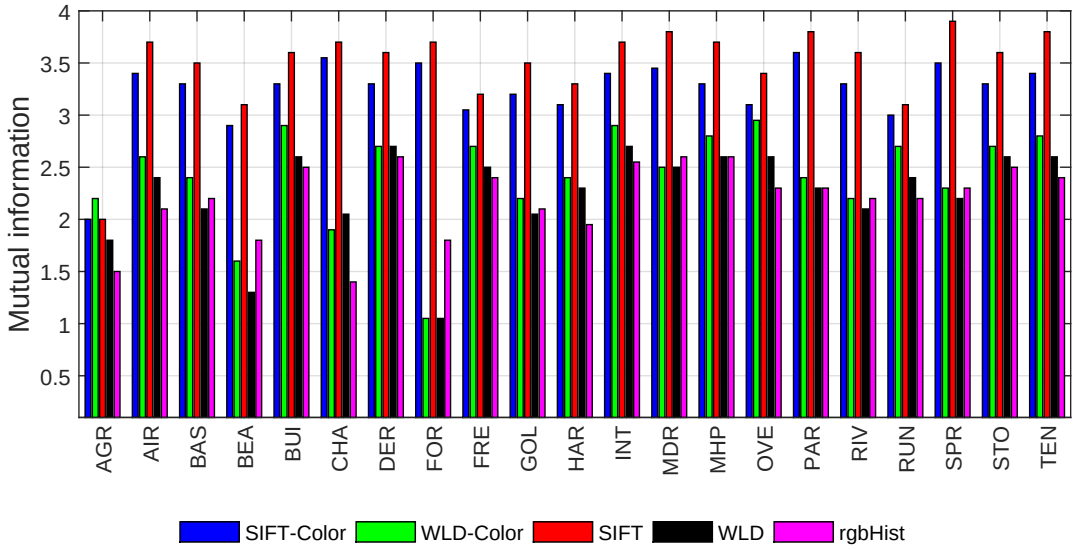


Figure 4.3: The mutual information computed for every individual class of the UC Merced Land Use dataset. The classes are: Agricultural (AGR), Airplane (AIR), Baseball diamond (BAS), Beach (BEA), Buildings (BUI), Chaparral (CHA), Dense residential (DER), Forest (FOR), Freeway (FRE), Golf course (GOL), Harbor (HAR), Intersection (INT), Medium density residential (MDR), Mobile home park (MHP), Overpass (OVE), Parking lots (PAR), River (RIV), Runway (RUN), Sparse residential (SPR), Storage tanks (STO), and Tennis court (TEN).

approach to the classification accuracies shown in Figure 4.2. (b), demonstrates that the channel's mutual information can predict the behavior of the feature descriptors applied to other image mining systems. As an example, since rgbHist carries less mutual information than the other descriptors, the images are less discriminable by their color features; and, consequently, the classification accuracy using rgbHist is also inferior to the other descriptors.

4.1.2.1 Class-wise Mutual Information

In this section, we compute the mutual information transmitted from each user labeled image class to the discovered topics. This indicates the dominant features of each class. The computation of mutual information is almost the same as it is explained in Section 4.1.1. The only difference is that for assessing an image class individually, in Equation 4.3 and Equation 4.5, we marginalize over the images belonging to that particular class. Figure 4.3 shows the mutual information for various features in different classes.

4.1.3 Summary

In this section, we introduce a communication channel-based approach to measure the information quantity that various feature descriptors provide to image mining systems. This method uses LDA as a communication channel, where images are the input, topics are the output, and the feature descriptors are the carriers of this channel. The channel's mutual information is then computed as the measure of the transmitted information quantity. Experimental results on the UC Merced Land Use dataset show that the SIFT descriptor carries the largest amount of mutual information in this dataset among the feature descriptors which has been analyzed in our experiments. This result is also confirmed by comparing it to the results of a classification task using SVM.

When we assess the different features of an image collection, using our proposed communication channel model, we get a general idea about the behaviors of image mining systems regardless of the user labeling. This allows us to develop specific feature descriptors to discriminate image classes in a given dataset. Moreover, the class-wise mutual information makes it possible to develop feature descriptors which are tuned to a specific image class across various datasets.

4.2 Feature Evaluation Based on Huffman Coding

Most of the existing image mining systems rely on primitive features which has led to the introduction of dozens of feature descriptors in recent years such as *RGB Color Histogram (rgbHist)*, *Weber Local Descriptor (WLD)* [3], and *Scale-Invariant Feature Transform (SIFT)* [70]. Although each feature descriptor is developed for extracting a particular kind of image features, it usually covers a range of feature types. That is why the image feature information provided by various feature descriptors may overlap, which we call *information overlap*. For example, SIFT which has been introduced mainly for representing geometry-based features such as edges and corners, provides also texture and color information to some degree. Thus, it may overlap with WLD (which mainly extracts texture features), or with rgbHist (which mainly extracts color features).

The information overlap is a key issue in feature fusion tasks. While the main idea behind the fusion of feature descriptors is to provide a comprehensive image description by integrating the information about different image properties, feature fusion may provide no additional information due to considerable information overlap between the feature descriptors. Usually users are not aware of the information overlap and they may try to improve the feature informativeness by combining more feature descriptors. However, increasing the number of feature descriptors usually also increases the final feature complexity of the fused descriptor and reduces the efficiency of learning algorithms. By considering the information overlap in feature fusion tasks we obtain a more comprehensive understanding of images and can select less but more diverse feature descriptors. This increases not only the discriminability of the given images but also scalability of image miming systems.

In this section, using information theory, we propose an approach based on *Huffman Coding (HC)* [123] to measure the overlaps between the information obtained by different feature descriptors. An overview of our proposed approach is shown in Figure 4.4. In our approach, the image features are extracted using various feature descriptors. Then the images are represented by BoW models. Following this, a lossless compression algorithm, a so-called Huffman Coding (for details, please refer to Section 1.4.2.1), is used to code all local points of each image by strings of binary digits ("0" and "1"). HC is based on the occurrence statistics of the visual words. Different feature descriptors provide non-identical visual word distributions which result in different Huffman codes of the images. The similarities between the codes are then quantified using the *Levenshtein distance* [151] as the information overlap measure. Since Huffman coding treats each image individually, our approach is independent of image labeling.

In order to validate the computed information overlap, we use it to explain the similarities between the clusterings of different feature descriptions of an EO image patch dataset, namely the UC Merced Land Use dataset. The results indicate that the similarity degree of the clusterings is proportional to the information overlap

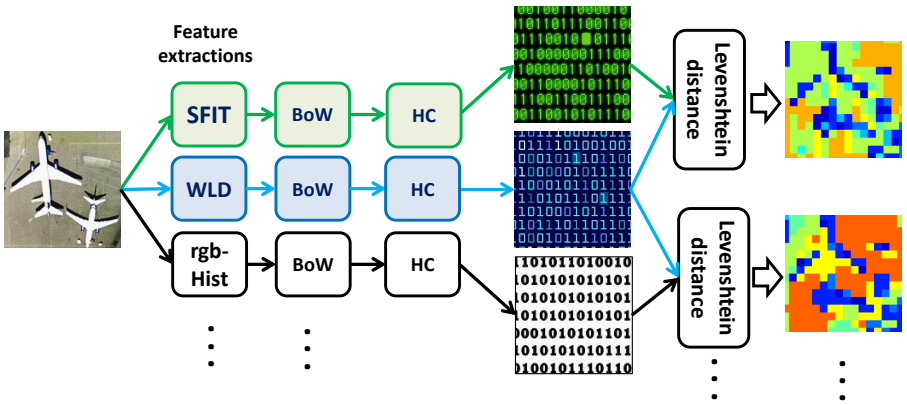


Figure 4.4: The pipeline of the proposed evaluation method: In this pipeline, for any given image, the point-wise distance between any pair of feature descriptors is computed using Huffman coding (HC) and Levenshtein distance. The output is the color coded distance of the local points.

of the employed feature descriptors. In addition, we use the information overlap to discuss the results of an image retrieval task on the UC Merced Land Use dataset. In this task, the image patches are represented by different feature descriptors. Then the retrieval results for a query using various image descriptions are compared. The results confirm that information overlap can justify the retrieval outcomes.

4.2.1 Methodology

In our proposed approach, the images are described by their local primitive features using different feature descriptors such as rgbHist, WLD, and SIFT. These descriptions are then used to represent every image patch $\mathbf{w}_d$ by a *Bag-of-Words (BoW)* model as a sequence of N_d visual words, with $\mathbf{w}_d = \{w_{d1}, w_{d2}, ..., w_{dN_d}\}$. Since every $\mathbf{w}_d$ is generated by the reoccurrence of the N_V visual words of a dictionary $V = \{v_1, v_2, ..., v_{N_V}\}$, the image patches can be represented by the probabilities of every v_l as $\mathbf{w}_d := \{p(v_1), p(v_2), ..., p(v_{N_V})\}$. This allow us to consider the image dataset as a data source. Under this assumption, the BoW models are considered as the documents and the visual words v_l are the data symbols. In order to code the image patches, we use HC due to its efficiency in mapping individual symbols to unique variable-length code strings. As a next step, a Huffman code table is generated for each image patch. Then every visual word is assigned a binary code based on its probability, which results in a binary tree model for each image patch. Since usually the visual words' semantics are from different levels, this tree representation allows a hierarchical ordering of the semantic image contents. In consequence, since the visual words refer to the local image points, the tree representation results in a hierarchical segmentation of each image patch. In this segmentation, the segments corresponding to the lower level visual words are usually components of the higher level segments. In order to place each visual word on the tree, its self-information

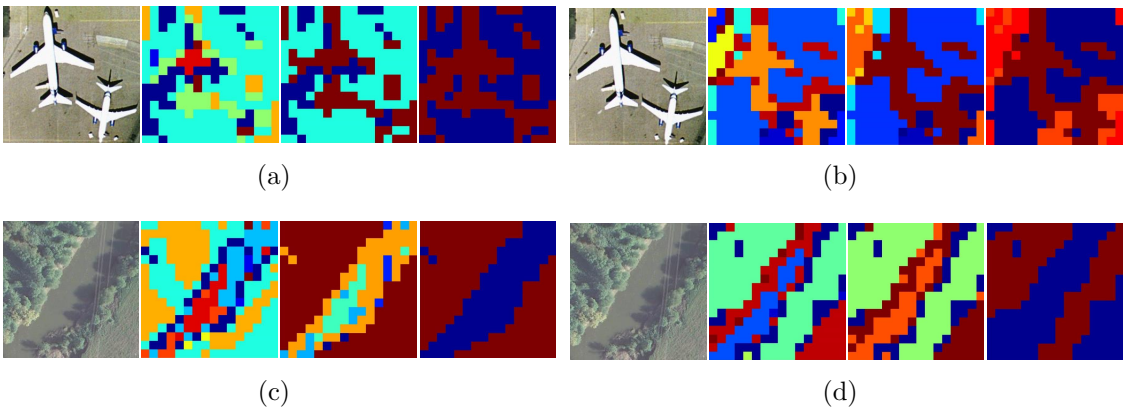


Figure 4.5: Huffman coding for two images using WLD-Color and rgbHist feature descriptors. For each image, the coded features are represented for 3 Huffman tree levels (from left to right, the feature levels increase), where the features belong to a segment are from the same semantic level. (a) WLD-Color. (b) rgbHist. (c) WLD-Color. (d) rgbHist.

(also known as symbol uncertainty) is considered, where the visual words with the highest and the lowest self-information are placed on the bottom and the top of the tree, respectively. For each image patch, the self-information $I(v_l)$ is then computed as:

$$I(v_l) = -log(p(v_l)). \tag{4.6}$$

According to this equation, less probable symbols have higher self-information and therefore, higher uncertainty.

From a semantic image understanding point of view, each feature level corresponds to a semantic level. In this way, the less frequent symbols correspond to the features with more uncertain semantic meanings (i.e., detailed features, object parts or properties). Traversing the Huffman tree from bottom to top, and grouping the lower level features, results in higher level features with less uncertain semantic meanings. Figure 4.5 shows the HC for two image patches from the UC Merced Land Use dataset. In order to code each image patch, it is represented by WLD-Color (obtained by concatenation of the WLD feature vectors computed for the RGB color channels) and rgbHist feature descriptors. In this figure, each image patch is followed by three segmented images, which visualize the Huffman tree based hierarchical segmentation for three different levels where, from left to right, the segment levels increase. Comparing the coded features in Figure 4.5. (a,b), the features represented by the two descriptors become more similar as the feature level increases. However, in Figure 4.5. (c,d), even the highest level features of the two descriptors reflect different image properties. In Figure 4.5. (c), for example, WLD-Color describes the image as “water body” and “non-water body” at the highest semantic level, while, in Figure 4.5. (d), rgbHist results in similar semantics for the water body and a part of the field (in the lower right corner of the image) due to their

similar content. More precisely, rgbHist assumes that the image patch contains the two semantic categories "forest green" and "field/water body green" at the highest level. As the examples show, according to the contents present in an image patch, different feature descriptors can reveal different degrees of similarity in representing semantic image contents.

Since each feature descriptor extracts a particular range of primitive features, the visual word distributions are different across various feature descriptors, which results in different image Huffman trees and codes. SIFT, for example, extracts geometry-based features such as edges and corners, and to some extent provides texture and color information, while rgbHist mainly extracts the color information of an image. Therefore, although they partially share a common type of information (e.g., color), their main focuses are rather diverse. In order to measure the amount of the overlapping and the disjoint part of the extracted information by any pairs of feature descriptors, we compare the HC of the two descriptions of each image. To this end, every local image point is assigned two binary codes based on its corresponding visual words of the two feature descriptors. Then the Levenshtein distance between the two binary codes of the local points is computed. The Levenshtein distance is a metric which measures the difference between any two character sequences (e.g., binary strings in our experiments) by computing the minimum number of required character insertions, deletions, and substitutions to convert one sequence into the other one [151]. In our experiments, this metric quantifies the similarity degree between the feature distributions with respect to their information level determined by the Huffman tree. More precisely, the visual words with similar information levels (i.e., with similar distributions over the image) are placed closely together on the Huffman tree. Therefore, the length and the sequence of the binary digits of their assigned code strings are similar, which results in a small Levenshtein distance.

4.2.2 Results and Discussion

In this section, we use our proposed approach to quantify the information overlaps between rgbHist, WLD-Color, and SIFT-Color feature descriptors on an EO image patch dataset, namely the UC Merced Land Use dataset. The descriptors are applied densely to extract primitive image features from 32×32 local windows with 50% overlap. In our setup, each rgbHist feature vector has 256 dimensions. WLD-Color and SIFT-Color are obtained by computing WLD and SIFT feature vectors for each of the three image channels (RGB) individually. The resulting feature vectors from the three channels are then concatenated to form vectors of 432 and 384 dimensions, respectively. In the next step, k-means is used to group the extracted feature vectors into 200 clusters. Each cluster center is then considered as a visual word. After applying a *Hard Voting (HV)* coding strategy, we assign the local feature vectors to the visual words and represent the images by a BoW model. In our experiments, each local image point is represented by three different types of visual words derived

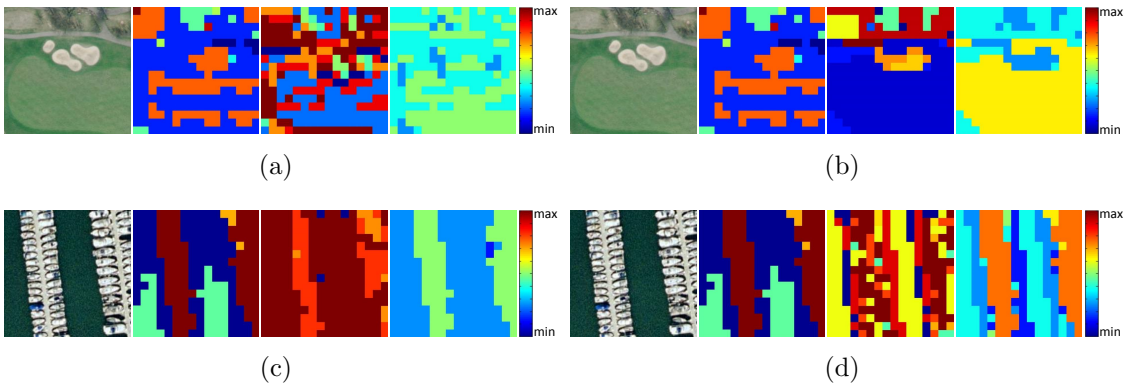


Figure 4.6: Information overlaps computed for feature descriptor pairs. In each sub-figure, the first image is the original image, the second and the third segmented images are the visualizations of the two Huffman coding variants of the original image, and the last image is the color map of the Levenshtein distances for the local image points. The feature pairs with smaller Levenshtein distances have higher information overlap. (a) Between WLD-Color and SIFT-Color, average Levenshtein distance = 3.39. (b) Between WLD-Color and rgbHist, average Levenshtein distance = 4.06. (c) Between WLD-Color and rgbHist, average Levenshtein distance = 2.71. (d) Between rgbHist and SIFT-Color, average Levenshtein distance = 3.71.

from the three feature descriptors. Based on the visual word distributions of the different image descriptions, Huffman code trees and tables are created for each image patch. Since the visual word distributions of the three feature descriptors vary in each image patch, the resulting Huffman trees and tables are dissimilar. Thus, each local image point is assigned three unalike binary code strings. The Levenshtein distance is then computed for every code pair assigned to every local image point. We consider the average Levenshtein distance over the local image points as the measure of the information overlap. Figure 4.6 shows the information overlaps computed for two image patches of UC Merced Land Use dataset. In this figure, each sub-figure compares a pair of feature descriptors, in which the first image is the original image patch, the second and the third segmented images are the visualization of the two HC variants, and the last image is the color map of the computed Levenshtein distances for every local image point. The average Levenshtein distances over the entire image then indicated as the mean distance between the two descriptions of the image patch. The closer the descriptions, the higher the information overlap. In Figure 4.6. (b), for example, while the green field looks homogeneous for rgbHist, WLD-Color discovers various textures there. Therefore, the Leveneshtein distance is large in this area (the information overlap is small). However, in the area with non-homogeneous colors, at the top of the green field, both descriptors discover similar patterns which cause smaller Levenshtein distances (a large information overlap).

In order to demonstrate the ability of the information overlap in predicting the

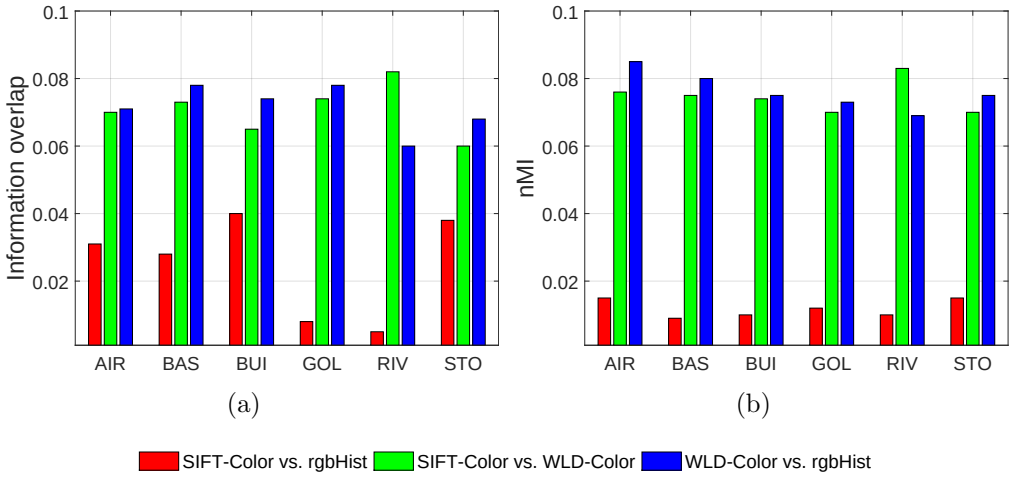


Figure 4.7: Demonstration of the correlation between information overlap and the similarity degrees between the clusterings of various image descriptions for sample classes of UC Merced Land Use dataset including "Airplane (AIR)", "Baseball diamond (BAS)", "Buildings (BUI)", "Golf course (GOL)", "River (RIV)", and "Storage tanks (STO)". (a) Information overlap. (b) nMI as the similarity measure of the clusterings.

behaviors of image mining systems, we use the information overlap to explain the similarities between the clusterings of the UC Merced Land Use dataset when the image patches are represented by the three feature descriptors. To this end, k-means clustering is applied to the BoW models of image patches. Then the *normalized Mutual Information (nMI)* clustering evaluation measure [152] is computed in order to compare the clusterings. We observe that for each feature descriptor pair, the similarity between their clusterings and their information overlap are correlated. Figure 4.7. (a) shows the information overlap for each pair of the three descriptions of some classes of the UC Merced Land Use dataset; Figure 4.7. (b) presents the nMI values as the similarity degrees between the clusterings of the classes. In order to compute the information overlap, we normalize the average Levenshtein distances over the three pairs. Since Levenshtein distance and information overlap are inversely related, we subtract the normalized Levenshtein distances from 1 to achieve information overlap. In order to make nMI comparable to the computed information overlap, we normalize the nMI values over the three descriptions as well. As the results show, for the class "Airplane", for example, WLD-Color has more information in common with rgbHist than with SIFT-Color. This causes the clusterings of the WLD-Color and rgbHist descriptions to be more similar than those of the rgbHist and SIFT-Color descriptions. Moreover, amongst the three feature descriptors, SIFT-Color and rgbHist extract the most diverse information. Therefore, if a fusion of two features for these classes is required, SIFT-Color and rgbHist can provide a broader range of new and diverse information than any other combinations of the

descriptors.

In order to show that the measured information overlap is applicable to various image mining tasks, we use it to explain the retrieval results of two image patches ("River" and "Baseball diamond") described by the three feature descriptors. Figure 4.8 shows the retrieval results in which, for each experiment, the first image shows the query, followed by the retrieval results sampled from the top 30 relevant image patches. When we analyze Figure 4.8. (a-c), the contents of the results from SIFT-Color and WLD-Color are more similar than those of rgbHist, which is consistent with the computed information overlap for the class "River" in Figure 4.7. (a). Both SIFT-Color and WLD-Color focus on the patterns of the trees and the water body (mostly in the forest and riverside areas of the query). In addition, WLD-Color is able to find the tree patterns in residential areas too, which is not possible for SIFT-Color. On the other hand, since rgbHist extracts the color information of the trees and the water body from the query, which could be found in a quite broad range of land cover types, the retrieval results include riversides, tennis courts, highways, and residential areas according to Figure 4.7. (c).

In Figure 4.8. (d-f), the results show that the contents of the features extracted by rgbHist and WLD-Color are very similar which verifies the high information overlap for the class "Baseball diamond" in Figure 4.7. (a) this means that the textures extracted by WLD-Color and the colors extracted by rgbHist from the given baseball field usually co-occur in land covers such as baseball diamonds, golf courses, and to some extent in airports. Therefore, in this dataset for example, if only a discrimination of the classes "Baseball diamond" and "Golf course" is required, replacing one of the descriptors (WLD-Color and rgbHist) with the other one may cause no significant change to the results. However, since the extracted features by SIFT do not occur in golf course images, replacing rgbHist or WLD-Color with SIFT-Color can result in a significant change in the discrimination performance. In order to verify our statement about the influences of the three feature descriptors in discriminating "Baseball diamond" and "Golf course" classes, we apply SVM classification to the three descriptions of these classes. The results in Figure 4.9 indicate that rgbHist and WLD-Color cause SVM to perform quite similarly for various number of training samples; however, using SIFT-Color helps to achieve a much higher discrimination of the classes.

4.2.3 Summary

In this section, using information theory, we propose a novel approach based on Huffman Coding (HC) to measure the overlaps between the information obtained by different feature descriptors. The information overlap is used as a measure of similarity between any two feature descriptors in representing an image dataset. In our proposed approach, we use Huffman coding to code the BoW model of every image patch according to the distribution of the visual words in the image patch.

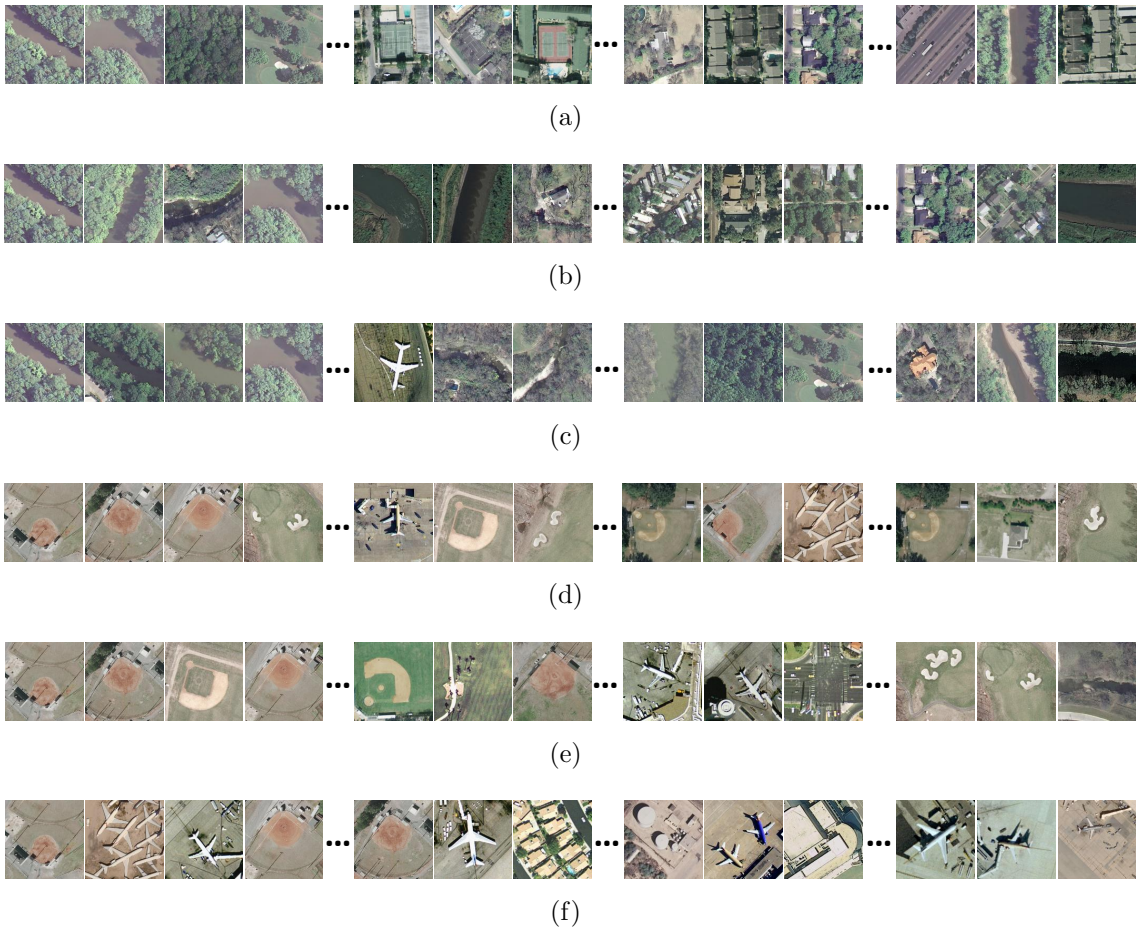


Figure 4.8: Retrieval results of two sample images of the UC Merced Land Use dataset ("River" and "Baseball diamond"), described by rgbHist, WLD-Color, and SIFT-Color feature descriptors. For each experiment, the first image shows the query, followed by the retrieval results sampled from the top 30 relevant images. The following results are represented in sequences, where the dots depict the discontinuity. (a) "River" described by rgbHist. (b) "River" described by WLD-Color. (c) "River" described by SIFT-Color. (d) "Baseball diamond" described by rgbHist. (e) "Baseball diamond" described by WLD-Color. (f) "Baseball diamond" described by SIFT.

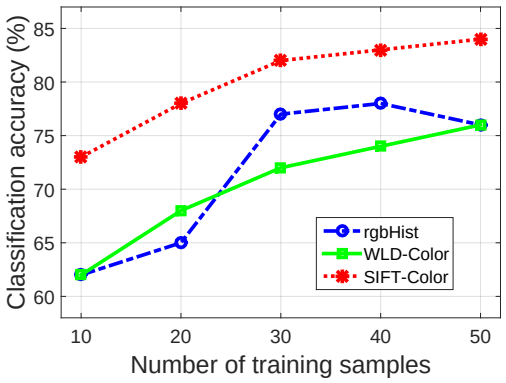


Figure 4.9: SVM classification accuracy in discriminating the images of the two classes "'Baseball diamond" and "Golf course". The images are described by rgbHist, WLD-Color, and SIFT-Color feature descriptors.

This coding provides a tree structure of the visual words for each image patch, where the lower level visual words correspond to the components of the objects represented by higher level visual words. Then the similarities between the trees obtained for various descriptions of the images (using rgbHist, WLD-Color, and SIFT-Color feature descriptors) are computed by their Levenshtein distances, and considered as the amount of shared information between the descriptions.

Experimental results show that the computed information overlap can justify the different performances of clustering methods when using various feature descriptors. Moreover, considering information overlap in feature fusion tasks allows providing a broader range of new and diverse information by a combination of less feature descriptors, which improves both the scalability and distinguishability of image mining systems.

5

Human Image Understanding: Evaluation of the Sensory and Semantic Gaps

5.1 Assessment of the Sensory Gap

In the literature, the sensory gap has been defined as the gap between an object in reality, and its representation based on the signals recorded by sensors [9]. A desired object can either be perceived directly by the user, or detected after processing the information in a machine learning application. In either case, a sensory gap exists. Causes behind the sensory gap can lie at the scene (e.g., clutter, occlusion) or sensor levels (e.g., perspective, resolution, field of view, perceptual spectra). In EO, the sensory gap is rather wide due to sensors (e.g., radar, multi- and hyper-spectral instruments) which record visual information very differently from the human visual system [11].

The sensory gap is affected by the complexities of the EO images, such as the resolution, perspective, or scale of the visual information [11]. The perspective of the images is a particular challenge in EO, since they present a bird's eye view. As described in the *recognition by components* theory [111], objects can be segmented into their *geometric components (geons)*, and we recognize them based on the identification of their geons and their structural relationships, which we then match to mental representations. Object recognition should be perspective invariant, so long as the structural relationship between geons can be identified from the different perspective. This is not the case when objects are viewed from above, since major object components can be occluded, making it harder to match the object to the stored mental description. Therefore, from this perspective, object identification is more difficult [111].

The sensory gap is also affected by the *Field of View (FOV)* presented. The FOV is the extent of the observable scene that is seen at any given moment. For the

EO imagery instruments, FOV refers to a solid angle through which a sensor can record signals from the scene. The larger the FOV, the more information is present in the image, specifically contextual information, which the user can apply when trying to identify and recognize an object within the image. Research on object detection and recognition in humans has shown the importance of context [153]. Context can provide information on spatial relations, semantic associations, global scene properties, and pose [154, 155]. When the object is not easily discernible on its own (due to low resolution, for example), contextual information becomes increasingly important [156, 157].

In this section, we assess the causes of the sensory gap in EO images by a human perceptual evaluation and a computational evaluation. Figure 5.1 shows the process chain of our evaluation method. The human perceptual evaluation is assessed by user labeling (User annotation C in Figure 5.1) of EO scene image patches, using content labels (listed in Table 5.2) defined for the scene by a previous annotation (User annotation B in Figure 5.1). The assigned labels and user feedback are then analyzed (the procedure is shown in green in Figure 5.1). Results point to image properties that limit image understanding, such as resolution, which users report is not high enough to readily discriminate objects. Image perspective also presents a challenge, since users are not used to this bird's eye view. The scale of objects in the image patches is also difficult to assess. When users are uncertain of an object's identity, due to other image properties, such as resolution or perspective, they could turn to the context surrounding the object to gather clues to identify it. However, due to the FOV which is constrained by the patch size, users have limited contextual information. Since recent image mining approaches usually split EO images into patches in order to gain more locality of the image features, selecting patch size which can provide more locality while contain enough contextual information is a challenge. Since we present image patches to the users and computer, we consider the patch size as the extent of the scene (i.e., the whole EO image) which can be perceived by the users and computer. Therefore, we use the term FOV referring to the patch size.

The effect of FOV is then evaluated by a computational method, in which the acquired context from the content reference (derived from a manual labeling of the scene described in Section 5.1.1) of a certain FOV is statistically analyzed using LDA. The sensory gap is considered as the difference between the context of the scene discovered by LDA, based on the occurrence of content reference labels within a certain FOV (which corresponds to increasing the image patch size), and the scene reference context. The results indicate that increasing the FOV decreases the sensory gap.

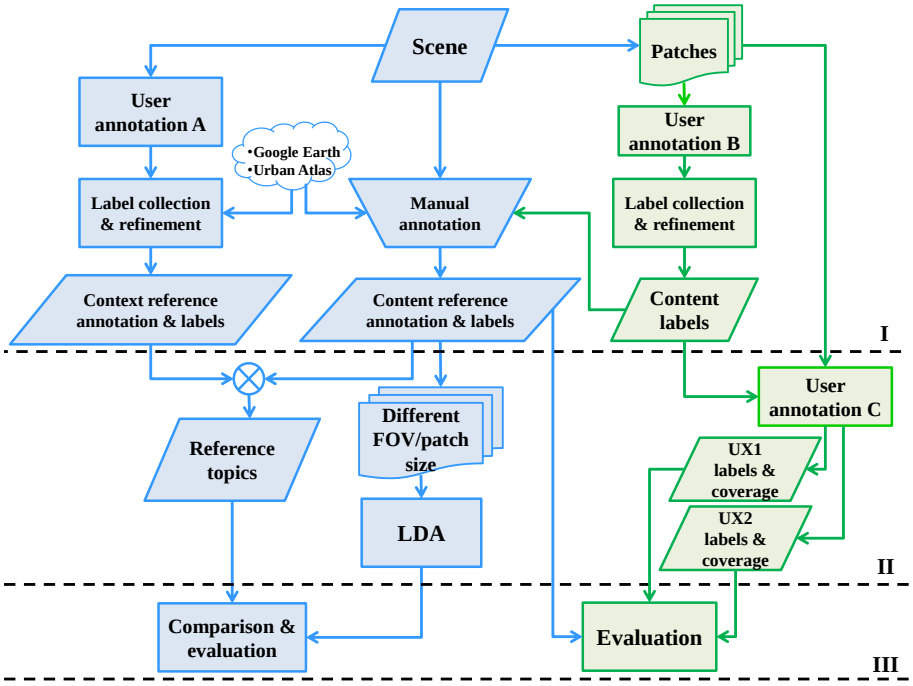


Figure 5.1: Process chain of the user perceptual evaluation (shown in green) and the computational evaluation (shown in blue) of the sensory gap.

5.1.1 Context and Content Reference Annotation and Label Generation

A multispectral scene of the Feldmoching area to the north of Munich, Germany, acquired on July 12th, 2010 (10:30 am UT) by the WorldView-2 satellite was used for annotation. The image has a resolution of 1.84 m, was trimmed to a size of 2000×1800 pixels, and three bands were displayed (RGB).

The process chain of the present study (please refer to Figure 5.1) shows an overview of the necessary steps, where each step is underlined and described in the following text. First, the EO Scene, using both human user experiments (the process steps shown in green) and computer experiments (shown in blue), to evaluate the sensory gap from both perspectives. In the initial phase, the scene was given to 9 human users (none of whom had a background in image processing), who were asked to annotate the image using the LabelMe tool [158]. This refers to User annotation A in the process chain in Figure 5.1. In this step, users were presented with the scene, and given a short demo of the tool. A *free text annotation* [159] was conducted - meaning that users were asked to label what they see, without using references or dictionaries. This approach was selected to gather labels based on user perceptions, without external influences. Each user generated an average of 19 unique labels. In the following step, Label collection & refinement, all unique labels (excluding

Figure 5.2: Context reference annotation, created for the 8 labels in Table 5.1.

1	Agricultural & semi-natural areas	5	Residential areas
2	Industrial/Commercial/Public/Military	6	Sport and leisure
3	Isolated structures	7	Transportation infrastructure
4	Natural areas	8	Water body

Table 5.1: Context labels

duplicates, plurals, synonyms) were identified, and polygons from the 9 annotations were compared to identify their commonalities. These annotations produced labels corresponding to higher level semantics, such as industrial areas and urban areas, indicating that users focused on the broader *gist* of the scene, as opposed to its details. These higher level semantics were gathered, loosely refined based on Urban Atlas[1] and 8 context labels were determined (please refer to Table 5.1). These context labels were used together with Google Earth[2] to manually annotate the image and create a Context reference annotation & labels (please refer to Figure 5.2 for a screenshot of this annotation).

For the user experiments, the scene was divided into 200 × 200 patches, with 50% overlap, resulting in 323 Patches. Then User annotation B was carried out,

[1]http://www.eea.europa.eu/data-and-maps/data/urban-atlas
[2]https://www.google.com/earth/

1	Agricultural field	7	Greenhouse	13	Railway
2	Building	8	Highway	14	Road
3	Crop	9	House	15	Soccer field
4	Factory	10	Isolated trees	16	Solar panels
5	Forest	11	Lake	17	Street
6	Grass	12	Parking lot	18	Tennis court

Table 5.2: Content labels

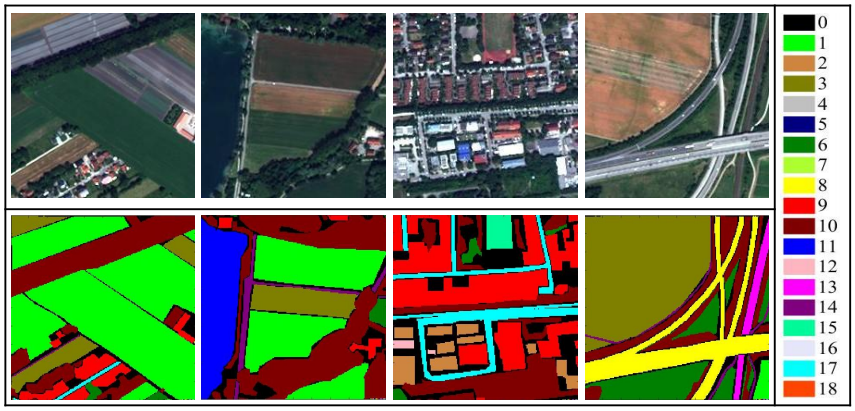


Figure 5.3: Sample image patches and their corresponding Reference content annotation. The legend shows the correspondence of the 18 labels in Table 5.2 to the annotated regions. The label "0" refers to the unlabeled areas.

where 3 different users did a free text annotation labeling, an average of 108 patches each. Next, Label collection & refinement took place, so that 18 labels describing the content of the patches were left (please refer to Table 5.2). In contrast to the labels given by the first 9 annotators, these labels corresponded to lower level semantic categories, such as lake and houses. These 18 Content labels were used as a lower level semantic dictionary, which provided a manageable set of terms, but was also rich enough to highlight the previously mentioned problems associated with the sensory gap. For example, the labels "road", "street", and "highway" illustrate perception problems due to scale; "agricultural field" and "grass" highlight issues with resolution; "building" and "house" highlight issues with perspective.

The 18 Content labels, together with Google Earth, were used in a manual annotation of the scene, creating a Content reference annotation. Figure 5.3 depicts samples of the image patches and their corresponding reference annotations. In this figure, the legend shows the correspondence of the 18 labels in Table 5.2 to the annotated regions; together with the label "0" which refers to the unlabeled areas. At this point we have both Context reference annotation & labels and Content ref-

erence annotation & labels (please refer to Phase I in Figure 5.1). Phase II describes our experimental procedure, which will be detailed in Section 5.1.2.1 from the user experiment side, and in Section 5.1.3.1 from the computer side. Phase III consists of the experimental outputs addressing the user perceptual and the computational evaluations of the sensory gap, which will be discussed in Sections 5.1.2.2 and 5.1.3.2.

5.1.2 User Perceptual Evaluation of the Sensory Gap

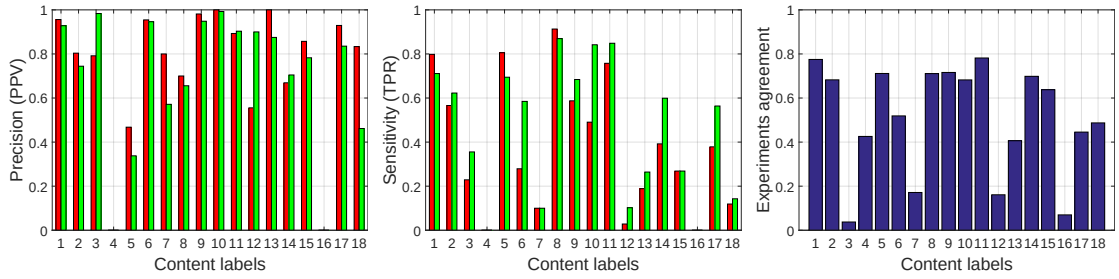
For a user perceptual evaluation of the sensory gap, User annotation C (please refer to Figure 5.1) was carried out. The experimental procedure and results will be discussed below.

5.1.2.1 Experimental Procedure

The 323 image patches previously described were divided into eight groups (seven groups of 40 patches, one group of 43 patches). Users were each given one group of patches, and a handout with the dictionary of content labels listed in Table 5.2, each assigned to a number code (e.g., 1=Agricultural field), and a second table with codes A-E, each corresponding to a percentage range (A=0-19%, B=20-39%, etc). Users were asked to look at each patch (zooming in as needed), and assign it at least one alphanumeric code, representing both the semantic content of the patch (the label), and the approximate area of the patch covered by each label (the coverage). For example, code 1A indicates there is an agricultural field, covering between 0-19% of the patch area. In this step, 16 users participated (the first 8 corresponding to *user experiment 1 (UX1)*, the second 8 to *user experiment 2 (UX2)*), so that each group of patches was labeled twice. This produced the UX1 labels & coverage and UX2 label & coverage. After labeling, participants were asked to fill out a short questionnaire, to gauge their perceptions on how confident they were of the correctness of their labels, and to give general feedback; all of which was used to further understand the results.

5.1.2.2 Results and Discussion

The similarity between the user and reference labels is computed by two measures, *Precision* and *Sensitivity*, which are formulated for three types of quantities, namely *true positive (TP)*, *false positive (FP)*, and *false negative (FN)* [160]. While Precision ($PPV = \frac{TP}{TP+FP}$) indicates the correctness of the user assigned labels to the patches, Sensitivity ($TPR = \frac{TP}{TP+FN}$) shows the percentage of the reference labels which have been identified by the users. In Precision, for each label, TP and FP are the number of times the label is correctly and incorrectly assigned by the users, respectively. In Sensitivity, TP and FN are the number of times a reference label is identified or missed, respectively.



(a) (b) (c)

Figure 5.4: Precision, sensitivity, and agreement of the labels for the two user experiments. In (a) and (b), UX1 and UX2 are depicted by red and green bars, respectively.

The average Precision rates (UX1=73% and UX2=70%) and Sensitivity rates (UX1=38% and UX2=46%) were similar for both user experiments (please refer to Figure 5.4 (a,b)). The relatively high Precision rates indicate that when users assigned labels, they were mostly correct. However, there was a large portion of missing labels, as reflected in the Sensitivity rates. Users reported their confidence in their labeling as an average of 3.7 on a *Likert* scale [161] (where 1 is not at all confident and 5 is very confident), indicating that they were aware of the potential inaccuracies of their annotations, including the unlabeled objects they could not identify or detect.

When asked to describe the difficulties of the labeling task, users cited problems with understanding the object scales, which then led to questions on how to distinguish semantically related terms (such as "road" and "street") which are typically differentiated by their size. Users also mentioned that the resolution of the image was not high enough to distinguish certain objects. The fact that they could not see the contextual information surrounding the patch, combined with the perspective of the image, made users unsure of what certain objects would look like. Therefore, they would have liked to use examples of labeled patches as a guide.

To further understand the patterns of errors, missing and correctly identified labels, we looked at the Precision and Sensitivity of each label, as well as the user feedback given. The results were analyzed with regard to the sensory gap. In terms of incorrect identification, two object classes stand out, "factory" and "solar panels". Although neither of these object classes were present in the image provided, users detected them. Due to the image's perspective, which does not provide height or depth information, the user can only see a cluster of similar buildings. Paths and small parking lots may be confused with factory infrastructure such as pipes connecting different sections of the factory. In the case of solar panels, the effect of perspective resulted in confusing greenhouses with solar panels (probably due to the way they reflect light).

Issues of scale are highlighted with the labels "highway", "road" and "street".

"Highway" is more likely to be confused with "road" or "street", than to go undetected. In the case of "street" or "road", users are more likely to miss them or to confuse them with each other. User feedback indicates that these objects are similar and distinguished based on size; however, the limited FOV of the patches makes this difficult for the users to judge.

The average user label agreement rate was found to be 50.6% among all categories (please refer to Figure 5.4 (c)). High agreement on several categories indicates they were easier to detect and discriminate (e.g., "lake" and "agricultural field"). "Solar panels" have a particularly low agreement rate, because this object category was not present in the image, and users confused it with different objects. In the case of "factory", although this object category was not present in the image, the user agreement rate is not as low because users confused the same objects with factory, indicating they have a similar mental representation of what it should look like. Two other categories have a particularly low agreement rate, "greenhouse" and "parking lot". These categories are hard to discriminate or hard to detect, and users mostly missed them, as can be seen by their corresponding Sensitivity rates. The category "crop" also had a very low agreement rate. Even though most of the labels for "crop" were correctly assigned, users did not label a large percentage of the crops in the image, as evidenced by the Sensitivity rates. User feedback reported confusion between the categories "crop" and "agricultural field", since the resolution of the image was not high enough for them to make this distinction. Users also expressed difficulties distinguishing between "building" and "house"; however, there was a high agreement rate for these categories, indicating that even if users express a degree of confusion between the terms, they share a similar mental representation of them.

These results highlight the ways in which the image perspective, resolution, scale and FOV are some of the causes behind the sensory gap from a human user perspective.

5.1.3 Computational Evaluation of the Sensory Gap

In Section 5.1.2, the sensory gap's causes (image resolution, perspective, scale, and FOV) are explored via a user perceptual evaluation. Among them, in the following sections, FOV is further assessed via a computational evaluation, in which LDA performs a statistical analysis of the contextual clues provided by a given patch with a certain FOV (a general overview is shown in Figure 5.1).

5.1.3.1 Methodology

As a first step in our experiments, the Context labels (please refer to Table 5.1) are represented as distributions over the Content labels (listed in Table 5.2), $W := \{w_1, w_2, ..., w_n\}$; this representation is called Reference topics, $\tilde{Z} := \{\tilde{z}_1, \tilde{z}_2, ..., \tilde{z}_m\}$. The distributions are obtained by overlapping the context and the content reference

annotations, and by pixel-wise measuring the overlap for each pair of context and content labels.

As a next step, the Reference content annotation is split into patches, where the patch size reflects the patch's FOV. The coverage of the content labels in each patch is considered as the occurrence probability, and this value is used to represent the patch as a histogram of the content labels. LDA is then applied to the histograms to discover the latent topics, $Z := \{z_1, z_2, ..., z_K\}$, behind the patch collection (reflecting the scene context), where each topic is a distribution over the content labels. Since the FOV limits the contextual clues, the resulting scene context differs from the reference context which is derived from the complete scene. This difference is then considered as the effect of FOV on the sensory gap. Figure 5.5 (a) exemplifies how changing the FOV limits contextual clues. For a 100 pixel patch, for example, roads cannot be well identified using the contextual clues.

In order to measure the difference between the two sets of topics, symmetrized Kullback-Leibler divergence [17] is used,

$$D_{KL}(R_i||Q_j) = \frac{1}{2}[\sum_{x=1}^{n} R_i(x)\ ln\frac{R_i(x)}{Q_j(x)} + \sum_{x=1}^{n} Q_j(x)\ ln\frac{Q_j(x)}{R_i(x)}], \tag{5.1}$$

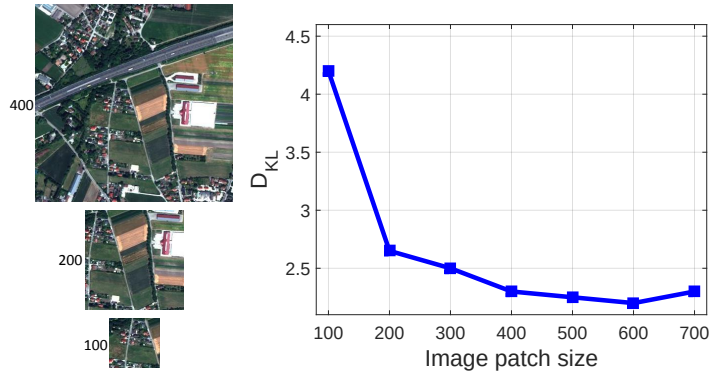
where $R_i(x) = p(w_x|\tilde{z}_i)$ and $Q_j(x) = p(w_x|z_j)$. For each LDA-topic, the closest reference topic is considered as its corresponding topic. The sum of the distances of the LDA-topics to their corresponding reference topics is then computed as the final distance between the two sets of topics, corresponding to the sensory gap.

5.1.3.2 Results and Discussion

In our experiments, LDA is applied to the content label representation of the image patches for various numbers of topics, $k \in [5, 12]$. Since LDA does not provide unique results, each experiment is repeated five times. The final sensory gap for a particular patch size is then obtained by averaging over all of the experiments. As Figure 5.5 (b) shows, increasing the FOV (patch size) significantly reduces the sensory gap up to a certain point (200 pixels). Further increasing the FOV causes no significant change to the sensory gap. This demonstrates that for the given scene considering all its properties (e.g., size, resolution, spectrum) a patch size of more than 200 pixels, statistically, does not add many contextual clues to each patch.

5.1.4 Summary

In EO, the sensory gap is rather wide due to sensor resolution, image perspective, scale and FOV. In this work, the sensory gap is assessed by human perceptual and computational evaluations. For a human perceptual evaluation, user labels describing image patch content are gathered and analyzed. The results highlight issues



(a) (b)

Figure 5.5: (a) Example of limitation of the contextual clues by changing the FOV (image patch size). (b) Influence of FOV (image patch size) on the sensory gap.

caused by the sensory gap. For example, the bird's eye view perspective of the image is one which humans are not accustomed to, and therefore affects object recognition. Resolution and scale present additional difficulties for object recognition. Users can disambiguate objects by gathering context from the image's FOV; therefore, a limited FOV makes issues such as resolution more serious. The effect of FOV on the sensory gap is also assessed via a computational evaluation, where the sensory gap is defined as the difference between the scene context discovered by LDA from content within a certain FOV (image patch size) and the reference context. The results indicate that increasing the FOV decreases the sensory gap. Future work could extend the research on FOV and how it interacts with other factors that cause the sensory gap (such as resolution).

5.2 Exploration of the Semantic Gap from User and Computer Perspectives

The semantic gap, in most of the previous research, has been defined as the difference between the user's understanding of objects in an image, and the computer's interpretation of those objects [9, 12, 13, 14]. However, each user will interpret images differently, and use different terms to label the objects within them, and this is what we call the *linguistic semantic gap*. While previous research addressed this as a "vocabulary problem" [15, 16], showing that it is unlikely for two people to assign the same label to a given object; this problem has not been considered in the context of the well-known semantic gap. Research on the semantic gap has considered differences between user and computer interpretations of an image, and proposed methods to bridge it, such as introducing various machine learning algorithms [99], using correlations among multiple data modalities (e.g., image, text, metadata) [117], discovering semantic rules between users and computers [12], and using interactive models [13]. The proposed methods have been verified either by comparing results to reference data, or by measuring the degree of user acceptance in the interactive systems. In this section, we show that since the *gold standard* is set by user created references or user acceptance, user subjective biases are included in this standard. Thus, although these methods result in a narrower semantic gap between computers and users, the linguistic semantic gap remains, therefore the resulting model for a specific user and search goal may still not be satisfactory to other users.

To overcome this problem, we propose that efforts to bridge the semantic gap should consider the linguistic semantic gap, and increase the diversity of data sets used in the domain (e.g., using various EO datasets for EO tasks), which will include different user perspectives and compensate for the individual subjective biases. Moreover, models derived from the proposed methods for bridging the semantic gap could be stored and further used by other systems, which would then be including other users' image interpretations.

Furthermore, we show the relationship between the sensory and the semantic gap. When users are presented with an image to annotate, they must both identify the objects in it, and label them. For every user, the task of object discrimination can be affected by the sensory gap, since users are limited to what they can perceive in an image, and this is influenced by image characteristics, such as resolution. Once objects have been identified, labeling them can also lead to different results for each user, due to their pre-existing knowledge, or the use of additional information (e.g., maps in EO), causing the linguistic semantic gap. Since users first perceive and identify objects, and then label them, it can be said that the semantic gap builds on the sensory gap (i.e., the sensory gap is one of the causes of the semantic gap).

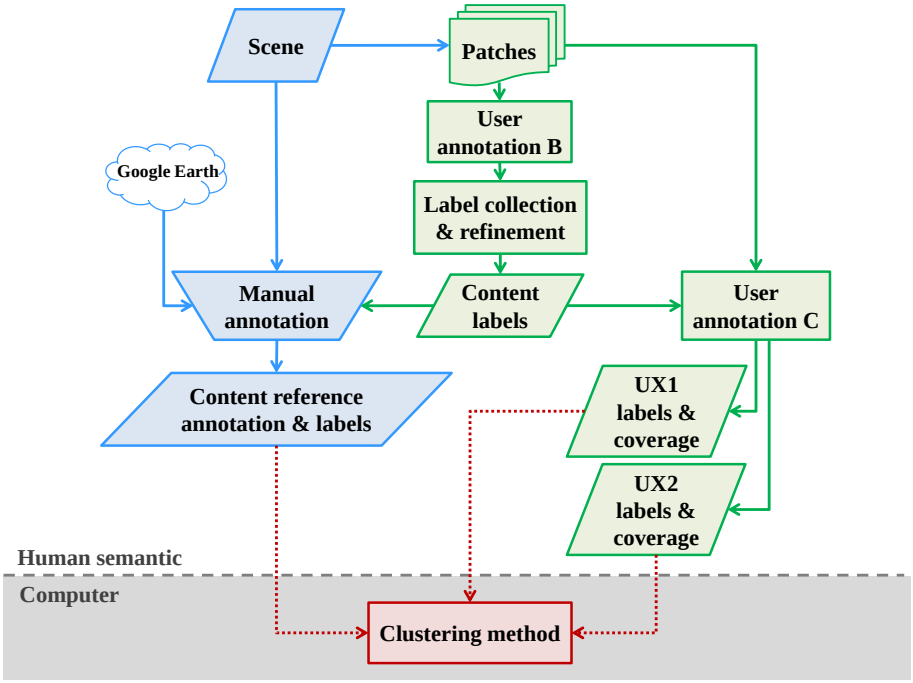


Figure 5.6: The process chain for the semantic gap assessment. For explanation, please refer to Section 5.2.1.

5.2.1 Experimental Procedure

In our work, we conduct a user and a computer experiment. A complete overview of the process chain followed for this section is depicted in Figure 5.6. In this diagram, first human semantics are gathered (as described in Section 5.1.1 and Section 5.1.2) in order to achieve UX1 labels & coverage, UX2 labels & coverage, and Content reference annotation & labels. In the following, we abbreviate the *content reference annotation* with *REF*.

5.2.1.1 Computer Experiments

Considering the semantic gap as the difference between the user and computer descriptions of the image, we measured it by comparing the distribution of the labels assigned by the users to the distribution of the labels assigned by a machine learning algorithm. From a user perspective, the image is described by its content in the form of semantic labels; and the distribution of the labels is based on the corresponding user assigned coverages. From a computer perspective, the image is described by a vector of its primitive features (e.g., shape, texture, color), and a learning algorithm is then performed on the feature space created by the integration of the feature vectors. Therefore, decision making in a computer is based on both the feature descriptors and learning algorithms.

To study the semantic gap, we fix the learning algorithm (using k-means clustering) and explore the effects of various feature descriptors. Thus, in order to obtain the distribution of the labels from a computer perspective, first we extract the primitive features. Secondly, k-means is applied to the primitive feature description of each image, where the number of clusters is set to 18 (corresponding to the labels in Table 5.2). The obtained clusters represent the different labels, and their size corresponds to their occurrence. We then normalize the cluster occurrences and the user assigned label coverages in order to represent each image by two probability mass functions from the computer and user perspectives, respectively. These functions are then compared by symmetrized Kullback-Leibler divergence [17]:

$$D_{KL}(L_i||C_i) = \frac{1}{2}[\sum_{x=1}^{18} L_i(x)\ ln\frac{L_i(x)}{C_i(x)} + \sum_{x=1}^{18} C_i(x)\ ln\frac{C_i(x)}{L_i(x)}], \tag{5.2}$$

where L_i and C_i are the probability mass functions representing the distributions of the labels and clusters in an image i. In these functions, x is a discrete random variable indicating a label or a cluster in the label and cluster distributions, respectively. Due to the unsupervised nature of the k-means, the correspondence between the x in L_i to the ones in C_i is not clear. In our experiments, we fix L_i and shuffle the x in C_i and compare the resulting functions to find the best fitting one. We then consider the D_{KL} between the L_i and the fitted function as the distance between the label and cluster distributions.

5.2.1.2 Feature Descriptors

In order to process the images from their different properties (e.g., shape, texture, color), they are represented by 3 different types of feature descriptors and their combinations: SIFT, WLD, rgbHist, SIFT-Color, WLD-Color. Features are extracted in a dense way at every location on every image using a sliding window of 32×32 pixels with 50% overlap. SIFT represents the geometry-based features of the images such as edges and corners by 128 dimensional vectors (please refer to Section A.4). WLD descriptor represents textural patterns of an image as a vector, where the resulting feature vectors in our experiments have 144 dimensions (please refer to Section A.5). To obtain the SIFT and WLD descriptors the methods are applied to the gray-value of the images, while to generate SIFT-Color and WLD-Color, the methods are applied to the RGB channels separately. The resulting vectors are then concatenated to achieve the final feature vectors. Thus, the SIFT-Color and WLD-Color features are 384 and 432 dimensional, respectively. rgbHist extracts color information of an image. For each local window, it concatenates the color histograms of the RGB channels and represents it as a vector. The resulting rgbHist vector is 768 dimensional, composed of three 256 dimensional vectors (please refer to Section A.1).

5.2.2 Results and Discussion

In this section, based on our experimental results, we discuss object discrimination and labeling; followed by a discussion on the relationship between the sensory and the semantic gap. In addition, we explain the effects of the semantic gap on biasing image mining systems.

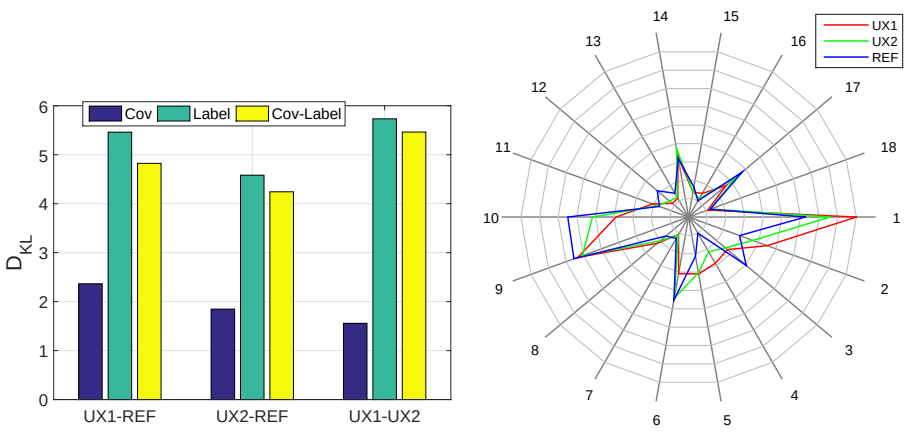
5.2.2.1 Object Discrimination and Object Labeling

In our experiment, users were asked to identify the objects in each patch, approximate what percentage of the patch area the object covered, and then label the object based on a given dictionary. This can be viewed as two tasks: one is a more perceptual task of visual segmentation of the patch into areas, grouping pixels according to similarity. Here the user is making a relative judgment- is each pixel like the neighboring one? And what overall area of the patch does this object cover? This task is affected by the sensory gap due to patch characteristics, such as resolution.

The second task is a more conceptual one- the user must identify what the object is and assign it a label from the dictionary in Table 5.2. This task is more difficult, since it involves making an absolute, as opposed to a relative, judgment. Previous research has found that annotators find ranking tasks (in which they make relative judgments) easier than assigning a precise score or classifying an image; and this type of task also produces a higher inter-annotator agreement [162].

The semantic gap associated with the visual segmentation task (identifying objects and assigning them percentages), and the labeling task is exemplified in Figure 5.7 (a). This figure shows the D_{KL} as a measure of the semantic gap between any two label distributions of the patches given by both UXs and the REF. D_{KL} is computed by first considering only the coverages (ignoring the label correspondences and by finding the best fitting distributions explained in Section 5.2.1.1); then only the labels (assuming the same probability for the occurred labels); and finally both together. Results show there is a higher degree of agreement, and lower semantic gap, when only the coverages are considered; whereas comparing only the labels results in a higher semantic gap. This demonstrates that visual segmentation and identification of objects is performed in a similar way by all users, compared to the object labeling.

It has been proposed that in order to determine the identity of an object, humans will turn to their memory to find an analogy - asking "what is it like?" (as opposed to "what is it?"). These analogies will result in memory associations, where additional information (e.g., context) will be considered, resulting in a prediction as to what the object is [163]. Considering the role of memory in prediction-making, it is natural that a person's background and experiences could play a role in their predictions [163]. It has also been noted that this prediction of what an object is can also affect what users see and where they consider the object's contours [164];



(a) (b)

Figure 5.7: (a) The semantic gap as the difference between two descriptions of an object, considering UX1 and REF, UX2 and REF, or UX1 and UX2. (b) Radar chart showing the average distributions of different labels in the patches based on UX1, UX2, and REF. Each number corresponds to a label as shown in Table 5.2.

and consequently, how they name the object. This brings us to the "vocabulary problem", which arises when people use different terms to describe the same object [15, 16]. A study involving spontaneous word choice for different domains revealed that there was less than a 20% probability of two people assigning the same label for a given object [15], exemplifying the linguistic semantic gap. This is reflected in Figure 5.7 (a), where the largest semantic gap is between UX1 and UX2, showing how even among users given the same images with a defined dictionary, there will not always be a consensus with regards to the semantics of an object.

These diverging understandings of label meanings can be further explained using Figure 5.7 (b). This radar chart shows the average distribution of different labels in the patches based on the REF, as well as the average distributions of the user assigned labels. The average distribution for the REF is calculated based on the coverages for all the patches (referring to the number of pixels corresponding to each label). For UX1 and UX2, the user-assigned coverages for each label were utilized. The deviation between the distribution of the UXs to REF for each label can tell us about missing labels, and confusion patterns between different labels. Since the area inside each plot is constant (because the values form a probability mass function), an increase in one dimension causes a proportional decrease in another dimension. A positive deviation of the UXs from REF (for example, in the case of "Agricultural field") indicates that the label was incorrectly assigned to different objects, or the coverage was overstated. The positive deviation in "Agricultural field" is compensated by the negative deviation of "Crop" for both UXs. A negative deviation of the UXs from the REF indicates that objects belonging to this category

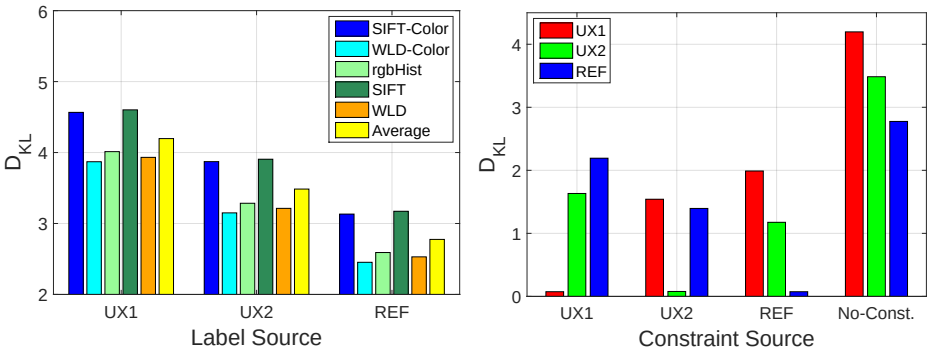


Figure 5.8: (a) The semantic gap between k-means and UX1, UX2, or REF; using 5 different feature descriptors (their average "Avg" is also depicted). (b) The average semantic gap over the 5 features, for UX1, UX2, and REF; while k-means is constrained by UX1, UX2, or REF.

were not detected, other labels were incorrectly assigned to this object, or the label coverage was understated. Furthermore, if we turn to "Grass", it is possible to observe a negative deviation for UX1. This is consistent with the user feedback in the questionnaires. Users commented that they were not always able to distinguish between the labels "Agricultural field", "Crop" and "Grass"; that the resolution of the image made the distinction between these semantic classes difficult, and that the labels themselves were difficult to define and differentiate. Taking all this together, it is possible to conclude that users from UX1 assigned the label "Agricultural field" to some objects that REF considered crops and grass. In the case of UX2, there is a small negative deviation for "Grass"; therefore, we can conclude that they misassigned the "Agricultural field" to crops in most cases.

Taking the differences in the semantic gap when comparing coverages to labels, and considering the user feedback regarding difficulties in labeling, it is possible to observe both a sensory gap (which is influenced by resolution and affects what is perceived in the image), and a semantic gap (which is influenced by confusion between labels, affecting the semantic labels given).

5.2.2.2 The Relationship between the Sensory and Semantic Gaps

In this section, we explain the relationship between the semantic and sensory gaps. Considering the semantic gap as the difference between the user and computer semantic understandings of images, we measure D_{KL} to compare the label distribution given by the UXs and the REF, to the distribution obtained by clustering the primitive feature descriptors for each image. The Y-axis in Figure 5.8. (a) shows this difference for 5 feature descriptors.

Semantic understanding is composed of both object perception and object naming. According to Section 5.1, the sensory gap affects object perception, which is in-

fluenced by the scene parameters (e.g., resolution), and the visual perceptual system. In our experiments, scene parameters are fixed; however, how objects are perceived by humans and computers is different. From the user side, since the perceptual system across humans is similar, the user sensory gap is considered to be similar for all three groups (UXs and REF) and is consequently disregarded. Therefore, only additional information (in the case of the REF) and user background (e.g., existing knowledge) can affect object naming, and thus semantic understanding. From the computer side, feature descriptors play the main role in object identification. By changing the feature descriptors, we can obtain different measures for the sensory gap, and consequently the D_{KL}. The other factor which affects the object classification and therefore the semantic understanding, is the learning algorithm, which we fixed to k-means. Therefore, in our experiments, two factors affect the semantic gap: the user background (or use of additional information), and the computer sensory gap (feature descriptors). In Figure 5.8. (a), the same pattern of D_{KL} for the feature descriptors across UXs and REF indicates the effect of the sensory gap from the computer side. By taking the average measure, we disregard the influence of the features to show the influence of user background or additional information. It shows that the average semantic gap for the REF is smaller than the average semantic gap for both UX1 and UX2, with UX1 being larger than UX2. This is consistent with the linguistic semantic gap shown in Figure 5.7. (a) where the distance between UX1 and REF is larger than the distance between UX2 and REF.

5.2.2.3 Effects of the Semantic Gap on Biasing Image Mining Systems

The demand for developing more efficient data mining systems has been met with methods usually performing based on human supervision in the form of annotated data, either for training or validation. Thus, different manually annotated datasets have been created; and are used for various purposes. However, according to research by Torralba & Efros [114], relying too much on a specific dataset for training and validating the proposed image information mining methods narrows down the research focus. The authors showed that in spite of efforts devoted to creating general and unbiased datasets, due to subjective and objective reasons (e.g., the purpose of the datasets), they suffer from strong built-in biases. The authors also doubted whether existing datasets reflect the expected real world scenarios. As a result, the verified systems based on reference datasets still do not provide results satisfactory to user requirements [14]. This has also been confirmed in [114] by training a model on a dataset and then testing it on another one. The results showed that the agreement is low even between datasets which appear to be similar.

The semantic gap, as the gap between user image understanding and that of computers, has been noted in previous research [13, 14] as a main reason behind the unsatisfactory results of current image information mining systems. Various schemes have been proposed to bridge the gap, which have been verified either by comparing

results to reference data, or by the degree of user acceptance in the interactive systems. Thus, although the proposed methods help bridging the semantic gap, they are biased to a dataset or to a user. Considering the interactive methods, for example, the gap between the system and a user become shorter as the user refines his request in each iteration; however, the resulting model may still not provide satisfactory results to other users. Based on our discussion in Section 5.2.2.1, the disagreement between the users' assigned labels can be due to the users' different needs and background knowledge.

In order to show the effects of this gap, in a new set of experiments we consider the effect of human interaction with the learning algorithm by constraining the k-means to the number of labels, given either by UX1, UX2, or REF. Figure 5.8. (b) shows the average semantic gap over the 5 features, for both UXs and the REF. The x-axis shows the group that defined the constraint. As the figure shows, when the learning algorithm is constrained by a group (e.g., UX1), the semantic gap between the learning algorithm and all the groups decreases, compared to the average when it is unconstrained. However, there is a significant decrease for the semantic gap between the learning algorithm and the group used to set the constraints (e.g., UX1). These results indicate that user interaction generally helps to shorten the semantic gap due to a basic common understanding between users; however, it biases the learning algorithm to that specific users' understanding of the image semantic.

To clarify what is meant by a common understanding between users, we will present an example. It has been shown in previous literature that using texture features improves the performance of the learning systems to a high degree in remote sensing tasks such as classification and segmentation [165]. In order to measure the performance of a learning system, its results are compared to a human-created reference data, which is biased by human perception, semantic understanding, and the task objective (what is expected of the data). Considering the reference as the basis for comparison, and considering the learning system's performance comes closest to it when using texture, we can conclude that texture features help humans in object identification which in turn biases the reference data. This is reflected in our experimental results in Figure 5.8. (a), which shows that the WLD-Color feature (which extracts textures) has the smallest semantic gap across all the groups.

Altogether, all existing methods proposed for bringing a system closer to a reference data or to a user decision, in principle shorten the semantic gap, although only some authors directly pointed this out in their publications [12, 13, 99, 117]. Moreover, only part of the high improvement achieved by bridging the gap is generalized, the bigger part is subjective and specific to that reference data or to the particular user.

5.2.3 Summary

The results of content based searches are not always satisfactory for users, due to the sensory and semantic gaps. Research on the semantic gap has considered differences between user and computer interpretations of an image, and proposed methods to bridge it. The proposed methods have been verified either by comparing results to reference data, or by measuring the degree of user acceptance in the interactive systems. Although these methods result in a narrower semantic gap between computers and users, the resulting model for a specific user and search goal may still not be satisfactory to other users. We show that the subjective biases present in the bridging methods, which we refer to as the linguistic semantic gap, cause this discrepancy. Furthermore, we show that the semantic gap builds on the sensory gap.

In order to overcome this problem, our proposal is that efforts to bridge the semantic gap should consider the linguistic semantic gap, and increase the diversity of data sets used in the domain (e.g., using various EO datasets in EO tasks), which will include different user perspectives and compensate the individual subjective biases. Moreover, models derived from proposed methods for bridging the semantic gap could be stored and further used by other systems, which would then be including other users' image interpretations.

5.3 Measuring the Semantic Gap Using an LDA-based Method

In this section, we introduce a method based on *Latent Dirichlet Allocation (LDA)* topic model [18] to measure the semantic gap, as the difference between the human understanding of the objects in an image and the computer interpretation of those objects. While bridging or shortening the semantic gap has been the focus of research in recent years [113, 118, 119], quantifying the semantic gap has been less often attempted [120]. However, developing a high performance image mining system with a narrow semantic gap requires quantifying and analyzing the gap.

In our proposed method, we extract the primitive features of the images using different feature description methods such as *rgbHist*, *Weber Local Descriptor (WLD)* [3], *Scale-Invariant Feature Transform (SIFT)* [70], *WLD-Color*, and *SIFT-Color*. The diversity of the employed feature descriptors allow us to study the content of the given image dataset from different aspects (color, texture, shape, and their combinations). In a next step, based on the extracted primitive features, the images are represented by a *Bag-of-Words (BoW)* model. Then LDA is applied to the BoW models of the images in order to discover the hidden semantic structure behind the image dataset as a set of *topics*. The images are then represented based on the occurrence of the topics, the so-called *Bag-of-Topics (BoT)* model. This model has been introduced and explained in detail in Section 3.1. Since the topics are discovered automatically and usually refer to semantics (detected objects and parts thereof), we consider the BoT models of the images as the computer's semantic interpretation of the images. Consequently, we measure the semantic gap as the similarity between the BoT models of the images and the users' semantic labeling of the images.

In order to verify our semantic gap measurement, we use the obtained semantic gap for various primitive features extracted from the UC Merced Land Use dataset [5], to predict the classification accuracy of a *Support Vector Machine (SVM)* [136] applied to the dataset. The experimental results show that the automatically measured semantic gap is able to predict the classification performance of the SVM for each primitive feature descriptor. Since the classification is supervised by the user labeling of the dataset, the classification error represents the deviation of the computer results from the user semantics. This is obtained based on the object discrimination using the image primitive features and the SVM algorithm (for details, please refer to Section 5.2). Therefore, measuring the semantic gap allows us to predict the user acceptance of the image mining systems (e.g., SVM classification together with image feature descriptors) when they are verified based on a reference dataset created by users.

5.3.1 Methodology

In our proposed method for measuring the semantic gap, we use LDA to automatically discover the semantics of image patches in form of topics. To this end, based on the extracted primitive features, image patches are represented by a BoW model as vectors in an Euclidean space, the so-called *visual word space*. LDA is then applied to the BoW models of the image patches to discover the latent semantics as a set of topics. In a next step, each image is represented as a distribution over the discovered topics (BoT). By assuming the topic distributions as vectors, BoT models of the image patches form an Euclidean space, the so-called *topic space*. Since usually the topics refer to semantics, BoT models of the image patches are considered as the semantic descriptions of the image patches. Thus, the images containing similar semantics are located close to each other in the topic space. Moreover, due to the automatic discovery of the topics, the distribution of the images in the topic space can be assumed as the semantic interpretation of the image dataset by a computer, which is based on the image primitive features (for details, please refer to Section 5.2).

In order to compare the computer's semantic interpretations of the images to those of the users', we use a user labeled image dataset. Thus, the images within each image class contain the same semantics as decided by the users. As a next step, the *geometric median* [166] of the images within each class is computed for the topic space and for the visual word space. While a median point in the topic space indicates the computer's semantic interpretation of a particular class, a median point in the visual word space indicates the users' semantic interpretation of the class. The semantic gap is then measured as the distance between the two geometric median points of each class (obtained from the BoW model and from the BoT model). However, since the two points are not in the same Euclidean space, we first map the geometric median points within the topic space, into the visual word space using the generative property of LDA.

The mathematical explanation of our proposed method is as follows. Assume each image $\mathbf{w}_d$ is represented by a BoW model as $\mathbf{w}_d = \{w_{d1}, w_{d2}, ..., w_{dN_d}\}$, where each visual word w_{di} is drawn from a fixed dictionary of N_V visual words (codewords), $V = \{v_1, v_2, ..., v_{N_V}\}$. The probability of each visual word v_l within each image patch, $p(v_l|\mathbf{w}_d)$, is then computed based on the visual word's occurrence. Thus, the BoW model of each image patch can be represented as a probability mass function over the visual words within the dictionary V. LDA is then applied to the images' BoW models to discover the latent semantics of the images as a set of K topics, $T = \{t_1, t_2, ..., t_K\}$. Each topic is then defined as a probability mass function over the visual words v_l ($l \in [1, N_V]$). Each image patch $\mathbf{w}_d$ is then represented as a mixture of topics defined by the multinomial distribution θ_d. Thus, the geometric

median of each image patch class L_i is computed by:

$$\theta_{med_{L_i}} = \arg\min_{\mathbf{y}} \sum_{\theta_d \in L_i} \|\theta_d - y\|_2 \,. \tag{5.3}$$

The geometric medians are then mapped to the visual word space through the generative process of LDA (for details, we refer the reader to Section 1.4.1.3.1) which represents each geometric median point as a probability mass function over the visual words of the dictionary V, where the occurrence of the visual word v_l for a geometric median point $\theta_{med_{L_i}}$ is obtained by:

$$p(v_l|\theta_{med_{L_i}}) = \sum_{j=1}^{K} p(v_l|t_j)p(t_j|\theta_{med_{L_i}}), \tag{5.4}$$

where $p(v_l|t_j)$ is the probability of the visual word v_l within topic t_j which is obtained in the parameter estimation phase of the LDA method. Furthermore, $p(t_j|\theta_{med_{L_i}})$ is the probability of every topic according to the geometric median $\theta_{med_{L_i}}$, and it is simply equal to $\theta_{med_{L_i}j}$ (for details, please refer to Section 1.4.1.3). The geometric median of each class is computed for the BoW models of the images. Assuming the geometric medians of each class (obtained from the BoW model and from the BoT model) as vectors in N_V-dimensional visual word space, the Euclidean distance between them is measured as the semantic gap for that particular class.

5.3.2 Results and Discussion

In our experiments, we use various local features of the images of the UC Merced Land Use dataset (for details please refer to Section B.3) using different feature extraction methods such as rgbHist, WLD, SIFT, SIFT-Color, and WLD-Color. These represent different properties of the images such as color, texture, shape, and their combinations. This diversity of feature descriptors allows us to study the effects of different image properties on the semantic gap. The images' local primitive features are extracted densely using a sliding window of 32×32 pixels. The rgbHist feature vectors are generated by concatenating the histograms of the pixel values for the RGB color channels (with 256 bins each), which results in 768-dimensional feature vectors. The WLD and SIFT feature descriptors are computed from the gray values of the images according to their original papers [3, 70]. Moreover, in order to build WLD-Color and SIFT-Color, WLD and SIFT are applied to each color channel separately; and the resulting feature vectors are then concatenated (for details please refer to Appendix B).

In a next step, we build the BoW model of the images based on the primitive feature vectors. In order to assess the influence of the number of visual words on the semantic gap, the BoW models are built for different codebook sizes (20, 50, 100,

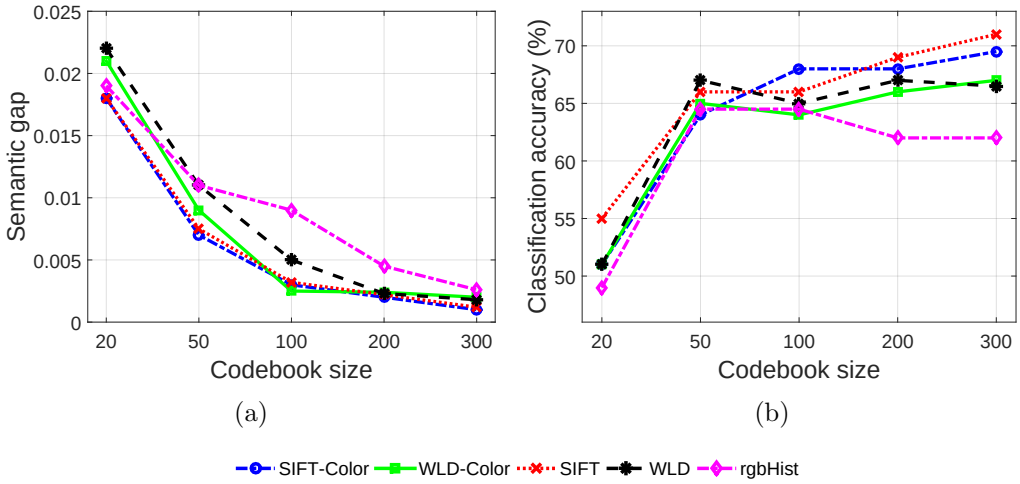


Figure 5.9: (a) Measuring the semantic gap as the Euclidean distance between the two geometric median points in visual word space, versus the codebook size for different primitive feature descriptors. (b) Classification accuracy of an SVM method for various feature descriptors versus codebook size.

200, and 300 codewords). Then LDA is applied to the BoW models of the images to discover the latent semantic structure of the image dataset as a set of topics. In UC Merced Land Use dataset the images has been grouped into 21 image classes by human users, which show how many semantic concepts has been considered by the users to discriminate the images. Thus, we preset the number of topics in LDA to 21, in order to discover the same number of semantics as that of the users, which makes the two sets of semantics (the computer's and the users') comparable. In a next step, we describe each image as a mixture of the discovered topics (BoT). This image description is then used to compute the geometric median of each image patch class. Using the generative property of LDA, in a next step, the geometric medians are mapped to visual word space. In addition, we compute the geometric medians of the image patch classes using the BoW models, in visual word space. The semantic gap for each class is then computed as the Euclidean distance between the two geometric medians (the one computed using the Bot models and the one computed using the BoW models).

Figure 5.9. (a) shows the computed semantic gap for different feature descriptors and for different codebook sizes. The results show that various feature descriptors cause different semantic gaps. For example, rgbHist caused the largest semantic gap which indicates that the users relied on the images' color properties less than their other properties (e.g., texture) during the manual image labeling process. This can be further justified by the diversity of the color features within each image class of the UC Merced Land Use dataset. Moreover, the results show that increasing the codebook size up to a certain point (100 codewords in our experiments) significantly

decreases the semantic gap for all the feature descriptors. According to Section 3.1.2, since topic representations of the images (BoT) build upon the images' BoW models, the larger the number of the visual words helps LDA to better estimate the latent semantics of the dataset. However, after a certain number of visual words, increasing the codebook size does not contribute significant new semantics to the topic discovery process.

Furthermore, we use the obtained semantic gap, to predict the classification performance of an SVM applied to the dataset. Figure 5.9. (b) shows the accuracy of SVM for the five different images' primitive feature descriptions. Comparing the classification results to the measured semantic gaps for different feature descriptors indicates that a decrease in the semantic gap tends to an increase in the classification accuracy. Moreover, for the primitive features which causes a larger semantic gap the classification accuracy is lower. For example, by using rgbHist we have on average the largest semantic gap and the lowest average classification accuracy. The results verify that the proposed method for measuring the semantic gap can be used to predict the performance of image mining systems when their results are compared to a user created reference dataset.

5.3.3 Summary

In this section, we proposed a method based on an LDA topic model for measuring the semantic gap, where the semantic gap is defined as the difference between human understanding of the objects in an image and the computer interpretation of those objects. In this method, LDA is applied to the BoW models of the images in order to represent the latent semantic of a given dataset as a set of topic. Using the discovered topics, each image is then represented as a mixture of the topics, the so-called BoT model. Since the topics are discovered automatically and usually refer to semantics, we assume the BoT models of the images as the computer's semantic interpretation of the images. For each class, the geometric median of the images' BoT models is computed, which represent the computer's semantic interpretation of the class. This point is then mapped to a visual word space, an Euclidean space formed by the BoW model of the images. Moreover, in visual word space, for each class a geometric median point is computed using the BoW model of the images, which is assumed to be the users' semantic interpretation of the class. The Euclidean distance between this point and the geometric median point computed from the BoT models is then considered as the semantic gap for that particular class.

We verified the obtained semantic gap to predict the classification accuracy of an SVM method using various primitive feature descriptions of the images. The experimental results showed that a decrease in the semantic gap tends to an increase in the classification accuracy. Moreover, the primitive features which cause smaller semantic gaps lead to higher classification accuracy.

6

Feature Space Exploration and Evaluation

In this chapter, we study the semantic content of image datasets by evaluating and exploring the topology of the feature space generated from their extracted features.

As an evaluation method, we propose a new approach for using clustering techniques to assess the topology of the feature space. In this approach both internal and external clustering evaluations are employed. When we use a particular feature descriptor to represent image patches, an internal evaluation demonstrates the degree of homogeneity and the discriminability of the image patch descriptions, while an external evaluation measures the similarity between the resulting clusters and the user labeling of the image patches (user acceptance).

For exploring the feature space, we propose to use the environment of a *Visual Data Mining (VDM)* system, the so-called *Cave Automated Virtual Environment (CAVE)*. This system allows users to navigate inside the feature space and explore the structure of the image semantics.

Hence, the proposed methods allows the selection of the most appropriate feature descriptors for representing an image dataset. In addition, they help to develop new feature descriptors so as to classify image datasets into semantically meaningful categories.

6.1 Evaluation of Feature Space Based on Clustering

Most existing image mining systems rely on feature spaces representing various properties of EO images. Therefore, assessing the topology of the feature space is essential in developing new feature descriptors and image mining systems. In this section, we propose a new approach for using clustering methods to evaluate the feature space structure. The main idea behind this method is to cluster the given image

patch dataset for different numbers of clusters; then we can evaluate the clusterings both internally and externally. The internal evaluation relies on the unsupervised nature of clustering to find optimum clusters. The optimum clusterings obtained for different feature descriptors such as *Gabor* filter banks [167], *Weber Local Descriptor (WLD)* [3], and *Rand_Feat* [28] (for details please refer to Appendix A), indicate the homogeneity and discriminability of the image patches according to their different properties (e.g., texture, color). In addition, the external evaluation allows comparing the resulting clusters to a user image labeling. Considering the image labeling as the user image understanding, the external evaluation indicates how different image properties are understood (consciously or unconsciously) by the user during image semantic recognition to group or discriminate the image patches. While previous works rely on either internal or external clustering evaluation [168, 169, 170, 171], in our work, we show that both evaluations are essential to investigate the feature space structure.

Evaluating the feature space structure not only allows choosing more descriptive feature descriptors, but also helps to develop more sophisticated descriptors in order to categorize image patch datasets into semantically meaningful categories. Furthermore, as a practical application, our proposed approach can be used in data annotation tasks, before and after annotation to ease and validate the annotation procedure. While an internal evaluation provides the user with an overview of the structure of the dataset, an external evaluation validates the annotated classes.

In order to demonstrate our approach, we apply it to the Seven Class TerraSAR-X image patch dataset (please refer to Section B.1). Then quantitative results as well as visualizations of feature space for three different feature descriptors are provided for the internal and the external evaluations of the feature space structure.

6.1.1 Internal and External Clustering Evaluations

Due to the unsupervised nature of clustering, the validation of the resulting set of clusters is challenging. Since the fundamental purpose of using clustering methods is to discover the unknown structure of data, validation without using external knowledge such as image labels, is highly important. In addition, in real world scenarios, since usually the data structure is unknown, image labels do not always correspond to the natural grouping of the images. Therefore, a given user labeling is not sufficient for validating a clustering result [168]. Internal clustering evaluation methods allow us to analyze and to approximate the structures behind given image datasets by finding an optimal clustering of the images regardless of already existing image labels. A variety of methods have been proposed for internal clustering evaluation such as the *S_Dbw* validity index [169], the *Calinski-Harabasz (CH)* approach [172], and the *Davies-Bouldin (DB)* index [173]. Moreover, some previous works compared and validated these methods from their different aspects [174, 175]. Based on their conclusions and our experimental results, S_Dbw provides a more effective cluster

analysis than the other methods due to considering at the same time both main clustering criteria, namely compactness and separability of the clusters [176]. Moreover, the authors of [174] concluded that the resulting index of S_Dbw is rather stable against monotonicity, noise, density problems, sub-clusters, and skewed distributions. Considering all the above, we use S_Dbw as the internal evaluation method in feature space assessment.

In addition to the internal evaluation, we use an external cluster indexing method to assess the closeness of user assigned image labels to the image clusters. Several measures have been introduced for external cluster indexing such as *Adjusted Random Indexing (ARI)* [177], *F-measure* [171], and *Adjusted Mutual Information (AMI)* [170]. Among them, we use ARI which has been proved in [178] to provide reasonable external clustering evaluation.

6.1.1.1 The S_Dbw Validity Index

The main idea behind any clustering method is to partition a set of points so that the points within a single cluster are similar, whereas the points in different clusters are distinct. S_Dbw is an internal validation measure which considers the two essential clustering criteria, namely within-cluster similarity and between-cluster distinguishablity [169]. In order to measure the within-cluster similarities, S_Dbw computes the average scattering of the data points in clusters as:

$$Scat = \frac{\frac{1}{k}\sum_{i=1}^{k}||\sigma(s_i)||}{||\sigma_D||}, \tag{6.1}$$

where k is the number of clusters, $||\sigma(s_i)||$ is the variance of cluster s_i, and $||\sigma_D||$ is the variance of the entire dataset. In order to measure the distinguishability of the clusters, the average density between the clusters is computed as follows:

$$Dens = \frac{1}{k.(k-1)}\sum_{i=1}^{k}[\sum_{j=1,j\neq i}^{k}\frac{f(s_i,p_{ij})+f(s_j,p_{ij})}{max\{f(s_i,c_i),f(s_j,c_j\}}], \tag{6.2}$$

where $f(s_i, p_{ij})$ is the number of points grouped in cluster s_i which lie at a given distance from a point p_{ij}. In our experiments, this distance is defined as the average variance of the dataset ($\frac{1}{k}\sum_{i=1}^{k}||\sigma(s_i)||$). Moreover, p_{ij} is the middle point on the line connecting the two cluster centers c_i and c_j. Finally, the S_Dbw measure is computed as follows:

$$S\_Dbw = Scat + Dens. \tag{6.3}$$

6.1.1.2 The Adjusted Random Index

Comparing different clusterings of a set of data points has been always a challenge. ARI is a partition comparison method based on a $k \times k'$ co-occurrence matrix, where

k and k' are the number of clusters in the two given clusterings $S = \{s_1, s_2, ..., s_k\}$ and $\tilde{S} = \{\tilde{s}_1, \tilde{s}_2, ..., \tilde{s}_{k'}\}$, respectively. The matrix values show the number of co-occurrences among the data points in the clusters. Basically, this method measures how different pairs of data points are treated in each clustering. More precisely, how many pairs are grouped together in both clusterings (GG), how many of them are separated in both clusterings (SS), and how many of them are grouped in one clustering while being separated in the other one (GS). Using these values, a *Random Indexing (RI)* is then computed as follows:

$$RI(S, \tilde{S}) = \frac{GG + SS}{GG + SS + GS}. \tag{6.4}$$

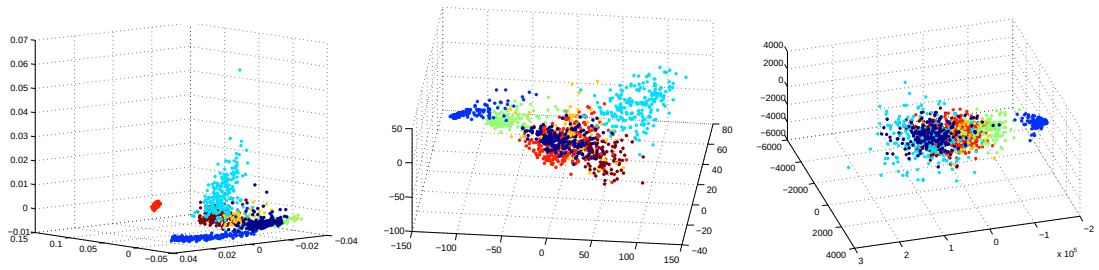
Since the expected value of RI is changing in every experiment, which leads to unfair comparison of the clusterings, an adjusted version, the so-called ARI has been introduced by Hubert and Arabie [177] defined by:

$$ARI(S, \tilde{S}) = \frac{RI(S, \tilde{S}) - Expected\ Index}{Max\ RI(S, \tilde{S}) - Expected\ Index}, \tag{6.5}$$

where the *Expected Index* is the expected value computed from (GG + SS) in a fixed experimental setup (e.g., for two clusterings, the original number of clusters and the data point cluster assignments are considered as a random assignment). Moreover, $Max\ RI(S, \tilde{S})$ is equal to 1. Thus, ARI is equal to 1 when the two clusterings are identical and is equal to 0 when the RI of the two clusterings is equal to the *Expected Index*.

6.1.2 Results and Discussion

In our experiments, we assessed the feature space structure of an EO image patch dataset, namely our Seven Class TerraSAR-X image patch dataset, both internally and externally. Three different feature spaces were created using Gabor, WLD, and Rand_Feat feature descriptors (for a detailed explanation about the descriptors, please refer to Appendix A). In our experimental setup, the Gabor features are constructed for 3 scales and 6 rotations resulting in feature vectors of 36 elements. Moreover, for WLD, we set the number of excitations and orientations to 6 and 8, respectively; this results in a feature vector of 144 elements. In order to exploit the pixel brightness values, usually histograms of pixel values are constructed. Since the brightness range of SAR images is rather wide, which results in a very large vector, constructing pixel value histograms is not trivial. Therefore, we put all the pixel values of an image $I_{i \times j}$ into a vector of size $m = i \times j$. In this way, the resulting vector is much smaller than the brightness histogram; however, it is still too large to be used efficiently in a clustering process. In order to reduce the dimensionality of this vector from m to $\tilde{m}$, we use *Random Projection (RP)* [27] which computes



(a) (b) (c)

Figure 6.1: Three-dimensional visualization of the feature space for WLD, Gabor, and Rand_Feat feature descriptors. Different colors represent different image classes (e.g., Forest, Water, Medium density urban area, Forest + Water, Road, High density urban area, and Urban area + Road). (a) WLD. (b) Gabor. (c) Rand_Feat.

the product of the high dimensional vector an a $m \times \tilde{m}$ random matrix ($\tilde{m} \ll m$). In our experiments, by setting $\tilde{m} = 32$, the resulting vectors have 32 dimensions. Figure 6.1 shows a three-dimensional visualization of the feature spaces built by the three feature descriptors. In this figure, the different colors represent different classes according to the user assigned labels.

In the next step, we apply k-means clustering to the feature descriptors and then evaluate the clusterings using the S_Dbw and ARI measures. The ideal clusters are compact with a low density of feature points between clusters. Since the S_Dbw measure does not depend on prior image labeling, the clustering is not penalized for discovering new clusters or finding a different structure to the user's understanding of the images. In addition, the ARI measure allows finding the similarity degree between the feature space structure discovered by the clusterings and the user assigned image labels. Moreover, measuring ARI for different types of feature descriptors shows that according to which image properties the users labeled the image patches.

6.1.2.1 Internal Cluster Evaluation

In our experiments, we use the S_Dbw measure to evaluate the clusterings of Gabor, WLD, and Rand_Feat descriptors for different numbers of clusters. Figure 6.3 shows some clusterings of the feature spaces built by the three feature descriptors. In this figure, the different clusters are depicted by distinct colors. The internal evaluation results of the clusterings are demonstrated in Figure 6.2. This illustration shows the average scattering of the clusters (Scat), the average between-cluster density (Dens), and their combination which is S_Dbw.

As the results of Figure 6.2. (a) show, the average scattering decreases monotonically when the number of clusters is increased; however, beyond a certain number of clusters the variations become insignificant. Comparing the three feature descrip-

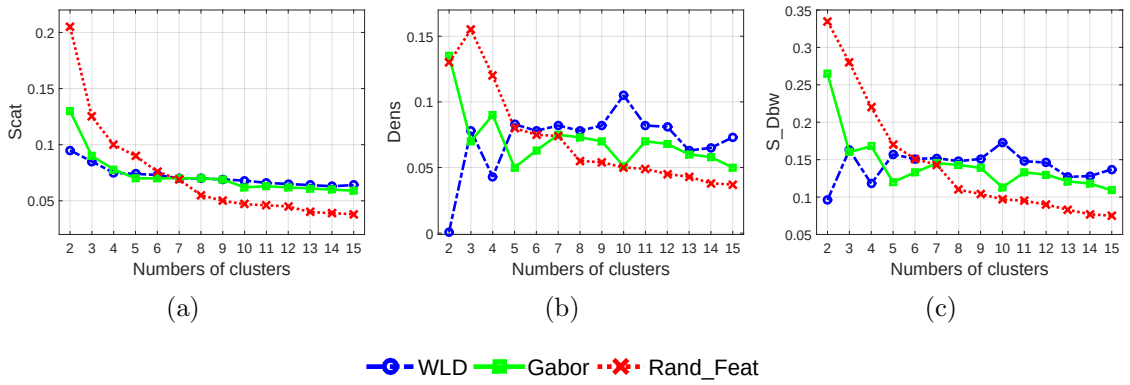


Figure 6.2: Internal evaluation of clusters for WLD, Gabor, Rand_Feat descriptors and various numbers of clusters. (a) Average scattering. (b) Average density between clusters. (c) S_Dbw measure.

tors, the average scattering decreases more rapidly for Rand_Feat than for Gabor and WLD, which means that the feature space structure for the Rand_Feat method is sparser than those of the two others. Moreover, Figure 6.2. (b) shows that the average between-cluster density does not change monotonically when the number of clusters increased. In addition, the overall behaviors of the curves are rather different among the feature descriptors. This indicates that the average between-cluster density is highly sensitive to the feature space structure. An optimum clustering would provide clusters with a minimum density of points between clusters. According to Figure 6.2. (b) the optimum clustering for WLD is obtained by partitioning the feature space into 2 partitions. This is consistent with the visualization of the feature space as two clearly separated clouds of feature points as shown in Figure 6.3. (a). Furthermore, by increasing the number of clusters to 3 as depicted in Figure 6.3. (b), the average between-class density increases significantly due to the splitting one of the separate clouds into 2 clusters, resulting in a large interfacing region. By further increasing the number of clusters to 4 as shown in Figure 6.3. (c), the average density is decreased to some extent, because the newly added cluster introduces a small interface to the other clusters, which decreases the average between-cluster density. Comparing Figure 6.2. (b) to Figure 6.3, the behaviors of the between-cluster density curves provide a general intuition about the structure of corresponding Gabor and Rand_Feat feature spaces. Figure 6.2. (c) shows the S_Dbw measure of the clustering in which the optimum clusterings are achieved at the minima of the curve. According to this figure, for WLD and Gabor, there could be multiple local minima which means there are multiple best clusterings (e.g., for WLD we have optimum solutions for 2, 4, 9, and 13 clusters).

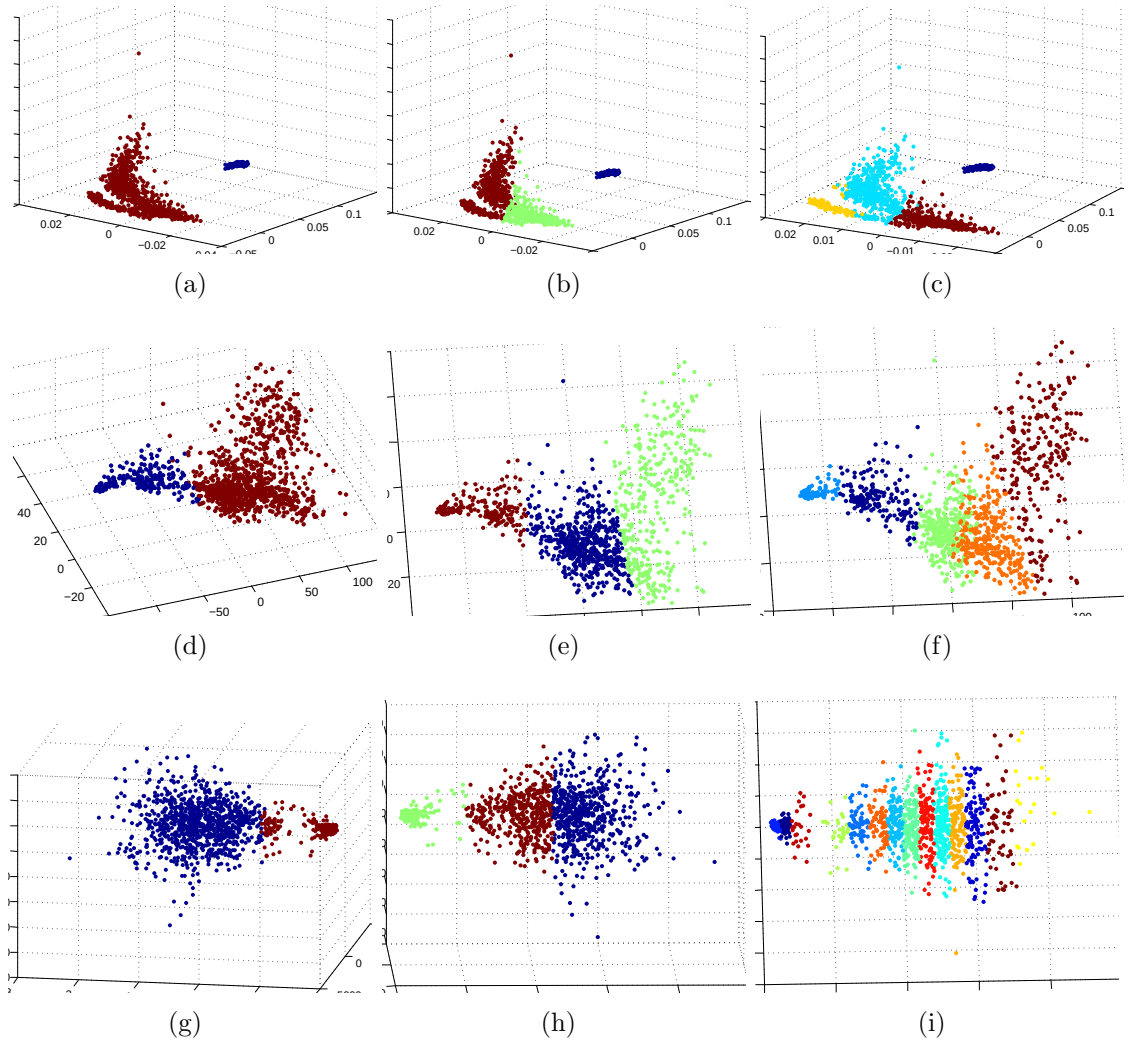


Figure 6.3: 3D visualization of sample clusterings of WLD, Gabor, and Rand_Feat feature spaces. The different colors denote distinct clusters. (a) WLD for 2 clusters. (b) WLD for 3 clusters. (c) WLD for 4 clusters. (d) Gabor for 2 clusters. (e) Gabor for 3 clusters. (f) Gabor for 5 clusters. (g) Rand_Feat for 2 clusters. (h) Rand_Feat for 3 clusters. (i) Rand_Feat for 13 clusters.

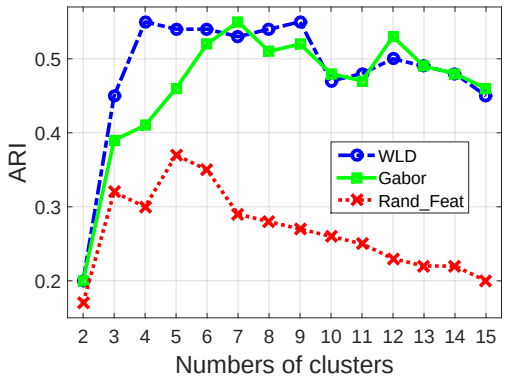


Figure 6.4: ARI measure computed for WLD, Gabor, and Rand_Feat descriptors for various numbers of clusters.

6.1.2.2 External Cluster Evaluation

As the results for internal clustering show, the optimum number of clusters is not necessarily equal to the number of classes introduced by the image labeling. Moreover, a comparison between the clusters in Figure 6.3 and the image labels in Figure 6.1 shows that the points which are assigned to one single class by the users are not necessarily grouped in the same cluster. This demonstrates that user semantic labeling does not necessarily correspond to the properties represented by feature descriptors. In order to compare the clustering results and the user labeling, we performed an external clustering evaluation using the ARI measure. Figure 6.4 shows ARI computed for WLD, Gabor, and Rand_Feat descriptors for different numbers of clusters, where a larger ARI value means the clustering and the labeling are closer. According to the figure, the structures represented by WLD and Gabor are more similar to the user labeling (user image understanding). In other words, users rely more on image textures than on average image brightness to discriminate and label image patches. Moreover, Figure 6.2. (c) and Figure 6.4 indicate that the optimum clusterings are not necessarily the most similar ones to the user labeling.

6.1.3 Summary

In this section, we introduced a clustering-based approach to evaluate EO image content structures. In this approach, after representing image patches by feature descriptors, we perform clustering on the feature space (built by the feature descriptors) for different number of clusters. The resulting clusterings are then evaluated both internally and externally (i.e., without and with using prior annotation, respectively). While an internal evaluation shows the feature space structure by finding the optimum clusters, an external evaluation allows comparing the obtained clusters to the user labeling of the image patches.

Experimental results indicate that there are multiple optimum clusterings of image patches depending on different levels of image properties (e.g., from low level shape and texture to higher level contexts). Furthermore, an external evaluation demonstrates that the data structures discovered through clusterings do not necessarily correspond to user labeling (i.e., user image understanding). Moreover, comparing the clusterings across different feature descriptors shows that users consider some image properties more than other ones, for example, users consider texture features more than average brightness of images.

6.2 Exploring the Feature Space Using a Visual Data Mining System

The exponential growth in the amount of various types of visual data such as EO and multimedia images has made exploration, analysis, visualization, and understanding of the available data a gigantic challenge. In addition, issues such as the sensory and semantic gaps limit the user understandings of the image semantics. One strategy which has been proposed and followed by a number of researchers during the past decade is to employ *Visual Data Mining (VDM)* [179] systems in order to provide users an interactive scenario for the image mining process. The existing VDM systems are usually composed of a user interface to visualize images and allow users to interact with the the system (either by manipulating the visualization or providing feedbacks to a machine learning algorithm), and a machine learning algorithm which controls the visualization process by learning from user feedback.

In this section, we focus on the visualization part of VDM systems and discuss the importance of visualization for a better understanding of the semantic image contents. Visualization of the images based on their extracted features represents images to the users similar to how a computer interprets these images (i.e., in the form of feature vectors). The knowledge which users gain by this visualization and interactive exploration into the images helps to shorten the semantic gap (i.e., the gap between users and a computer) in two ways: 1) by creating reference data which are closer to the computer's semantics; 2) by designing and developing feature descriptors that interpret images in the feature space in a way more understandable to human.

This part of our research has been conducted in the framework of a joint project of the *Munich Aerospace* faculty[1], namely *Immersive Visual Information Mining for the TerraSAR-X/TanDEM-X Archive* between the *Remote Sensing Technology Institute (IMF)* of the *German Aerospace Center (DLR)*[2] and the *Institute for Human-Machine Communication (MMK)*[3] of the *Technical University of Munich*[4]. The goal of the project was to develop a new interactive image mining system to provide users with a more effective visualization of images and an efficient exploration and interaction capability. Our contribution to this project was extracting various features from images and analyzing the structure of the resulting feature spaces based on the visualization provided by MMK's virtual reality lab.

[1] http://www.munich-aerospace.de/index.php/en/
[2] http://www.dlr.de/dlr/en/desktopdefault.aspx/tabid-10002/
[3] http://www.mmk.ei.tum.de/en/home/
[4] https://www.tum.de/en/homepage/

6.2.1 Visual Data Mining Systems

A *Query by Example (QE)* is the simplest human-machine interaction system which allows users to query a sample image as the only interaction to the system. QE then visualizes images similar to the user's query; however, due to the semantic gap, the visualized images might be unsatisfactory to the users. In order to deliver a human understanding of the images to the system, *Relevance Feedback (RF)* [180] and *Active Learning* systems [82, 181] have been proposed. Even though the visualization part in these systems is the same as for QE (i.e., presenting a number of relevant images to the users), users are allowed to feedback the system by accepting or declining the visualized images. Based on the users' feedbacks, the machine learning algorithm of the system is tuned to the users' understanding of the images. The first significant improvement in the visualization part of VDM systems has been conducted by Yang *et al.* [182], where the features of the images are extracted in form of high-dimensional feature vectors. The dimensionality of the feature vectors is then reduced and images are visualized in a two-dimensional environment. This visualization allows users to browse the images based on their semantic contents, derived from the image features.

The increasing interest in visualizing the images based on their features has led to proposing various VDM systems with two-dimensional or three-dimensional environments [179] for visualization. For example, a VDM system developed based on *Virtual Reality (VR)* technology visualizes a given image collection as a cloud of points in a stereoscopic three-dimensional environment on a computer screen [183, 184] according to their extracted features. In addition, the points are colored according to the previously assigned labels. These systems are considered to be a step forward in image analysis by allowing users to navigate within the images and explore them based on their features. Despite their advantages, these proposed systems suffer from a number of limitations. For example, the effectiveness of the visualization of the images on the stereoscopic screens become limited by increasing the number of images (data points). Moreover, illustrating the images as points in the feature space cannot fully reveal the semantic content of the images.

In order to deal with the limitations of the previous VDM systems, Nakazato and Huang [185] proposed an active learning system which visualizes a number of the most relevant images to a given query based on their extracted features within a three-dimensional virtual environment, the so-called *Cave Automated Virtual Reality (CAVE)*. The users are then allowed to navigate inside the CAVE's environment and interact with the images and the system. In addition, visualizing the images (as opposed to the colorful points of the previous systems) helps the users to better understand the semantic content of the images. However, visualizing only a number of relevant images to the users, biases the users' image understanding to their primary assumption of the image semantics; which causes not only a large semantic gap between the users and the computer, but also a large gap between

different users' image understandings, i.e., the linguistic semantic gap (for details, please refer to Section 5.2). In order to remedy the shortcomings of the existing VDM systems, a *Learning by Immersive Visual Data Mining* system was developed in MMK [186, 187, 188, 189]. In this system, the performance of the machine learning algorithm as well as the dimensionality reduction technique are leveraged in order to provide an adaptive visualization mechanism. This mechanism visualizes the entire image collections in the CAVE and helps users to better understand the semantic structure of the images. By using the visualization provided by the developed VDM system, in this section, we assess the structure and the semantics of the images based on their extracted features. In the following section, we review the visualization environment of the developed system. The details of the learning and visualization process of the system are out of the scope of this dissertation.

6.2.2 Feature Space Visualization within the CAVE

In order to represent the semantic contents of images, we visualize the images in the CAVE based on their extracted features. The CAVE is an interface assisting users in better understanding data structures by allowing them to navigate and explore the data. It is composed of four large walls which are used as four display screens. The walls are aligned to form a cubic space which provides the users with a 180 degree horizontal view. For providing a stereoscopic scenario, two projectors are used for displaying a scene on each wall. In addition, six infrared cameras are mounted on top of the walls surrounding the environment. These cameras, together with a *Personal Computer (PC)*, form a real-time tracking system which tracks the users' positions and activities inside the environment. A user can then navigate inside the data and interact with the data (e.g., by selecting an image) inside the CAVE's environment by wearing polarized glasses (to produce a stereoscopic effect) and using a hand-held remote control. Users are able to navigate in four directions (left, right, forward, and backward), and to rotate 180 degrees horizontally. In addition, by changing the orientation of the control, users are able to navigate along the direction of the control allowing them to move toward a desired point. Furthermore, users are allowed to zoom in and out over a desired area in the direction of the control by pressing a button on it. As a visualization software, "3DVIA Studio4"[5] (an interactive three-dimensional visualization software widely used in VR scenarios) is employed in the CAVE which used for our visualizations.

In order to visualize the images based on their features, we first represent images as high-dimensional vectors of their extracted features (i.e., feature vectors) using a feature extraction method (e.g., rgbHist, WLD, SIFT). For positioning the feature vectors within the three-dimensional environment of the CAVE, a dimensionality reduction technique (e.g., *Principal Component Analysis (PCA)* [190], *Nonnegative*

[5] http://www.3dvia.com/studio/

Matrix Factorization (NMF) [191], *Farness preserving Nonnegative Matrix Factorization (NMF)* [192]) is then employed. Figure 6.5 shows the visualization of given image patch collections within the CAVE, where a user navigates and explores the image semantics by wearing polarized glasses and holding a control.

6.2.3 Results and Discussion

In this section, we show example visualizations of images within the CAVE. Since taking pictures from the stereoscopic environment of the CAVE is rather complicated, we simulate the visualizations on a computer screen using "DataVis3" [193], a VDM system. By using a dimensionality reduction technique, each extracted high-dimensional feature vector which describes an image patch, is mapped to the three-dimensional environment of the visualization system.

Figure 6.6 illustrates the visualization of multispectral image patches based on their color features extracted by rgbHist method. The images have been acquired by the WorldView-2 satellite from the two cities of Munich (in Germany) and Venice (in Italy). As illustrated in Figure 6.6. (a), the image patches from the two cities are separately grouped within the feature space. Exploring the image patches helps us to understand the semantics of the image patches, for example, the similarities between the landscapes and architectures of the two cities. By navigating through the images as illustrated in Figure 6.6. (b), we can observe that the general landscapes of the two cities are diverse to a high degree. While Venice is surrounded by sea water, Munich is located within fields and forests. On the other hand, these two cities are similar in urban areas where the buildings have have red roofs. If we have more samples from other European cities with common features (i.e., the red roofs in urban areas), one could, for example, use this feature to develop feature extraction methods used to discriminate European cities from other cities around the globe. The developed feature descriptors should be able to consider the reddish feature of the roofs while neglecting the features from the landscapes.

Figure 6.7 and 6.8 exemplify the visualizations of two SAR image patch collections based on their texture features extracted using the Gabor method. The images have been acquired by the TerraSAR-X satellite. The image patch collection visualized in Figure 6.7 is grouped into three clusters. While two clusters are spherical (i.e., the ones with lower brightness), the third cluster is elongated, displaying image patches from brighter to the darker ones. By having a closer look into the image patches in Figure 6.7. (b,c), we can observe that the images within the spherical clusters are rather homogeneous in texture (e.g., containing water surfaces); whereas the elongated cluster extends from the patches with highly structured contents such as industrial areas and storage tanks, to the more homogeneous image patches which partially to totally containing water surfaces, such as ocean. This part of the cluster is very close to the cluster containing the image patches of water surfaces. Furthermore, although the image patches within the two spherical clusters

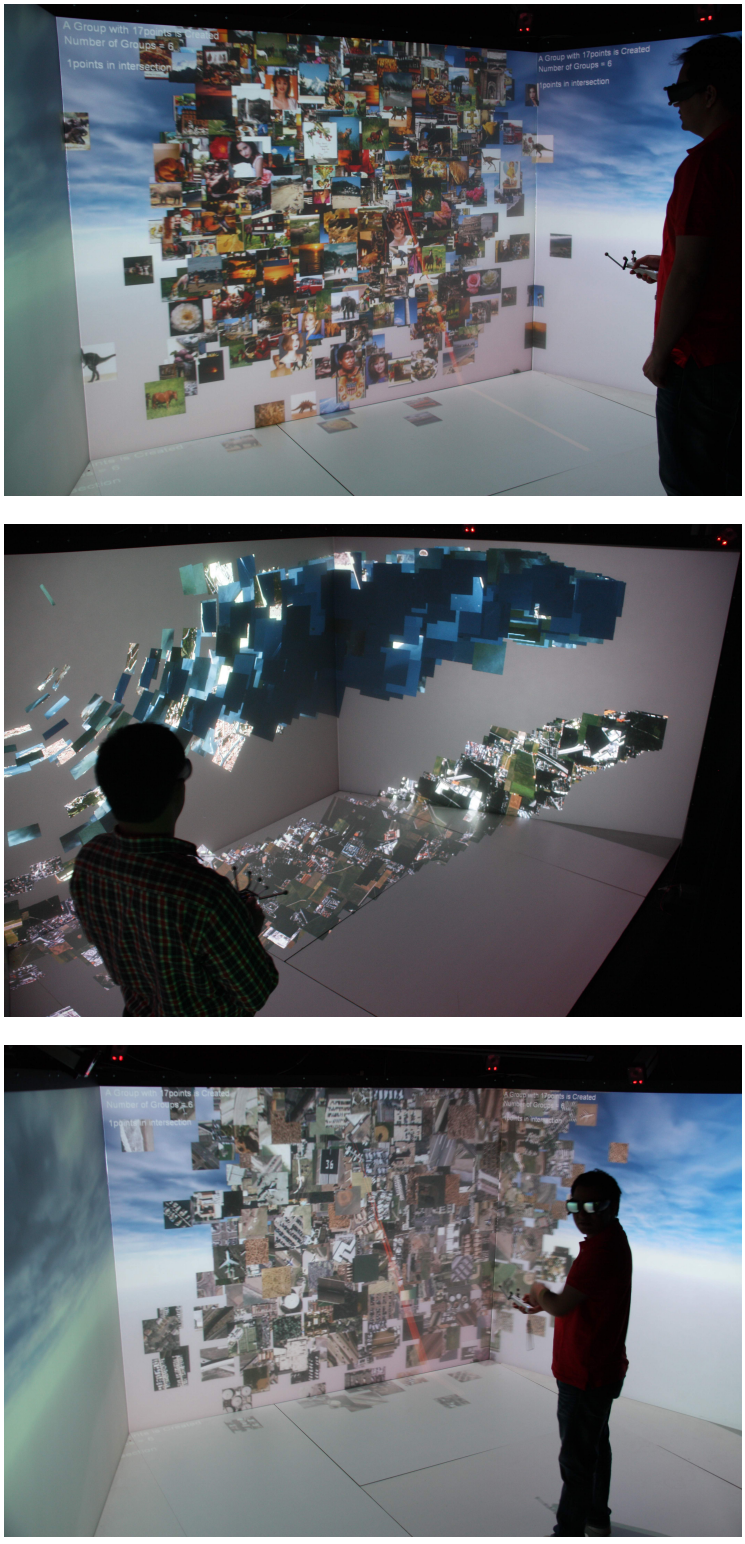
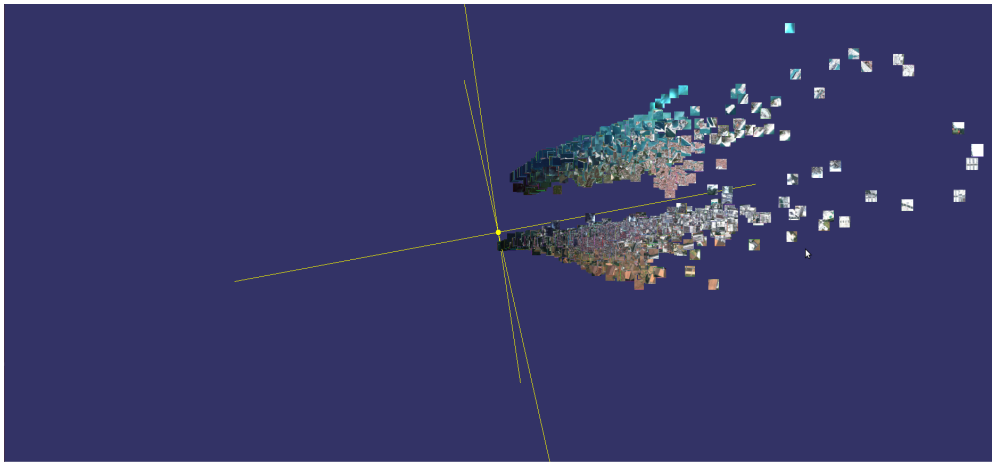
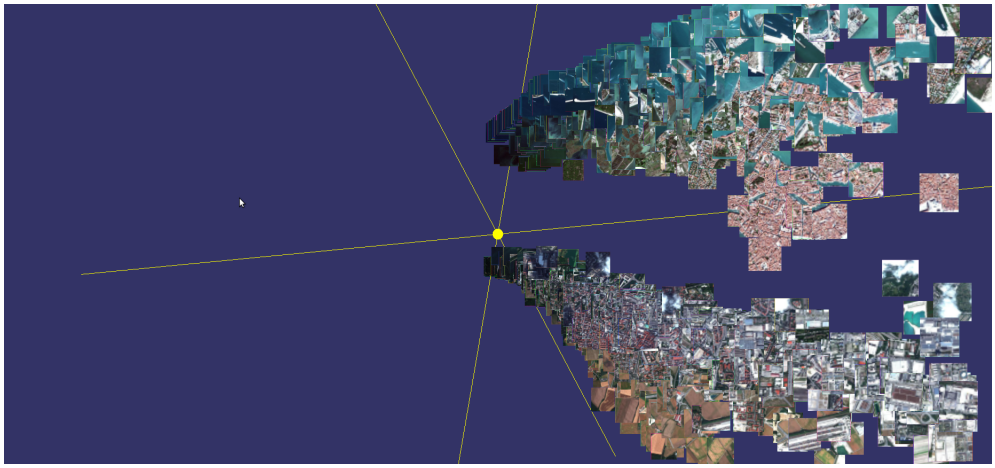


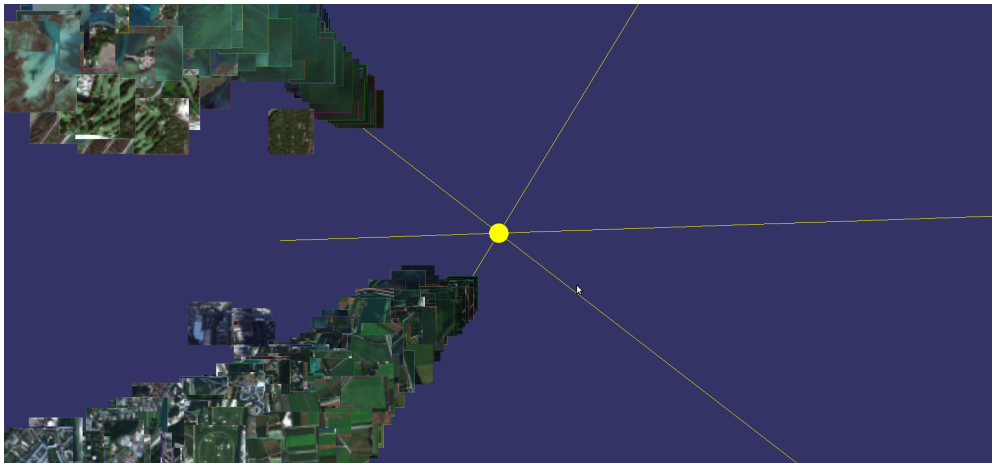
Figure 6.5: Visualization of image collections in the CAVE. The images have been provided by Dr.-Ing Mohammadreza Babaee, Human-Machine Communication Institute (MMK), Technical University of Munich, Munich, Germany.



(a)



(b)



(c)

Figure 6.6: Visualization of multispectral image patches in a three-dimensional environment based on their extracted color features using the rgbHist feature descriptor.

look very similar as they contain the same content class such as water surfaces, they are separated in the feature space. This can be a result of the difference between human object discrimination and that of a computer. While a human observer may consider both clusters as containing the same object class, a computer using Gabor feature descriptors discriminated them. These image patches may contain a feature which can be only recognized by the computer. This kind of knowledge obtained from the feature space exploration can explain an unsatisfactory result given by an image mining system. Being aware of the feature space structure, users can either replace the feature descriptor or develop a new feature extraction method to fulfill their needs, for example, to unify the two clusters.

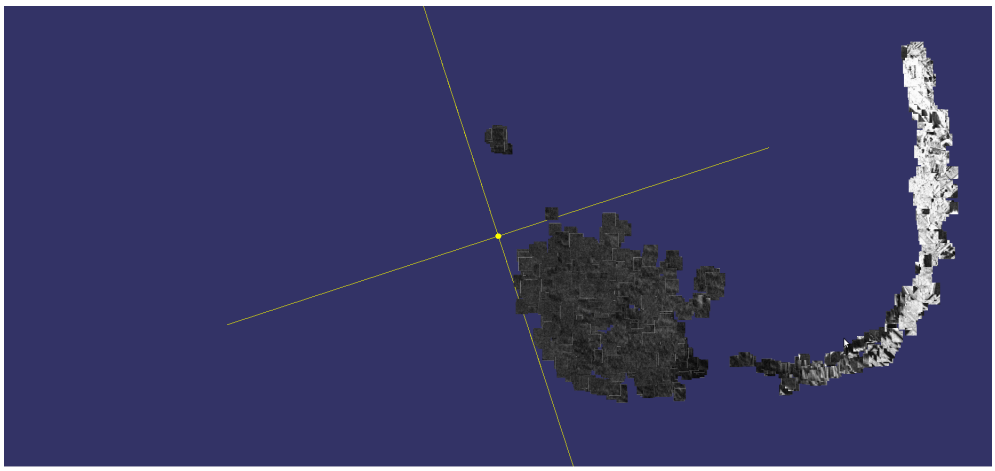
Figure 6.8 shows a visualization of another example of SAR image patches. As illustrated in the figure, the feature space generated from the extracted Gabor features of the image patches consist of one elongated cluster. In this cluster the homogeneity of the patches varies from one end to the other one (containing highly structured land uses such as industrial areas to the homogeneous image patches containing agricultural fields or water surfaces). Being aware of the structure of the feature space one can, for example, explain the poor performance of an image mining system, which considers a Gaussian distribution for each object class, when categorizing various land uses is required.

As discussed above, visualization of an image collection helps users not only to understand the image semantics, but also to figure out the right set up of an image mining system (e.g., selecting an appropriate feature descriptor) to fulfill their expectations. In addition, users can further utilize this knowledge for annotating an image collection considering various semantics, even from different levels. Considering the example of Figure 6.7, at a higher semantic level a user can split the data into three categories, while at a lower semantic level, the elongated cluster can be further split into various land uses.

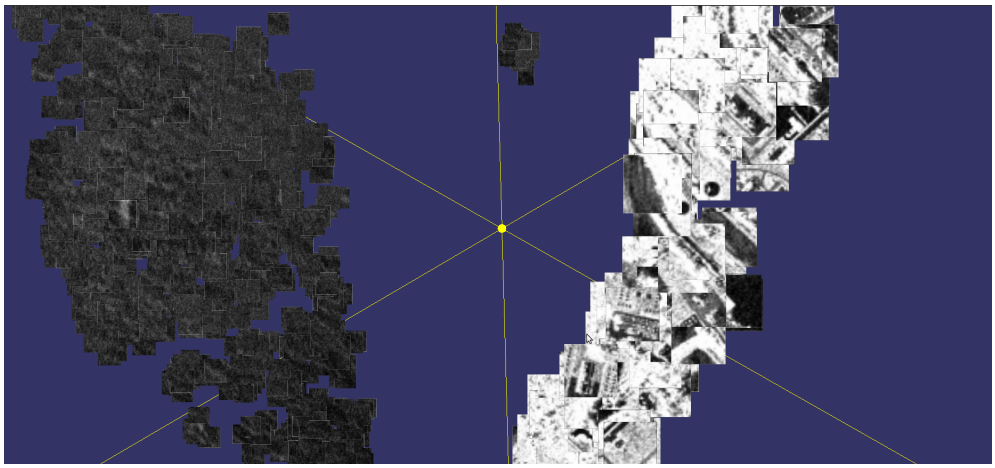
6.2.4 Summary

In this section, we discussed the importance of visualization and exploration of image patches based on their extracted features, to understand their semantics. For exploring the image feature space we used a recently developed *Visual Data Mining (VDM)* system, the so-called *Learning by Immersive Visual Data Mining* system. This system utilize the three-dimensional environment of the *Cave Automated Virtual Environment (CAVE)* to visualize images. Users are then allowed to navigate, explore, and interact with the images.

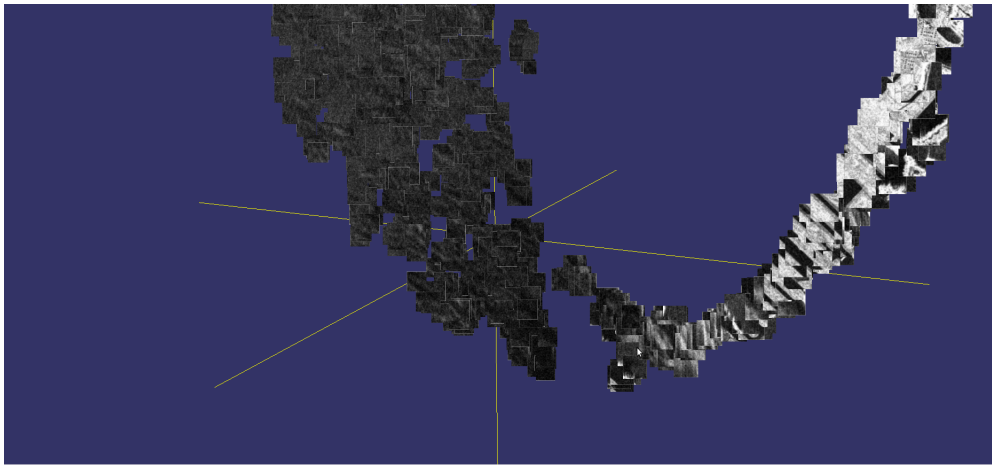
For our discussion, we visualize various EO images such as multispectral and SAR images. The knowledge gained through exploring the image feature space can help to shorten the semantic gap, defined as the gap between users' and computers' image understanding. By visualizing the images based on their extracted features users can see the images from an image mining system perspective. Thus, they can



(a)

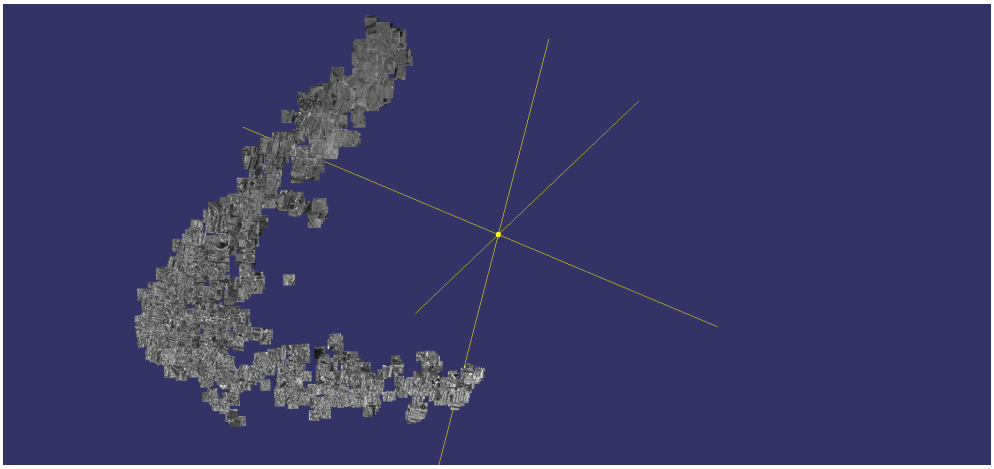


(b)

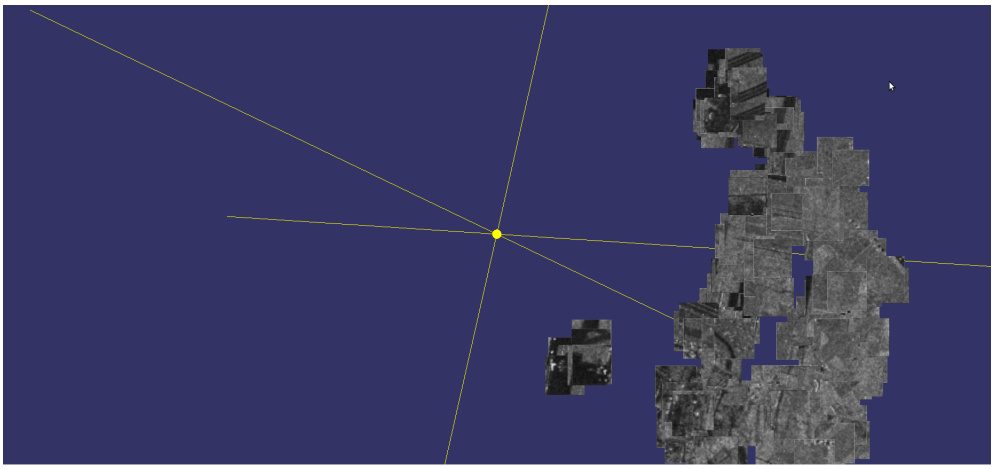


(c)

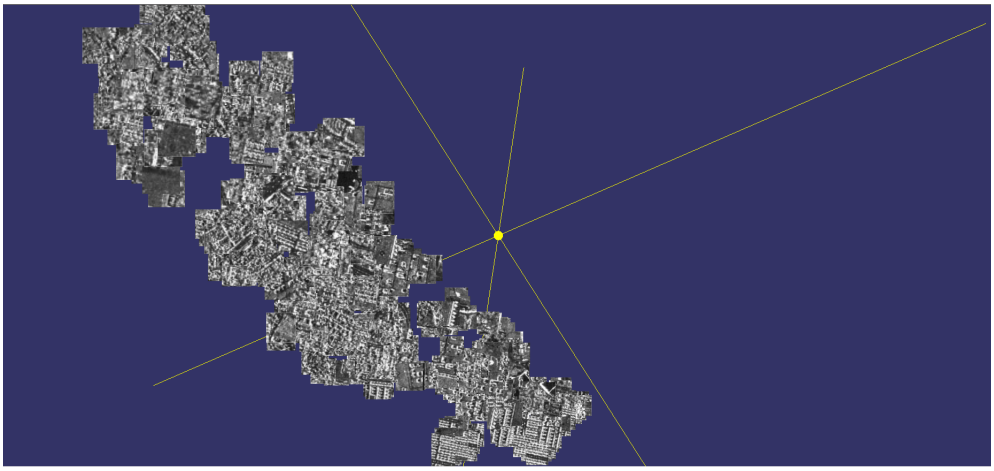
Figure 6.7: Visualization of SAR image patches in a three-dimensional environment based on their extracted texture features using the Gabor feature descriptor.



(a)



(b)



(c)

Figure 6.8: Visualization of SAR image patches in a three-dimensional environment based on their extracted texture features using the Gabor feature descriptor.

either adjust the system parameters (e.g., the feature extraction method) to provide a result closer to their expectations, or modify their object annotations according to their findings of the image semantics from the feature space. In addition, new feature extraction methods can be designed and developed to provide user-expected semantics in the feature space, which can be utilized later by image mining systems as the semantic interpretations of the images.

7

Summary, Conclusions, and Future Work

In this chapter, we summarize this dissertation and conclude the discussed points, followed by potential future work.

7.1 Summary

In this dissertation, we dealt with the semantics of *Earth Observation (EO)* images based on the description of their extracted features.

First, we proposed an efficient feature coding strategy, the so-called *Locally Linear Salient Coding (LLSaC)* method, to provide a compact but descriptive representation of image features on the *Bag-of-Words (BoW)* level. The compactness of the coded feature descriptors increases the scalability of the applied learning systems. LLSaC combines two important characteristics of a feature space, namely saliency and interrelationships between feature points. Experimental results indicated that LLSaC significantly outperforms other coding strategies in describing the feature space structure with more compact coded feature descriptors for both multimedia and EO image patch datasets.

Next, we introduced a new approach, *Bag-of-Topics (BoT)*, to model EO image patches according to semantically meaningful features. To generate image BoT models, we applied *Latent Dirichlet Allocation (LDA)*, a topic model, to the BoW models of images and discovered their latent semantics as a set of topics. Each image is then represented by a mixture of the discovered topics (BoT). Experimental results demonstrated that the BoT model can provide results comparable to those of the BoW model; however, the description of data is much more compact in the BoT model. Consequently, BoT not only increases the scalability of image mining systems, but also discriminates various image classes to a higher degree. Furthermore, since the BoT model builds upon the BoW model, the improvements

in the efficiency of the BoW model such as using a LLSaC strategy scales up the discriminablity of the topics in the BoT model, too.

Additionally, we proposed a communication channel-based approach for measuring the information quantity that various feature descriptors extract from a given image collection and deliver to image mining systems, regardless of user labeling. In this approach, we model LDA as a communication channel, where images are the input, topics are the output, and the feature descriptors are the carriers in this channel. The transmitted information quantity is then measured by computing the channel's mutual information. Experimental results demonstrated that feature descriptors which carry a larger amount of mutual information provide a higher discrimination capability when they are applied to other image mining tasks such as a classification using an SVM approach.

Following this, we used information theory to propose a novel technique based on *Huffman Coding (HC)*, a lossless compression technique, to measure the overlaps between the information obtained by different feature descriptors. The information overlap is used as a measure of similarity between any two feature descriptors in representing an image dataset. Experimental results showed that the computed information overlap can predict the degree of similarities between the performances of an image mining system using various feature descriptors. Additionally, considering information overlap in feature fusion tasks allows for the provision of a broader range of new and diverse information by a combination of less feature descriptors, which improves both the scalability and distinguishability of image mining systems.

Next, we conducted user studies for assessing the issues which limit users' understanding of the EO image semantics, namely the *sensory* and *semantic* gaps. In EO images, the sensory gap is rather wide due to sensor resolution, image perspective, scale and FOV (patch size). For our study, we assessed the sensory gap by evaluating human perception and by employing a computational approach. For the human perceptual evaluation, user assigned labels describing image patch content were gathered and analyzed. The results highlighted issues caused by the sensory gap such as the bird's eye view perspective of the EO images which humans are not accustomed to, and therefore affects their object recognition. Additionally, resolution and scale present difficulties for object recognition. Users can disambiguate objects within an image patch by gathering context from the object surroundings, which is limited by the FOV of the image patch (i.e., the image patch size). Therefore, a limited FOV makes issues such as resolution more serious. The effect of FOV on the sensory gap was also assessed via a computational evaluation where the sensory gap is defined as the difference between the scene context discovered by LDA from content within a certain FOV, and the reference context. The results indicated that increasing the FOV decreases the sensory gap.

We then further studied the semantic gap and its influences on the results of image mining systems. The existing methods for bridging the semantic gap usually consider the gap as the differences between user and computer interpretations of an

image. These methods are verified either by comparing results to a reference dataset, or by measuring the degree of user acceptance in interactive systems. Although these methods result in a narrower semantic gap between computers and users, the resulting model for a specific user and search goal may still not be satisfactory to other users. We showed that the subjective biases present in the bridging methods, which we refer to as the *linguistic semantic gap*, causes this discrepancy. In addition, we showed that the sensory gap is one of the causes of the semantic gap. In order to overcome this problem, we suggest to consider the linguistic semantic gap in designing methods to bridge the semantic gap. Additionally, the diversity of data sets being used in the domain (e.g., using various EO datasets in EO tasks) should be increased, which will include different user perspectives and compensate the individual subjective biases. Moreover, models derived from proposed methods for bridging the semantic gap could be stored and further used by other systems, which would then be including various users' image interpretations.

Next, we developed a clustering-based approach to evaluate EO image semantics, in which we apply clustering to the extracted features of image patches for a different number of clusters. The resulting clusterings are then evaluated both internally and externally (i.e., without and with using prior image labels, respectively). While an internal evaluation shows the feature space structure by finding the optimum clusters, an external evaluation allows for the comparison of the obtained clusters to the user labeling of the image patches. Experimental results indicated that multiple optimum clusterings can exist for a given image patch collection depending on different image property levels (e.g., from low level texture to higher level contexts). Furthermore, an external evaluation showed that the structures of the feature space discovered by the clusterings do not necessarily correspond to user labeling (i.e., user image understanding). Moreover, comparing the clusterings across different feature descriptors demonstrates that users consider some image properties more intensively than others.

Finally, we showed the importance of image visualization and exploration, based on their extracted features, in understanding of their semantics. We visualized example EO image patches inside a virtual environment, the so-called *Cave Automated Virtual Environment (CAVE)*. Experimental results demonstrated that the knowledge gained through exploring the image feature space can lead to reduce the semantic gap. The visualization of images based on their extracted features allows users to see the images from the point of view of an image mining system. Thus, the users can either adjust the system parameters to bring the results closer to their expectations, or modify their opinions about the image semantics according to their new findings in the feature space.

7.2 Conclusions

As a main contribution, we showed that higher level feature descriptors represent images by their semantics. These descriptors are usually more compact and improves the accuracy and scalability of image mining systems. Furthermore, the image mining results are usually verified through a comparison to a user generated reference dataset, or user acceptance in an active learning scenario, where in both cases user image understanding plays a main role. Therefore, efforts for developing new feature descriptors should consider issues which cause discrepancies between user image understanding and the computer's interpretation of the images, such as the sensory and semantic gaps.

According to our studies, the computer sensory gap is affected by the feature descriptors being used. Object discrimination, which is mostly affected by the sensory gap, is a basic step for object identification; therefore, the sensory gap has an effect on the semantic gap. The semantic gap is a measure of the relevance of the information provided by the computer for the user. Therefore, studying the sensory gap provides a way to identify feature descriptors which present the most relevant information for the specific task. In addition, while image product properties are fixed in the *Earth Observation (EO)* domain, different tasks require different image properties. Thus, studying the sensory gap helps to find appropriate image properties for annotation and learning tasks (e.g., optimal image patch size) and to find the best combination of data sources from different sensors (SAR, multi-spectral), with different properties (e.g., resolution).

Moreover, analyzing the feature space generated by the description of the extracted features from images can help to reduce the semantic gap, defined as the gap between a user's image understanding and a computer's image interpretation. In this dissertation, we introduced two methods for assessing the image feature space: 1) employing a computational method, 2) visualizing the feature space and explore it. The obtained knowledge can then be used to develop feature descriptors (either primitive level or higher level descriptors) to restructure the feature space according to the users' image understanding. Furthermore, since user object perception and recognition is different than that of a computer's, analysis of the feature space can introduce new semantics to the users (which the human visual system is not able to perceive), which can help the users to leverage their image understanding.

In spite of efforts for reducing the semantic gap between users and computers, the resulting model for a specific user and search goal may still not be satisfactory to other users. We further showed that the main reason behind this discrepancy is the existing subjective biases in the bridging methods. These biases are caused by image understanding across users due to their diverse background knowledge, and the goals of each specific image mining task, which we refer to as the *linguistic semantic gap*. In order to overcome this problem, we suggest to consider the linguistic semantic gap in developing feature descriptors and designing methods to bridge the semantic

gap. Moreover, the diversity of datasets being used in the domain (e.g., using various EO datasets in EO tasks) should be increased, which will include different user perspectives and compensate for the individual subjective biases. In addition, models derived from proposed methods for bridging the semantic gap could be stored and further used by other systems, which would then be including various users' image interpretations.

7.3 Future Work

For a future work in feature coding, the locally linear reconstruction technique can be seen as an independent wrapper which can be applied to other codebook-based coding strategies (e.g., LLC, SV, the variants of SaC). This allows them to use the local information of the feature points in order to discover the global structure of the feature space with fewer codewords.

This dissertation demonstrates the superior performance of the *Bag-of-Topics (BoT)* in classification tasks. However, the selection of an optimized number of topics for the LDA model still deserves more detailed investigations.

Research on the interaction between the causes of the sensory gap such as image patch size should be extended since considering the sensory gap is necessary for the semantic gap assessment. In addition, the relationships between the sensory and semantic gaps should be further studied. Additionally, various subjective biases in the existing image mining systems and EO datasets should be studied. Moreover, more user studies should be conducted in order to determine various aspects of the linguistic semantic gap.

7.4 Related Publications

- R. Bahmanyar, A. Murillo Montes de Oca, and M. Datcu, "The semantic gap: An exploration of user and computer perspectives in Earth Observation images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 2046–2050, October 2015.

- R. Bahmanyar and A. Murillo Montes de Oca, "Evaluating the sensory gap for Earth Observation images using human perception and an LDA-based computational model," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 566–570, September 2015.

- R. Bahmanyar, S. Cui, and M. Datcu, "A comparative study of Bag-of-Words and Bag-of-Topics models of EO image patches," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 1357–1361, June 2015.

- R. Bahmanyar, M. Datcu, and G. Rigoll, "Comparing the information extracted by feature descriptors from EO images using Huffman coding," in *Proc. International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, June 2014.

- M. Babaee, G. Rigoll, R. Bahmanyar, and M. Datcu, "Locally linear salient coding for image classification," in *Proc. International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–4, June 2014.

- M. Babaee, R. Bahmanyar, G. Rigoll, and M. Datcu, "Farness preserving non-negative matrix factorization," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 3023–3027, October 2014.

- M. Babaee, R. Bahmanyar, G. Rigoll, and M. Datcu, "Interactive clustering for SAR image understanding," in *Proc. European Conference on Synthetic Aperture Radar (EUSAR)*, pp. 1–4, June 2014.

- R. Bahmanyar, G. Rigoll, and M. Datcu, "A clustering-based approach for evaluation of EO image indexing," in *Proc. ISPRS Sensors and Models in Photogrammetry and Remote sensing (SMPR)*, pp. 79–84, October 2013.

- R. Bahmanyar and M. Datcu, "Measuring the semantic gap based on a communication channel model," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 4377–4381, September 2013.

- M. Babaee, R. Bahmanyar, G. Rigoll, and M. Datcu, "Immersive visual information mining for exploring the content of EO archives," in *Proc. ESA Living Planet Symposium*, (Edinburgh), September 2013.

A

Primitive Feature Extraction Methods

In this appendix, we briefly explain the primitive feature extraction methods which have been used in our experiments.

A.1 RGB Color Histogram

An *RGB Color Histogram (rgbHist)* feature descriptor is created by concatenating the histograms of the pixel values for the RGB color channels where each histogram has 256 bins, resulting in a histogram of 768 bins. This histogram is then considered as a vector in Euclidean space, the so-called rgbHist feature vector. In this dissertation, rgbHist feature vectors are computed locally using a square sliding window with a given dimension.

A.2 Random Features

In order to exploit the pixel brightness values of images, usually histograms of pixel values are constructed (e.g., color histograms in multimedia color images). Since the brightness range of SAR images is rather wide, which results in a very large vector, constructing a pixel value histogram is not trivial. Thus, we put all the pixel values of an image $I_{i \times j}$ in a vector of size $m = i \times j$. In this way, the resulting vector is much smaller than the brightness histogram; however, it is still too large to be used efficiently in a clustering algorithm. Therefore, we use *Random Projection (RP)* [27] which decreases the dimensionality of the resulting feature vector m to a lower dimensional vector of size $\tilde{m}$. The resulting feature vector is called *Random Features (Rand_Feat)*. In this method, the product of the high dimensional feature vector and a $m \times \tilde{m}$ random matrix is computed. This method has been recently successfully applied to SAR image segmentation by Hou *et al.* [28].

A.3 Mean and Variance

The simplest use of statistics for extracting image local patterns is computing statistical *mean* and *variance* of local pixel neighborhoods. In our experiments, for each pixel, we employ a sliding window of 3×3 pixels and compute the mean and variance of the pixels within the window, resulting in two-dimensional feature vectors.

A.4 Scale-Invariant Feature Transform

The *Scale-Invariant Feature Transform (SIFT)*, proposed by Lowe [70], is an image interest point detector and local feature descriptor. SIFT applies a *Difference of Gaussians (DoG)* algorithm to a series of smoothed images in scale space and detects maxima and minima of the resulting image as the interest points of the image. The interest points are usually located at high contrast edges and corners and are invariant to image scale and rotation. The interest point selection of SIFT is usually used for image registration and matching.

In our experiments, we extract the SIFT feature descriptors in a dense way, which means we use a regular grid as opposed to only the interest points. In order to describe local features of an image, SIFT generates 16 orientation histograms on neighborhoods of 4×4 pixels, where each histogram consists of 8 bins corresponding to 8 different directions. Concatenation of these histogram results in feature vectors with 128 elements.

A.5 Weber Local Descriptor

The *Weber Local Descriptor (WLD)* is a feature descriptor developed based on Weber's law, a psychological law [3]. According to this law, a human notices the change in a stimulus as a valid signal if its ratio to the original brightness of the stimulus is above a certain constant value. WLD is constructed by a two-dimensional histogram of: 1) *Differential Excitation*, the brightness difference ratio between each pixel x and its neighbors; 2) *Orientation*, which is the gradient orientation of each pixel x. The final feature vector is constructed by building a one-dimensional histogram of the computed two-dimensional histogram, after quantizing it to M excitations and T orientations.
In our experiments, we set M and T equal to 6 and 8, respectively, which results in a feature vector of 144 elements.

A.6 Gabor Feature Descriptors

The *Gabor* feature descriptors, proposed for texture analysis, are achieved by filtering a given image patch using a set of linear band-pass filters, the so-called Gabor filters [167]. These filters are generated by scaling and rotating a mother wavelet filter. The impulse response of this filter is a two-dimensional modulated Gaussian function. The final Gabor feature vector is constructed by using the means (μ_{sr}) and the standard deviations (σ_{sr}) of the image patch filtered by S scales and R rotations, $F_{Gabor} = [\mu_{11}\ \sigma_{11}\ \mu_{12}\ \sigma_{12}\ ...\ \mu_{SR}\ \sigma_{SR}]$.
In our experiments, the Gabor features are constructed for 3 scales and 6 rotations which leads to a vector of 36 dimensions.

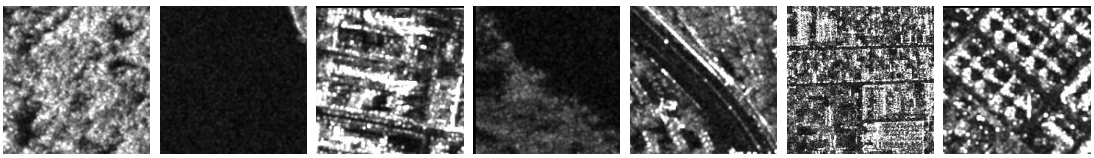
B

Datasets

In this appendix, we introduce the image patch datasets which have been used for our experiments.

B.1 Seven Class TerraSAR-X Image Patches

This dataset is a collection of 1230 TerraSAR-X image patches each comprising 160×160 pixels[1]. The patches are cut out from multi look ground range detected and radiometrically enhanced TerraSAR-X image products; their ground sample distance is 1.2 m. The image patches are grouped into seven non-equal size classes, namely forest (198 images), water (210 images), medium density urban area (204 images), forest & water (114 images), roads (67 images), high density urban area (279 images), and urban area & roads (158 images). The images within the classes are rather homogeneous which allows us to study the difference between the annotation and the resulting clusters. This dataset is not publicly available. Figure B.1 illustrates some sample image patches of the dataset.



(a) (b) (c) (d) (e) (f) (g)

Figure B.1: Seven Class TerraSAR-X image patch dataset. (a) Forest, (b) Water bodies, (c) Medium density urban, (d) Forest & Water bodies, (e) Roads, (f) High density urban, (g) Urban & Roads.

[1]The images have been collected by Shiyong Cui, Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Germany, shiyong.cui@dlr.de.

B.2 Fifteen Class TerraSAR-X Image Patches

This dataset contains 3434 TerraSAR-X image patches of 160×160 pixels manually grouped into 15 non-equal size classes[1]. The number of images in the classes are between 118 and 420 images. Six classes are created from urban areas, four classes from different kinds of agricultural fields, and the rest are from forests, mountains, industrial areas, highways, and water bodies. The patches are derived from multi look ground range detected and radiometrically enhanced TerraSAR-X image products; their ground sample distance is 1.2 m. Since this dataset is not publicly available, four representative samples of each class are shown in Figure B.2.

B.3 UC Merced Land Use

This dataset is a manually labeled image collection gathering 21 classes of land use scenes [5]. Each class contains 100 image patches of 256×256 pixels from aerial orthography with a ground resolution of about 0.3 m. In this dataset, the classes are selected such that they contain a rich variation of spatial patterns. Thus, there are classes being homogeneous in color, classes homogeneous in texture, classes homogeneous in shape, and classes containing images which have no shared features. Figure B.3 shows samples of the UC Merced Land Use dataset.

B.4 Fifteen Natural Scenes

The Fifteen Natural Scenes dataset is a collection of 4485 gray value images of outdoor and indoor scenes[2]. The images are grouped into 15 non-equal size categories, where each contains between 200 and 400 images. Figure B.4 shows sample images of this dataset.

[2] http://www-cvr.ai.uiuc.edu/ponce_grp/data/

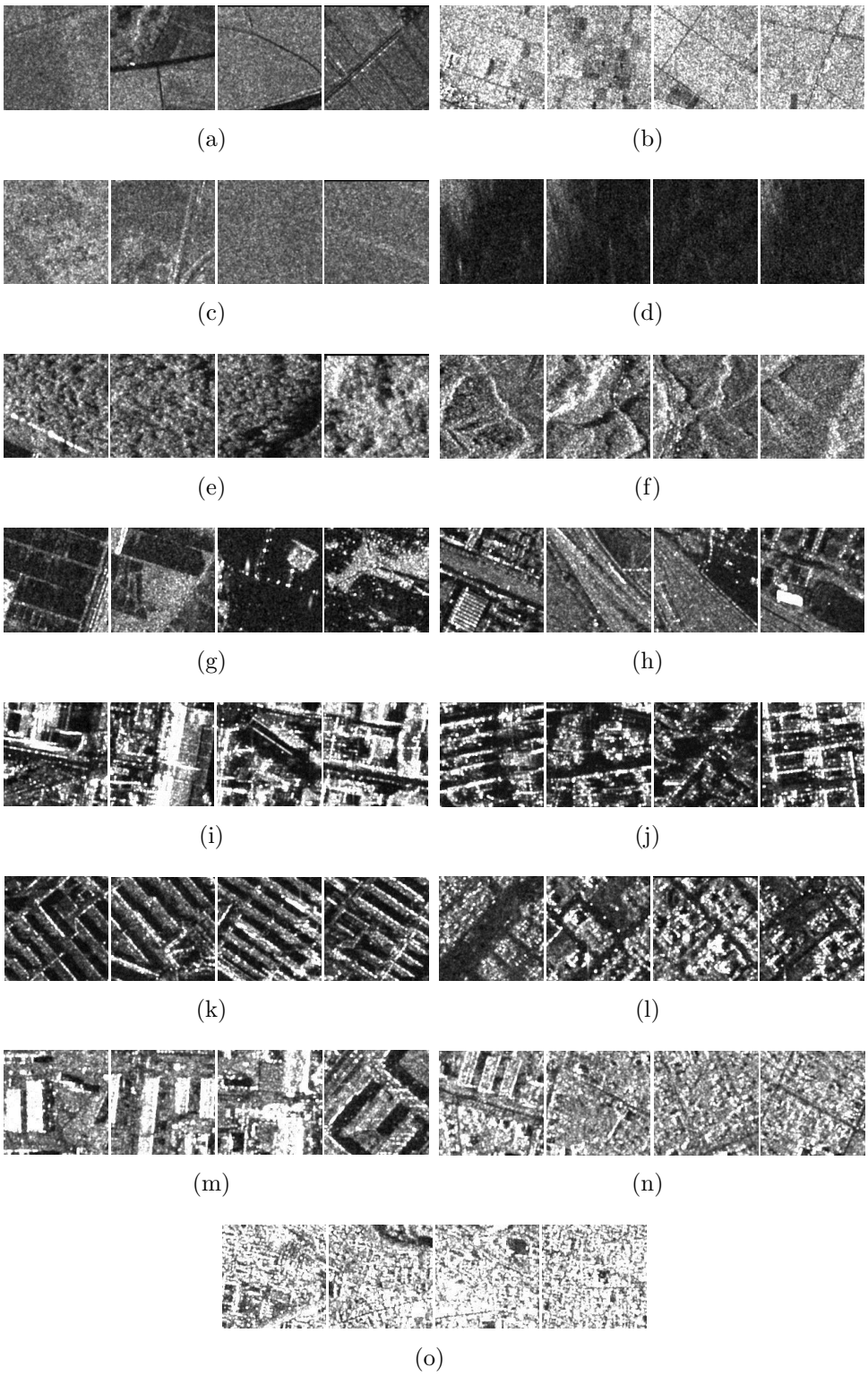


Figure B.2: Four representative samples from each of the fifteen class TerraSAR-X image patches. (a) and (b) Agricultural fields, (c) Grass fields, (d) Water bodies, (e) Forests, (f) Mountains, (g) Flooded fields, (h) Highways, (i) Industrial areas, (j) - (o) Different kinds of urban areas.

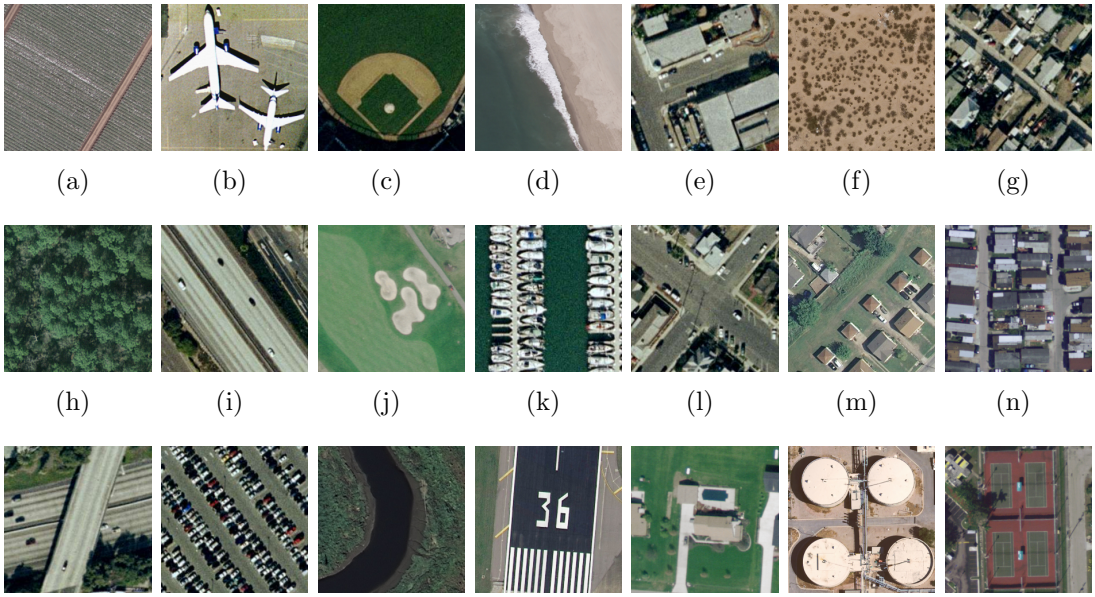


Figure B.3: UC Merced Land Use dataset. A sample from each of the 21 groups is shown in this figure. (a) Agricultural, (b) Airplane, (c) Baseball diamond, (d) Beach, (e) Buildings, (f) Chaparral, (g) Dense residential, (h) Forest, (i) Freeway, (j) Golf course, (k) Harbor, (l) Intersection, (m) Medium density residential, (n) Mobile home park, (o) Overpass, (p) Parking lots, (q) River, (r) Runway, (s) Sparse residential, (t) Storage tanks, (u) Tennis court.

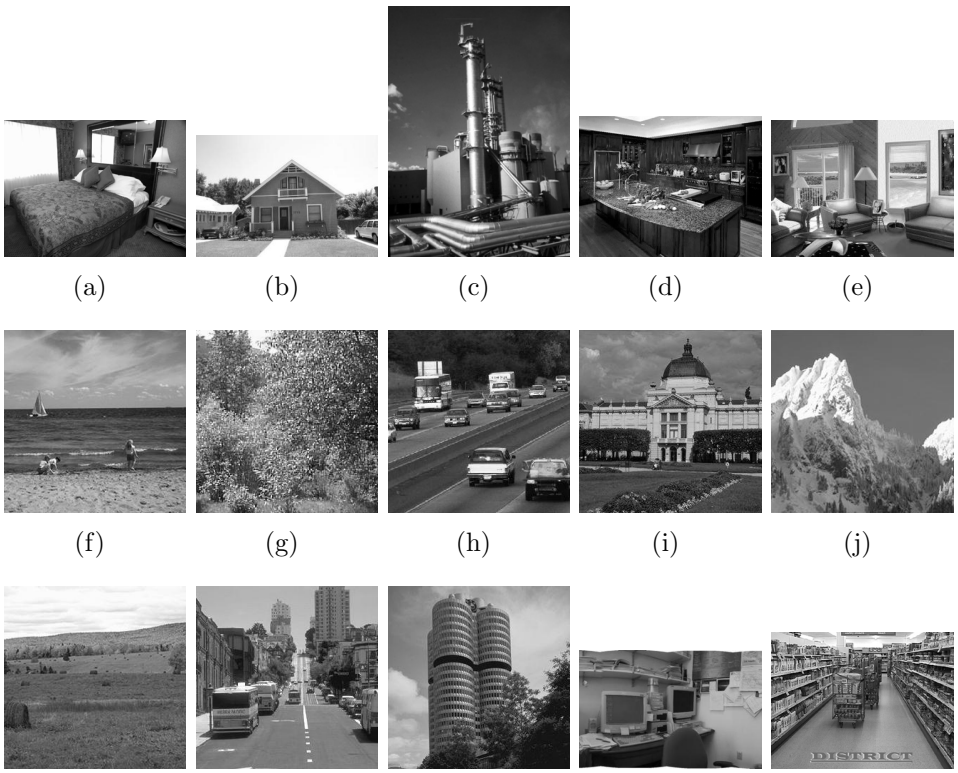


Figure B.4: Samples from the Fifteen Natural Scenes dataset. (a) Bedroom (b) Suburb, (c) Industrial, (d) Kitchen, (e) Living room, (f) Coast, (g) Forest, (h) Highways, (i) Inside city, (j) Mountain, (k) Open country, (l) Street, (m) Tall building, (n) Office, (o) Store.

Acronyms

AB-SIFT ..... Adaptive Binning Scale-Invariant Feature Transform
AMI ..... Adjusted Mutual Information
ARI ..... Adjusted Random Indexing
AWLD ..... Adapted Weber Local Descriptor
BoT ..... Bag-of-Topics
BoW ..... Bag-of-Words
CASI ..... Digital Compact Airborne Spectrographic Imager
CAVE ..... Cave Automated Virtual Environment
CH ..... Calinski-Harabasz
DB ..... Davies-Bouldin
Dens ..... Average Between-cluster Density
DoG ..... Difference of Gaussians
EM ..... Expectation Maximization
EO ..... Earth Observation
FCD ..... Fast Compression Distance
FN ..... False Negative
FOV ..... Filed of View
FP ..... False Positive
FrFT ..... Fractional Fourier Transform
GG ..... Grouped, Grouped
GLCM ..... Gray Level Co-occurrence Matrix

GMRF ..... Gaussian Markov Random Field
GS ..... Grouped, Separated
GSC ..... Group Salient Coding
HC ..... Huffman Coding
HV ..... Hard Voting
IEEE ..... Institute of Electrical and Electronics Engineers
LBP ..... Local Binary Patterns
LCC ..... Local Coordinate Coding
LDA ..... Latent Dirichlet Allocation
LIBSVM ..... Library for Support Vector Machines
LiDAR ..... Light Detection And Ranging
LLC ..... Local-constraint Linear Coding
LLE ..... Locally Linear Embedding
LLSaC ..... Locally Linear Salient Coding
LZW ..... Lempel-Ziv-Welch
MAP ..... Maximum A Posterior
MFT ..... Matched Fourier Transform
ML ..... Machine Learning
MLPH ..... Multilevel Local Pattern Histogram
MMK ..... Institute for Human-Machine Communication
MV ..... Mean-Variance
MVR ..... Mean-Variance-Ratio
NMF ..... Nonnegative Matrix Factorization
PCA ..... Principal Component Analysis
pLSA ..... probabilistic Latent Semantic Analysis
PPV ..... Precision Measure
QE ..... Query by Example
QMF ..... Quadrature Mirror Filter
Rand_Feat ..... Random Features
REF ..... Content Reference Annotation

RF . . . . . Relevance Feedback
RGB . . . . . Red-Green-Blue Color Channels
rgbHist . . . . . RGB Color Histogram
RI . . . . . Random Indexing
RMD . . . . . Ratio of Mean Difference
RP . . . . . Random Projection
SaC . . . . . Salient Coding
SAR . . . . . Synthetic Aperture Radar
SAR-SIFT . . . . . Synthetic Aperture Radar Scale-Invariant Feature Transform
SC . . . . . Sparse Coding
Scat . . . . . Average Scattering on the Clusters
SIFT . . . . . Scale-Invariant Feature Transform
SIFT-Color . . . . . Scale-Invariant Feature Transform for RGB Color Channels
STFT . . . . . Short Time Fourier Transform
SS . . . . . Separated, Separated
SURF . . . . . Speeded Up Robust Features
SV . . . . . Soft Voting
SVM . . . . . Support Vector Machine
TP . . . . . True Positive
TPR . . . . . Sensitivity Measure
UC . . . . . University of California
UT . . . . . Universal Time
UX1 . . . . . User Experiment Group 1
UX2 . . . . . User Experiment Group 2
VDM . . . . . Visual Data Mining
VR . . . . . Virtual Reality
WLD . . . . . Weber Local Descriptor
WLD-Color . . . . Weber Local Descriptor for RGB Color Channel

List of Symbols

b ..... Interception term in an SVM method

c_i ..... Center of gravity of cluster s_i

C ..... Regularization parameter of an SVM method

D ..... A text or image corpus

$\mathbb{D}$ ..... Set of training data in a supervised machine learning algorithm

f ..... Decision function in an SVM method

k ..... Number of clusters in a clustering task

K ..... Number of topics in a topic model

$\hat{K}$ ..... Number of nearest codewords to x_i in feature space

$\bar{K}$ ..... Number of nearest neighbors to x_i in feature space

m ..... Dimensionality of a point or a vector

M ..... Number of documents or images in corpus D

N ..... Number of points in point set X

N_d ..... Number of textual or visual words in $\mathbf{w}_d$

N_V ..... Number of textual or visual words in dictionary V

$N(x_i)$ ..... Set of the nearest codewords to x_i in feature space

$NP(x_i)$ ..... Set of the nearest neighbors to x_i in feature space

p_{ij} ..... Middle point on the line connecting the two cluster centers c_i and c_j

$\mathbb{R}^m$ ..... A m-dimensional real coordinate space

S ..... Set of clusters

s_i ..... A cluster in a cluster set S

T ..... Set of discovered topics by LDA

t_j ..... A discovered topic by LDA

U ..... Set of linear coefficients

u_{il} ..... Linear coefficient corresponding to the pints x_i and x_l

V ..... Dictionary of textual or visual words

v_i ..... A textual or visual words in dictionary V

$\mathbf{w}_d$ ..... A text document or BoW model of an image

w_{dn} ..... A word-token in $\mathbf{w}_d$

X ..... Set of m-dimensional points

x_i ..... A m-dimensional point in a point set X

y ..... Label vector assigned to point set X through a classification task

$\mathbf{z}_d$ ..... Set of topic-tokens generating $\mathbf{w}_d$

z_{dn} ..... The topic-token corresponding to the word-token w_{dn}

λ ..... Lagrange multiplier

$\phi(.)$ ..... A Kernel approximation function

$\overline{\phi}(.)$ ..... A Kernel approximation function

$\Phi(.)$ ..... A monotonically decreasing function

γ ..... A Kernel parameter

$\Gamma(.)$ ..... A Gamma function

ξ_i ..... Slack variable corresponding to point x_i in an SVM method

θ ..... Latent topic mixing weight in an LDA model

α ..... Dirichlet parameter

β ..... An $N_V \times K$ matrix parameterizing the word probabilities within the topics in an LDA model

ω ..... Weight vector in an SVM method

η_{ij} ..... Salient response of point x_i to codeword v_j

σ ..... Variance

Bibliography

[1] A. R. Brenner, L. Roessing, and P. Berens, "Potential of very high resolution SAR interferometry for urban building analysis," in *Proc. European Conference on Synthetic Aperture Radar (EUSAR)*, pp. 1–4, June 2010.

[2] R. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *IEEE Transactions on Systems, Man and Cybernetics*, vol. SMC-3, pp. 610–621, November 1973.

[3] J. Chen, S. Shan, C. He, G. Zhao, M. Pietikainen, X. Chen, and W. Gao, "WLD: A robust local image descriptor," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1705–1720, September 2010.

[4] M. Walessa and M. Datcu, "Model-based despeckling and information extraction from SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, pp. 2258–2269, September 2000.

[5] Y. Yang and S. Newsam, "Bag-of-Visual-Words and spatial extensions for land-use classification," in *Proc. ACM International Conference on Advances in Geographic Information Systems (GIS)*, pp. 270–279, November 2010.

[6] S. Cui, *Spatial and Temporal SAR Image Information Mining.* PhD thesis, University of Siegen, 2014.

[7] R. D. Zilca and Y. Bistritz, "Feature concatenation for speaker identification," in *Proc. European Signal Processing Conference*, pp. 1–4, September 2000.

[8] F. S. Tsai and K. L. Chan, "Dimensionality reduction techniques for data exploration," in *Proc. International Conference on Information, Communications Signal Processing*, pp. 1–5, December 2007.

[9] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, December 2000.

[10] "SAR-EDU remote sensing education initiative." https://saredu.dlr.de/.

[11] A. Murillo Montes de Oca, N. Nistor, and M. Datcu, "Creating a reference data set for satellite image content based retrieval," in *Proc. Conference on Big Data from Space (BiDS)*, pp. 71–75, November 2014.

[12] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE Journal of Selected Topics in Applied Earth Observation and Remote Sensing*, vol. 4, pp. 193–204, March 2011.

[13] J. Tang, Z.-J. Zha, D. Tao, and T.-S. Chua, "Semantic-gap-oriented active learning for multilabel image annotation," *IEEE Transactions on Image Processing*, vol. 21, pp. 2354–2360, April 2012.

[14] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. ACM International Conference on Multimedia*, pp. 157–166, November 2014.

[15] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Communications of the ACM*, vol. 30, pp. 964–971, November 1987.

[16] H. Chen, "Collaborative systems: Solving the vocabulary problem," *Computer*, vol. 27, pp. 58–66, May 1994.

[17] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, pp. 79–86, March 1951.

[18] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, March 2003.

[19] B. Julesz, "Textons, the elements of texture perception, and their interactions," *Nature*, vol. 290, pp. 91–97, January 1981.

[20] C. O. Dumitru and M. Datcu, "Information content of very high resolution SAR images: Study of dependency of SAR image structure descriptors with incidence angle," *International Journal on Advances in Telecommunications*, vol. 5, pp. 239–251, June 2012.

[21] C. O. Dumitru and M. Datcu, "Study and assessment of selected primitive features behaviour for SAR image description," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3596–3599, July 2012.

[22] A. A. Popescu, I. Gavat, and M. Datcu, "Contextual descriptors for scene classes in very high resolution SAR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 9, pp. 80–84, January 2012.

[23] K. van de Sande, T. Gevers, and C. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1582 –1596, September 2010.

[24] S. Cui, C. Dumitru, and M. Datcu, "Ratio-detector-based feature extraction for very high resolution SAR image patch indexing," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, pp. 1175–1179, September 2013.

[25] I. I. Amr, M. Amin, P. El-Kafrawy, and A. M. Sauber, "Using statistical moment invariants and entropy in image retrieval," *International Journal of Computer Science and Information Security*, vol. 7, pp. 160–164, January 2010.

[26] M. J. E. Salami, A. Khorshidtalab, A. Baali, and A. M. Aibinu, "Classification of retinal images based on statistical moments and principal component analysis," in *Proc. International Conference on Computer and Communication Engineering (ICCCE)*, pp. 92–95, September 2014.

[27] L. Liu and P. Fieguth, "Texture classification from random features," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, pp. 574–586, March 2012.

[28] B. Hou, N. Li, S. Wang, and X. Zhang, "SAR image segmentation based on random projection and signature frame," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3726–3729, July 2014.

[29] D. Gong, S. Li, and Y. Xiang, "Face recognition using the Weber Local Descriptor," in *Proc. Asian Conference on Pattern Recognition (ACPR)*, pp. 589–592, November 2011.

[30] R. Touzi, A. Lopes, and P. Bousquet, "A statistical and geometrical edge detector for SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 26, pp. 764–773, November 1988.

[31] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 971–987, July 2002.

[32] A. Lucieer, P. Fisher, and A. Stein, "Texture-based segmentation of highresolution remotely sensed imagery for identification of fuzzy objects," in *Proc. GeoComputation Conference*, September 2003.

[33] C. Song, P. Li, and F. Yang, "Multivariate texture measured by local binary pattern for multispectral image classification," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 2145–2148, July 2006.

[34] C. Song, F. Yang, and P. Li, "Rotation invariant texture measured by local binary pattern for remote sensing image classification," in *Proc. International Workshop on Education Technology and Computer Science (ETCS)*, vol. 3, pp. 3–6, March 2010.

[35] M. Musci, R. Queiroz Feitosa, G. Costa, and M. Fernandes Velloso, "Assessment of binary coding techniques for texture characterization in remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 10, pp. 1607–1611, November 2013.

[36] D. Dai, W. Yang, and H. Sun, "Multilevel local pattern histogram for SAR image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 8, pp. 225–229, March 2011.

[37] M. De Martinao, F. Causa, and S. Serpico, "Classification of optical high resolution images in urban environment using spectral and textural information," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 1, pp. 467–469 vol.1, July 2003.

[38] L.-K. Soh and C. Tsatsoulis, "Texture analysis of SAR sea ice imagery using gray level co-occurrence matrices," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 780–795, March 1999.

[39] W. Lv, Q. Yu, and W. Yu, "Water extraction in SAR images using GLCM and support vector machine," in *Proc. IEEE International Conference on Signal Processing (ICSP)*, pp. 740–743, October 2010.

[40] P. M. Treitz, P. J. Howarth, O. R. Filho, and E. D. Soulis, "Agricultural crop classification using SAR tone and texture statistics," *Canadian Journal of Remote Sensing*, vol. 26, no. 1, pp. 18–29, 2000.

[41] U. Kandaswamy, D. A. Adjeroh, and M. C. Lee, "Efficient texture analysis of SAR imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, pp. 2075–2083, September 2005.

[42] R. Chellappa and S. Chatterjee, "Classification of textures using Gaussian Markov random fields," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33, pp. 959–963, August 1985.

[43] I.-W. Tsai and D.-C. Tseng, "Segmentation of multispectral remote sensing images based on Markov random fields," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, vol. 1, pp. 264–266 vol.1, August 1997.

[44] D. Gleich and M. Datcu, "Wavelet-based despeckling of SAR images using Gauss Markov random fields," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 4127–4143, December 2007.

[45] D. Espinoza Molina, D. Gleich, and M. Datcu, "Evaluation of Bayesian despeckling and texture extraction methods based on Gauss Markov and autobinomial Gibbs random fields: Application to TerraSAR-X data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, pp. 2001–2025, May 2012.

[46] D. Gleich and M. Datcu, "Despeckling and information extraction from SLC SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, pp. 4633–4649, August 2014.

[47] J. Singh, *Spatial Content Understanding of Very High Resolution Synthetic Aperture Radar Images*. PhD thesis, University of Siegen, 2014.

[48] W. Xu and S. Wang, "SAR detection of moving targets based on matched Fourier transform," in *Proc. Asian and Pacific Conference on Synthetic Aperture Radar (APSAR)*, pp. 289–292, November 2007.

[49] J.-j. Chen, J. Chen, and S.-l. Wang, "Detection of ultra-high speed moving target based on matched Fourier transform," in *Proc. International Conference on Radar*, pp. 1–4, October 2006.

[50] C. Wang and S.-b. Li, "SAR imaging of moving targets based on second order match Fourier transform," in *Proc. IEEE International Conference on Signal Processing (ICSP)*, vol. 3, pp. 1971–1974, October 2012.

[51] S.-l. Wang, S.-g. Li, J.-l. Ni, and G.-y. Zhang, "A new transform–matched Fourier transform," *Acta Electronic Sinica*, vol. 29, no. 3, pp. 403–405, 2001.

[52] H.-B. Sun, G.-S. Liu, H. Gu, and W. min Su, "Application of the fractional Fourier transform to moving target detection in airborne SAR," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 38, pp. 1416–1424, October 2002.

[53] L. B. Almeida, "The fractional Fourier transform and time-frequency representations," *IEEE Transactions on Signal Processing*, vol. 42, pp. 3084–3091, November 1994.

[54] J. Singh and M. Datcu, "Mining very high resolution complex-valued SAR images using the fractional Fourier transform," in *Proc. European Conference on Synthetic Aperture Radar (EUSAR)*, pp. 135–138, April 2012.

[55] J. Singh and M. Datcu, "Parametric modeling of the fractional Fourier transform coefficients for complex-valued SAR image categorization," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 2882–2886, September 2013.

[56] J. Singh and M. Datcu, "SAR image categorization with log cumulants of the fractional Fourier transform coefficients," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 5273–5282, December 2013.

[57] A. Popescu, C. Patrascu, J. Singh, I. Gavat, and M. Datcu, "Spotlight TerraSAR-X data modeling using spectral space-variant measures, for scene targets and structure indexing," in *Proc. European Conference on Synthetic Aperture Radar (EUSAR)*, pp. 1–4, June 2010.

[58] T. Li and M. Ogihara, "Toward intelligent music information retrieval," *IEEE Transactions on Multimedia*, vol. 8, pp. 564–574, June 2006.

[59] E. P. Simoncelli and E. H. Adelson, "Non-separable extensions of quadrature mirror filters to multiple dimensions," *Proceedings of the IEEE*, vol. 78, pp. 652–664, April 1990.

[60] A. Croisier, D. Esteban, and C. Galand, "Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques," in *Proc. International Conference on Information Science and Systems*, 1976.

[61] D. Esteban and C. Galand, "Application of quadrature mirror filters to split band voice coding schemes," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, pp. 191–195, May 1977.

[62] C.-S. Li and V. Castelli, "Deriving texture feature set for content-based retrieval of satellite image database," in *Proc. International Conference on Image Processing*, vol. 1, pp. 576–579, October 1997.

[63] S. Fukuda and H. Hirosawa, "A wavelet-based texture feature set applied to classification of multifrequency polarimetric SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 37, pp. 2282–2286, September 1999.

[64] F. Farrokhnia and A. K. Jain, "A multi-channel filtering approach to texture segmentation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 364–370, June 1991.

[65] "MPEG-7." http://mpeg.chiariglione.org/standards/mpeg-7. [Online].

[66] S. Newsam, L. Wang, S. Bhagavathy, and B. S. Manjunath, "Using texture to analyze and manage large collections of remote sensed image and video data," *Journal of Applied Optics: Information Processing*, vol. 43, January 2004.

[67] L.-j. Du, "Texture segmentation of SAR images using localized spatial filtering," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1983–1986, May 1990.

[68] J. H. Lee and W. D. Philpot, "A spectral-textural classifier for digital imagery," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, May 1990.

[69] L. Shu, T. Tan, M. Tang, and C. Pan, "A novel registration method for SAR and SPOT images," in *Proc. IEEE International Conference on Image Processing (ICIP)*, vol. 2, pp. 213–16, September 2005.

[70] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157, September 1999.

[71] M. Hasan, M. Pickering, and X. Jia, "Modified SIFT for multi-modal remote sensing image registration," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 2348–2351, July 2012.

[72] A. Sedaghat and H. Ebadi, "Remote sensing image matching based on adaptive binning SIFT descriptor," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 5283–5293, October 2015.

[73] Y. Yuan and X. Hu, "Bag-of-Words and object-based classification for cloud extraction from satellite imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, pp. 4197–4205, August 2015.

[74] Y. Yang and S. Newsam, "Comparing SIFT descriptors and Gabor texture features for classification of remote sensed imagery," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 1852–1855, October 2008.

[75] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, September 1995.

[76] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, pp. 453–466, January 2015.

[77] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded up robust features," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 404–417, May 2006.

[78] Z. L. Song and J. Zhang, "Remote sensing image registration based on retrofitted SURF algorithm and trajectories generated from lissajous figures," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, pp. 491–495, July 2010.

[79] X. Tian, C. Wang, and H. Zhang, "Extraction of object features from high resolution SAR images based on SURF features," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 1802–1805, July 2011.

[80] B. Pang, H. Sun, Q. Yu, and P. Wu, "A hybrid SAR image registration algorithm base on SURF and mutual information," in *Proc. IEEE Asia-Pacific Conference on Synthetic Aperture Radar (APSAR)*, pp. 379–382, September 2015.

[81] C.-R. Shyu, M. Klaric, G. J. Scott, A. S. Barb, C. H. Davis, and K. Palaniappan, "Geoiris: Geospatial information retrieval and indexing system—content mining, semantics modeling, and complex queries," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 45, pp. 839–852, April 2007.

[82] S. Cui, M. Datcu, and P. Blanchart, "Cascade active learning for SAR image annotation," in *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 2000–2003, July 2012.

[83] L. Zhao, P. Tang, and L. Huo, "Land-use scene classification using a concentric circle-structured multiscale Bag-of-Visual-Words model," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2015, In press.

[84] M. Lienou, H. Maitre, and M. Datcu, "Semantic annotation of satellite images using Latent Dirichlet Allocation," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, pp. 28–32, January 2010.

[85] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature coding in image classification: A comprehensive study," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 493–506, March 2014.

[86] Y.-L. Boureau, J. Ponce, and Y. Lecun, "A theoretical analysis of feature pooling in visual recognition," in *Proc. International Conference on Machine Learning (ICML)*, pp. 111–118, June 2010.

[87] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV Workshop on Statistical Learning in Computer Vision*, pp. 1–22, May 2004.

[88] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders, "Kernel codebooks for scene categorization," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 696–709, October 2008.

[89] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1794–1801, June 2009.

[90] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Proc. Advances in Neural Information Processing Systems (NIPS)*, pp. 2223–2231, December 2009.

[91] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3360–3367, June 2010.

[92] Y. Huang, K. Huang, Y. Yu, and T. Tan, "Salient coding for image classification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1753–1760, June 2011.

[93] Z. Wu, Y. Huang, L. Wang, and T. Tan, "Group encoding of local features in image classification.," in *Proc. International Conference on Pattern Recognition (ICPR)*, pp. 1505–1508, November 2012.

[94] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, pp. 2323–2326, December 2000.

[95] M. Babaee, G. Rigoll, R. Bahmanyar, and M. Datcu, "Locally linear salient coding for image classification," in *Proc. International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–4, June 2014.

[96] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman, "Discovering object categories in image collections," in *Proc. International Conference on Computer Vision (ICCV)*, October 2005.

[97] A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via pLSA," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 517–530, May 2006.

[98] R. Lienhart and M. Slaney, "PLSA on large scale image databases," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 4, pp. 1217–1220, April 2007.

[99] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *Proc. ACM International Conference on Image and Video Retrieval (CIVR)*, pp. 17–24, July 2007.

[100] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *The Journal of Machine Learning*, January.

[101] X. Wei and W. B. Croft, "LDA-based document models for ad-hoc retrieval," in *Proc. ACM International Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 178–185, August 2006.

[102] L. Cao and L. Fei-Fei, "Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, October 2007.

[103] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 524–531, June 2005.

[104] R. Bahmanyar, S. Cui, and M. Datcu, "A comparative study of Bag-of-Words and Bag-of-Topics models of EO image patches," *Geoscience and Remote Sensing Letters, IEEE*, vol. 12, pp. 1357–1361, June 2015.

[105] J. Zhang, M. Marszaek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, pp. 213–238, June 2007.

[106] H. Bannour, L. Hlaoua, and B. Ayeb, "Survey of the adequate descriptor for content-based image retrieval on the web: Global versus local features," in *Proc. CORIA*, pp. 445–456, 2009.

[107] J. Heinly, E. Dunn, and J.-M. Frahm, "Comparative evaluation of binary features," in *Proc. European Conference on Computer Vision (ECCV)*, pp. 759–773, October 2012.

[108] R. Bahmanyar and M. Datcu, "Measuring the semantic gap based on a communication channel model," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 4377–4381, September 2013.

[109] R. Bahmanyar, G. Rigoll, and M. Datcu, "A clustering-based approach for evaluation of EO image indexing," in *Proc. ISPRS Sensors and Models in Photogrammetry and Remote Sensing (SMPR)*, pp. 79–84, ISPRS, October 2013.

[110] R. Bahmanyar, M. Datcu, and G. Rigoll, "Comparing the information extracted by feature descriptors from EO images using Huffman coding," in *Proc. International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, June 2014.

[111] I. Biederman, "Recognition-by-components: a theory of human image understanding," *Psychological Review*, vol. 94, no. 2, pp. 115–147, 1987.

[112] R. Bahmanyar and A. Murillo Montes de Oca, "Evaluating the sensory gap for earth observation images using human perception and an LDA-based computational model," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 566–570, September 2015.

[113] R. Zhao and W. Grosk, "Negotiating the semantic gap: from feature maps to semantic landscapes," *Pattern Recognition*, vol. 35, no. 3, pp. 593 – 600, 2002.

[114] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1521–1528, June 2011.

[115] Z. Theodosiou, C. Kasapi, and N. Tsapatsoulis, "Semantic gap between people: An experimental investigation based on image annotation," in *International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, pp. 73–77, December 2012.

[116] H. Ma, J. Zhu, M. R. T. Lyu, and I. King, "Bridging the semantic gap between image contents and tags," *IEEE Transactions on Multimedia*, vol. 12, pp. 462–473, August 2010.

[117] H.-Y. Ha, F. C. Fleites, and S.-C. Chen, "Building multi-model collaboration in detecting multimedia semantic concepts," in *Proc. International Conference Collaborative Computing*, pp. 205–212, October 2013.

[118] J. S. Hare, P. A. S. Sinclair, P. H. Lewis, K. Martinez, P. G. B. Enser, and C. J. S, "Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches," in *Proc. European Semantic Web Conference (ESWC)*, Springer Verlag, June 2006.

[119] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom, "Mind the gap: Another look at the problem of the semantic gap in image retrieval," in *Multimedia Content Analysis, Management and Retrieval*, vol. SPIE V, SPIE and IS&T, January 2006.

[120] C. Liu and G. Song, "A method of measuring the semantic gap in image retrieval: Using the information theory," in *Proc. International Conference on Image Analysis and Signal Processing (IASP)*, pp. 287–291, October 2011.

[121] R. Bahmanyar, A. Murillo Montes de Oca, and M. Datcu, "The semantic gap: An exploration of user and computer perspectives in earth observation images," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, pp. 2046–2050, October 2015.

[122] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Applied Statistics*, vol. 28, no. 1, pp. 100–108, 1979.

[123] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of the IRE*, vol. 40, pp. 1098–1101, September 1952.

[124] C. M. Bishop, *Pattern Recognition and Machine Learning*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[125] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006.

[126] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geoscience Frontiers*, 2015, In press.

[127] Z. Lv, Y. Hu, H. Zhong, J. Wu, B. Li, and H. Zhao, "Parallel k-means clustering of remote sensing images based on mapreduce," in *Proc. International Conference on Web Information Systems and Mining (WISM)*, pp. 162–170, October 2010.

[128] C. Huo, Z. Zhou, H. Lu, C. Pan, and K. Chen, "Fast object-level change detection for VHR images," *IEEE Geoscience and Remote Sensing Letters*, vol. 7, pp. 118–122, January 2010.

[129] C. O. Dumitru, S. Cui, G. Schwarz, and M. Datcu, "Information content of very-high-resolution SAR images: Semantics, geospatial context, and ontologies," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, pp. 1635–1650, April 2015.

[130] C. O. Dumitru, S. Cui, D. Faur, and M. Datcu, "Data analytics for rapid mapping: Case study of a flooding event in Germany and the tsunami in Japan

using very high resolution SAR images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, pp. 114–129, January 2015.

[131] C. Vaduva, T. Costachioiu, C. Patrascu, I. Gavat, V. Lazarescu, and M. Datcu, "A latent analysis of earth surface dynamic evolution using change map time series," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 2105–2118, April 2013.

[132] T. Costachioiu, I. Nita, V. Lazarescu, and M. Datcu, "A semantic framework for data retrieval in large remote sensing databases," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 5285–5288, July 2012.

[133] T. Costachioiu, R. Constantinescu, B. AlZenk, and M. Datcu, "Semantic analysis of satellite image time series," in *Proc. European Signal Processing Conference (EUSIPCO)*, pp. 2492–2495, August 2012.

[134] C. Vaduva, I. Gavat, and M. Datcu, "Latent Dirichlet allocation for spatial analysis of satellite images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 51, pp. 2770–2786, May 2013.

[135] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 1 ed., July 2008.

[136] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[137] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, August 1999.

[138] D. M. Blei, "Probabilistic topic models," *Communications of the ACM*, vol. 55, pp. 77–84, April 2012.

[139] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, 1948.

[140] G. Wade, *Signal Coding and Processing, (Second Edition)*. Cambridge University Press, September 1994.

[141] T. A. Welch, "A technique for high-performance data compression," *Computer*, vol. 17, pp. 8–19, June 1984.

[142] D. Cerra and M. Datcu, "Algorithmic cross-complexity and relative complexity," in *Proc. Data Compression Conference (DCC)*, pp. 342–351, March 2009.

[143] D. Espinoza-Molina, M. Quartulli, and M. Datcu, "Query by example in earth-observation image archive using data compression-based approach," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 6035–6038, July 2012.

[144] D. Espinoza-Molina and M. Datcu, "Urban area understanding based on compression methods," in *Proc. Joint Urban Remote Sensing Event (JURSE)*, pp. 174–177, April 2013.

[145] D. Espinoza-Molina, J. Chadalawada, and M. Datcu, "SAR image content retrieval by speckle robust compression based methods," in *Proc. European Conference on Synthetic Aperture Radar (EUSAR)*, pp. 1–4, June 2014.

[146] D. Cerra and M. Datcu, "A fast compression-based similarity measure with applications to content-based image retrieval," *Journal of Visual Communication and Image Representation*, vol. 23, no. 2, pp. 293 – 302, 2012.

[147] J. Chadalawada, D. Espinoza-Molina, and M. Datcu, "Assessment of earth observation data content based on data compression - application to settlements understanding," in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 6130–6133, July 2012.

[148] H. Liu, Z. Yang, Z. Wu, and X. Li, "A-optimal non-negative projection for image representation," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1592–1599, June 2012.

[149] L. K. Saul and S. T. Roweis, "An introduction to locally linear embedding," tech. rep., 2000. https://www.cs.nyu.edu/ roweis/lle/publications.html.

[150] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman, "The devil is in the details: an evaluation of recent feature encoding methods," in *Proc. British Machine Vision Conference*, pp. 76.1–76.12, August 2011.

[151] V. Levenshtein, "Binary codes capable of correcting deletions, insertions and reversals," *Soviet Physics Doklady*, vol. 10, p. 707, 1966.

[152] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proc. Annual ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–273, July 2003.

[153] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz, "Scene perception: Detecting and judging objects undergoing relational violations," *Cognitive Psychology*, vol. 14, no. 2, pp. 143–177, 1982.

[154] C. Green and J. E. Hummel, "Familiar interacting object pairs are perceptually grouped," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 32, pp. 1107–1119, October 2006.

[155] A. Oliva and A. Torralba, "The role of context in object recognition," *Trends in Cognitive Sciences*, vol. 11, pp. 520–527, December 2007.

[156] A. Torralba, "How many pixels make an image?," *Visual Neuroscience*, vol. 26, pp. 123–131, January 2009.

[157] E. Barenholtz, "Quantifying the role of context in visual object recognition," *Visual Cognition*, vol. 22, pp. 30–56, December 2013.

[158] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: A database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, pp. 157–173, October 2008.

[159] A. Hanbury, "A survey of methods for image annotation," *Journal of Visual Languages and Computing*, vol. 19, pp. 617–627, October 2008.

[160] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, pp. 861–874, June 2006.

[161] R. Likert, "A technique for the measurement of attitudes.," *Archives of Psychology*, vol. 22, no. 140, pp. 1–55, 1932.

[162] H. Hutt, R. Everson, M. Grant, J. Love, and G. Littlejohn, "How clumpy is my image? evaluating crowdsourced annotation tasks," in *Proc. UK Workshop on Computer Intelligence (UKCI)*, pp. 136–143, September 2013.

[163] M. Bar, "The proactive brain: using analogies and associations to generate predictions.," *Trends in Cognitive Sciences*, vol. 11, pp. 280–9, July 2007.

[164] C. L. Zitnick and D. Parikh, "The role of image understanding in contour detection," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 622–629, June 2012.

[165] J. Yuan, D. Wang, and R. Li, "Remote sensing image segmentation by combining spectral and texture features," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 52, pp. 16–24, January 2014.

[166] E. Weiszfeld and F. Plastria, "On the point for which the sum of the distances to n given points is minimum," *Annals of Operations Research*, vol. 167, pp. 7–41, March 2009.

[167] B. S. Manjunath and W. Y. Ma, "Texture features for browsing and retrieval of image data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 837–842, August 1996.

[168] I. FÃrber, S. GÃnnemann, H.-p. Kriegel, P. KrÃger, E. MÃller, E. Schubert, T. Seidl, and A. Zimek, "On using class-labels in evaluation of clusterings," in *Proc. International Workshop on Discovering, Summarizing and Using Multiple Clusterings (MultiClust) in conjunction with ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 2010.

[169] M. Halkidi and M. Vazirgiannis, "Clustering validity assessment: finding the optimal partitioning of a data set," in *Proc. IEEE International Conference on Data Mining (ICDM)*, pp. 187–194, November 2001.

[170] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *Journal of Machine Learning Research*, vol. 9999, pp. 2837–2854, December 2010.

[171] B. Larsen and C. Aone, "Fast and effective text mining using linear-time document clustering," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 16–22, June 1999.

[172] T. Caliński and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Statistics - Theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[173] D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 24–227, 1979.

[174] Y. Liu, Z. Li, H. Xiong, X. Gao, and J. Wu, "Understanding of internal clustering validation measures," in *Proc. IEEE International Conference on Data Mining (ICDM)*, (Washington, DC, USA), pp. 911–916, IEEE Computer Society, July 2010.

[175] E. Rendón, I. M. Abundez, C. Gutierrez, S. D. Zagal, A. Arizmendi, E. M. Quiroz, and H. E. Arzate, "A comparison of internal and external cluster validation indexes," in *Proc. American Conference on Applied Mathematics and WSEAS International Conference on Computer Engineering and Applications*, pp. 158–163, January 2011.

[176] P.-N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining, (First Edition)*. Addison-Wesley Longman Publishing Co., Inc., 2005.

[177] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, 1985.

[178] N. X. Vinh and J. Epps, "A novel approach for automatic number of clusters detection in microarray data based on consensus clustering," *Proc. IEEE International Conference on Bioinformatics & Bioengineering (BIBE)*, pp. 84–91, June 2009.

[179] S. J. Simoff, M. H. Bhlen, and A. Mazeika, *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*, vol. 4404. Springer, 2008.

[180] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, pp. 644–655, September 1998.

[181] B. Settles, "Active learning literature survey," tech. rep., University of Wisconsin, Madison, 2010.

[182] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky, and M. Ward, "Semantic image browser: Bridging information visualization with automated intelligent image analysis," in *Proc. IEEE Symposium on Visual Analytics Science and Technology*, pp. 191–198, October 2006.

[183] N. Férey, P. E. Gros, J. Hérisson, and R. Gherbi, "Visual data mining of genomic databases by immersive graph-based exploration," in *Proc. International Conference on Computer Graphics and Interactive Techniques in Australasia and South East Asia*, pp. 143–146, November 2005.

[184] H. Azzag, F. Picarougne, C. Guinot, and G. Venturini, "VRMiner: A tool for multimedia database mining with virtual reality," in *Database Technologies: Concepts, Methodologies, Tools, and Applications* (J. Erickson, ed.), pp. 1151–1167, IGI Global, 2009.

[185] M. Nakazato and T. S. Huang, "3D MARS: Immersive virtual reality for content-based image retrieval," in *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, p. 12, August 2001.

[186] M. Babaee, R. Bahmanyar, G. Rigoll, and M. Datcu, "Interactive clustering for SAR image understanding," in *Proc. European Conference on Synthetic Aperture Radar (EUSAR)*, pp. 1–4, June 2014.

[187] M. Babaee, G. Rigoll, and M. Datcu, "Immersive interactive information mining with application to earth observation data retrieval," in *Availability, Reliability, and Security in Information Systems and HCI* (A. Cuzzocrea, C. Kittl,

D. Simos, E. Weippl, and L. Xu, eds.), vol. 8127 of *Lecture Notes in Computer Science*, pp. 376–386, Springer Berlin Heidelberg, 2013.

[188] M. Babaee, M. Datcu, and G. Rigoll, "Assessment of dimensionality reduction based on communication channel model; application to immersive information visualization," in *Proc. IEEE International Conference on Big Data*, pp. 1–6, October 2013.

[189] M. Babaee, R. Bahmanyar, G. Rigoll, and M. Datcu, "Immersive visual information mining for exploring the content of EO archives," in *Proc. ESA Living Planet Symposium*, (Edinburgh), September 2013.

[190] B. Schölkopf, A. J. Smola, and K.-R. Müller, "Kernel principal component analysis," in *Advances in Kernel Methods* (B. Schölkopf, C. J. C. Burges, and A. J. Smola, eds.), pp. 327–352, Cambridge, MA, USA: MIT Press, 1999.

[191] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, October 1999.

[192] M. Babaee, R. Bahmanyar, G. Rigoll, and M. Datcu, "Farness preserving non-negative matrix factorization," in *Proc. IEEE International Conference on Image Processing (ICIP)*, pp. 3023–3027, October 2014.

[193] D. Teleaga, M. Buican, C. Cucu-Dumitrescu, F. Serban, and M. Datcu, "Visual data mining for exploration of EO images archives," in *Proc. ESA Living Planet Symposium*, (Edinburgh), September 2013.