Towards Semantic-Based Image Annotation

By

Andres Dorado

Supervised By

Prof. Ebroul Izquierdo

A Dissertation Submitted in Partial Fulfilment
of the Requirements for the Degree of
Doctor of Philosophy
in
Electronic Engineering

Department of Electronic Engineering
Queen Mary, University of London
2005

Abstract

This research work addresses the problem of using concept-related indexing of image content as a near-automatic way to perform semantic image annotation. The main objective is to provide a framework in which lexical information of visual interpretations and their components (concept-related indexes) can be used to perform content-based image annotation.

Several design phases starting from the formation of an MPEG-7 learning space to the construction of a robust semantic indexer were applied. Salient features of the proposed framework for concept-related indexing of image content are:

Provide a suitable combination of low-level visual features. A specific concern is on defining a structure in which MPEG-7 descriptor elements may be aggregated into feature vectors. A structure is proposed to preserve the semantics embedded in descriptors, avoid description overriding, and control the vector dimensionality using the minimum number of required elements.

Unambiguous interpretation is reduced using a built-in knowledge base consisting of concepts organized into a restrained lexicon.

A *learning procedure*, which applies partially supervised clustering, is an important component within the framework when presenting the exemplars to the learner. A fuzzy partition of the learning space is used not only to approximate semantically meaningful groups, but also to facilitate long-term learning.

Semantic profiles are proposed to incorporate conceptualisation of image content into clusters. The underlying relationship between features vectors is considered in defining the semantic profiles. In addition, a matching procedure is proposed to estimate distance between semantic profiles.

High scalability by using partitions rather than the entire learning space to add new image concepts, or new image representations, is provided. This issue relates long-term learning and contributes to improve generalization capabilities.

Experimental studies demonstrated how the proposed framework leads to a potential solution of equipping content-based image retrieval systems with learnable concepts useful to deal with meta-information.

Contents

1 Introduction **1**

1.1 Motivation . . . 1

1.1.1 Multimedia Computing . . . 1

1.1.2 Content-Based Image Retrieval (CBIR) . . . 2

1.2 Problem Definition . . . 2

1.2.1 The Semantic Gap . . . 2

1.2.2 Research Problem . . . 4

1.3 Objectives . . . 5

1.4 Research Contributions . . . 5

1.5 Structure of the Document . . . 7

2 Conceptual Image Analysis **8**

2.1 Image Abstraction and Interpretation . . . 8

2.2 Classification and Conceptual Analysis . . . 10

2.2.1 Semantic Classification . . . 11

2.2.2 The Lexicon . . . 13

2.2.3 Semantic Annotation . . . 14

2.3 State-of-the-Art . . . 16

2.3.1 Image Abstractions . . . 16

2.3.2 Capturing Knowledge and Interpretations . . . 17

2.3.3 Image Analysis Techniques . . . 18

2.3.4 CBIR Systems . . . 20

2.3.5 Summary of Selected Experimental Studies . . . 21

2.4 On the Road to Semantic Annotation . . . 21

3 An MPEG-7 Learning Space **23**

3.1 MPEG-7 . . . 24

3.2 Feature Space . . . 25

3.2.1 Colour and Texture Descriptors . . . 26

3.2.2 Selection of Descriptor Elements . . . 28

3.3 Empirical Evaluation of Feature Vector Quality . . . 30
3.3.1 Univariate Analysis . . . 30
3.3.2 Inter-Class Analysis . . . 31
3.3.3 Principal Components Analysis . . . 36
3.4 Summary . . . 38

4 User-Driven Fuzzy Clustering 39
4.1 Unsupervised Fuzzy Partition of the Feature Space . . . 40
4.2 Structural Data Analysis . . . 42
4.3 Partially Supervised Clustering for Semantic Classification . . . 44
4.4 Summary . . . 47

5 A Framework for Concept-related Indexing of Image Content Phase One: Preceding Case Studies 48
5.1 Case Study I: Fuzzy Inference Approach . . . 48
5.1.1 Learning Unit . . . 49
5.1.2 Indexing Unit . . . 50
5.1.3 Test Conditions . . . 51
5.1.4 Evaluation . . . 52
5.2 Case Study II: Radial Basis Function Network Approach . . . 56
5.2.1 Learning Unit . . . 56
5.2.2 Indexing Unit . . . 58
5.2.3 Test Conditions . . . 59
5.2.4 Evaluation . . . 61
5.3 Summary . . . 64

6 A Framework for Concept-related Indexing of Image Content Phase Two: Design and Evaluation 66
6.1 Framework Overview . . . 67
6.1.1 Learning Unit . . . 67
6.1.2 Indexing Unit . . . 69
6.2 Semantic Profiles . . . 70
6.2.1 Concept Profiling . . . 70
6.2.2 Proposed Matching Procedure for Performance Evaluation . . . 71
6.3 Evaluation . . . 74
6.3.1 Test Conditions . . . 74
6.3.2 Hierarchical Clustering . . . 75
6.3.3 Single-element instance . . . 76
6.3.4 Multiple-element instance: Enlarging the Lexicon . . . 79

6.4 Summary of Experimental Studies 84

7 Conclusions and Discussion **85**

7.1 Conclusions 85

7.2 Discussion and Further Work 87

8 Bibliography 89

A The Wilcoxon Test **96**

B Rule Generation **99**

B.1 Frequent Pattern Mining 99

B.2 An Abstract Example 100

C Fuzzy Reasoning Model **102**

C.1 Fuzzy Systems 101

C.2 Fuzzification 101

C.3 Fuzzy Inference 103

C.4 Defuzzification 104

List of Figures

2.1 Granules of information 9
2.2 RGB and HSV colour spaces 10
2.3 Semantic image classification 12
2.4 Selection of training images 12
2.5 Indexes as entries to ontology 13
2.6 Corel annotations 14
2.7 Semantic image annotation 15

3.1 Scope of MPEG-7 24
3.2 Sample of colour layout description in XML format 26
3.3 Mapping from RGB to HMMD space 27
3.4 Sample of colour structure description in XML format 27
3.5 Sample of scalable colour description in XML format 27
3.6 Sample of edge histogram description in XML format 28
3.7 Sample of homogeneous texture description in XML format 28
3.8 Accuracy combining descriptor elements 29
3.9 CLD –Contour plots and statistical moments 31
3.10 CSD –Contour plots and statistical moments 31
3.11 SCD –Contour plots and statistical moments 31
3.12 EHD –Contour plots and statistical moments 32
3.13 HTD –Contour plots and statistical moments 32
3.14 Retained indexes of salient features 33
3.15 Feature selection process 33
3.16 Effects of confidence level on feature selection 34
3.17 Quantification of salient features 34
3.18 Retained indexes at different confidence levels 35
3.19 Performance using salient features 36
3.20 Colour –Scree plots 37
3.21 Texture –Scree plots 38
3.22 Accuracy of the semantic grouping using PCA 38

4.1 High- and low-level similarity . . . 41
4.2 Structural analysis . . . 42
4.3 Cluster-class dependencies . . . 43
4.4 Semantic overlapping . . . 43
4.5 partial supervision . . . 46
4.6 Structural data analysis . . . 47

5.1 Fuzzy-based indexing unit . . . 51
5.2 Samples of TRECVID images . . . 52
5.3 Samples of building images . . . 53
5.4 Rule base EHD . . . 54
5.5 Selected sub-images . . . 54
5.6 Hidden neuron model . . . 57
5.7 Linear activation levels . . . 59
5.8 RBFN-based indexing unit . . . 59
5.9 Sample of images used in the data set II . . . 60
5.10 Fuzzification factor effects . . . 61
5.11 Performance of RBFN approach . . . 62
5.12 Final stage RBFN approach . . . 63
5.13 RBFN performance . . . 63
5.14 Samples of right classification . . . 64
5.15 Samples of wrong classification . . . 65

6.1 Learning unit . . . 67
6.2 Indexing unit . . . 69
6.3 Dendrogram CURE . . . 75
6.4 Samples of clusters generated by CURE . . . 76
6.5 Expected profile . . . 78
6.6 Misleading clusters . . . 78
6.7 Clusters with well-ranked misclassified images . . . 79
6.8 Outliers . . . 79
6.9 Cluster profiles mi700-c50-t0.4 (colour) . . . 80
6.10 Cluster profiles mi700-c50-t0.4 (texture) . . . 81
6.11 Cluster profiles mi700-c50-t0.4 (small) . . . 81
6.12 Cluster profiles mi700-c50-t0.4 (matching effects) . . . 82
6.13 Cluster profiles mi700-c50-t0.4 (saturation and brightness) . . . 83
6.14 Cluster profiles mi700-c50-t0.4 (variety) . . . 83

B.1 Fuzzy variables . . . 99
B.2 Neck condition . . . 100

C.1 Fuzzy System 101
C.2 Piecewise linear functions 102
C.3 Mamdani fuzzy inference system 103
C.4 Sugeno fuzzy inference system 104
C.5 Center of Area 105

D.1 Work plan 107

List of Tables

2.1 Summary of research on image classification 22

3.1 MPEG-7 visual descriptors 25
3.2 Setting feature vector dimension 29
3.3 Percentage of retained elements 35

4.1 Ranking of class pairs 43
4.2 Partially supervised clustering algorithm 46

5.1 Fuzzy inference learning procedure 50
5.2 Fuzzy inference indexing procedure 51
5.3 Summary evaluation case study I 52
5.4 Fuzzy variables of EHD 53
5.5 Example of inference rules 55
5.6 Example of feature and fuzzy values 55
5.7 Results of case study I 55
5.8 RBFN learning procedure 58
5.9 RBFN indexing procedure 58
5.10 Summary evaluation case study II 61
5.11 Confusion matrix RBFN approach 62

6.1 Clustering learning procedure 68
6.2 Clustering indexing procedure 69
6.3 Semantic profiling procedure 70
6.4 Sample of cluster profile 71
6.5 A sample of position-based matching procedure 73
6.6 Proposed matching procedure to compare semantic profiles 73
6.7 Corel categories 74
6.8 Concepts in third data set 75
6.9 Summary framework evaluation I 76
6.10 Cluster profiles si5-c10-t0.8 76

6.11 Cluster profiles si5-c15-t0.8 77
6.12 Cluster profiles si5-c20-t0.8 77
6.13 Summary framework evaluation II 80
6.14 Cluster profiles mi700-c50-t0.4 (colour) 80
6.15 Cluster profiles mi700-c50-t0.4 (texture) 81
6.16 Cluster profiles mi700-c50-t0.4 (small) 81
6.17 Cluster profiles mi700-c50-t0.4 (matching effects) 82
6.18 Cluster profiles mi700-c50-t0.4 (saturation and brightness) 82
6.19 Cluster profiles mi700-c50-t0.4 (variety) 82
6.20 Summary of experimental studies 84

A.1 Rank-sum calculation 96

B.1 Fuzzy values 99
B.2 Transaction set 100

C.1 Z-shape membership function estimation 102
C.2 S-shape membership function estimation 102
C.3 Π-shape membership function estimation 103
C.4 Λ shape membership function estimation 103
C.5 Operators of aggregation (T-norm) 104
C.6 Operators of composition (T-conorm or S-norm) 105

Chapter 1

Introduction

1.1 Motivation

The growing number of available multimedia information and the continuously increasing number of tools for image processing is not only leading to new more appealing multimedia functionalities, but also bringing new challenges to the research community. This section introduces the two main subjects in whose this research is placed on: multimedia computing and content-based image retrieval.

1.1.1 Multimedia Computing

The integration of media in consumer electronics is facilitating a wide variety of applications ranging from multimedia cell phones, passing throughout personal digital assistants (PDA's) and camcorders, to high-definition interactive television. Multimedia content is delivery to users located at offices, homes, and more often everywhere (i.e. commuters).

These trends have turned out the attention of researchers and practitioners in computer vision, machine learning, pattern recognition, and other related fields towards an area that is becoming increasingly important: *multimedia computing.*

Multimedia computing involves the capture, storage, manipulation, and transmission of multimedia objects such as text, handwritten data, audio objects, still images, 2-D/3-D graphics, animation, and full motion video [1]. It also relates to streaming media middleware, multimedia data abstractions, continuous media representations, and media coding.

Although quite impressive advances can be observed in this relatively new area, there is a growing need of technologies that can adapt multimedia content to diverse client devices in the emerging "ubiquitous" or "pervasive" computing era [2]. Besides, there is also a need of systems to enable computational interpretation and processing of multimedia content and the automated generation or extraction of semantic knowledge from them [3].

Regarding to the interpretation of multimedia content, most of the research has been addressed to the problem of automated retrieval of syntactic and semantic data as a way to overcome the information glut [4].

In this context, image understanding is finding a place within multimedia computing technologies whereat it is playing a key role in the use of images to interface people and systems (or machines) [5]. Accordingly, several systems such as QBIC, VisualSEEk, VisualGREP, Virage have been developed [6]-[11] These systems follow the content-based image retrieval paradigm proposed in the early 1990's, in which images are represented using a set of feature attributes for further indexing, browsing, and retrieval [12].

1.1.2 Content-Based Image Retrieval (CBIR)

Initially, CBIR systems used traditional methods to designate a passage from low-level visual features to human understanding (of the image content) in order to provide a way that a computer can execute the recognition process [13]. This bottom-up approach relies completely on matching procedures at the lowest level of content representation.

CBIR systems based only on low-level matching procedures have poor performance for semantically specific requests. Intuitively, two objects can be similar in their visual primitives but semantically dissimilar. Besides, users do not think in terms of low-level representations.

Consequently, advanced CBIR systems have incorporated reasoning and learning capabilities to establish an intelligent architecture, which enable them to produce effective and accurate results according to the users' requests. Accepting the role that high-level semantic concepts play in the way human perceive images and measure their similarity, these systems combine approaches to go from bottom to top and in the opposite way to generate interpretations of image content. Such a combination is the foundation of the critical paradigm of "bridging the semantic gap" [14].

1.2 Problem Definition

Content-based image retrieval has been the subject of ongoing research for several decades evidencing its importance and the fact that many problems remain open. This section presents one of the holy-grail challenges in multimedia computing, the sensory data gap, what is followed by a definition of the problem in the proposed research work.

1.2.1 The Semantic Gap

Bridging the semantic gap refers to overcoming the obstruction known as the *sensory data gap* [12], which expresses the difference between human perception and interpretation of audio-visual information in all its different forms and the interpretation founded on a formal description derived from an automatic analysis of the machine [4].

There are several challenges in incorporating high-level matching procedures (top-down approach) in CBIR systems to conciliate human and machine interpretations. Some of these challenges are:

- *Subjectiveness.* Interpretations of the image content are subjective due to the different physiological and psychological responses of each beholder to visual stimuli. Most of these interpretations expressed through semantic concepts are not directly related to image attributes. Subjectivity could introduce ambiguity and inconsistency in the interpretations.

- *Adding interpretations.* The most natural way to add high-level interpretations to images is manual annotation. In addition to the subjectivity mentioned before, there are more shortcomings when using manual annotation. It is either unfeasible or impractical to manually annotated large-scale image databases. It has an undesirable effect in the system's performance, which is stalled at certain point for lack of information in decisions making without the necessary human factor. Moreover, there is not an agreed-upon (or widely accepted) scene description language.

- *Feature selection.* Advanced CBIR systems require more robust schemes to identify salient features of images that capture certain aspect of semantic content of these images. This is a difficult task because only low-level features can be reliably extracted from images and there is limited or no contextual information about the image content.

- *High dimensional description space.* These systems deal with the problem of integrating heterogeneous sets of features and domain knowledge. Some representations of image content use high dimensional feature vectors. The curse of dimensionality has an effect upon the computational efficiency [15].

- *Poor generalization.* High dimensionality along with the tremendous variety of images found within most semantic concepts derives in poor generalization beyond the training set.

At this point, recognizing the toughness of these challenges, a natural question is regarding to the feasibility of bridging (or at least narrowing) the semantic gap. This question has motivated experimental studies based on results obtained in another disciplines. For instance:

Studies concerning visual recognition in pigeons support the hypothesis for a mechanism that looks at collective low-level features for making the retrieval results more satisfactory using comparatively high-level processing [16]. The birds are presented with large sets of pictorial stimuli, including colour photographs, as part of a training procedure to respond differentially to the stimuli according to an experimenter-defined class rule (e.g. "tree" and "non-tree").

As indicated by Huber [16], “this type of experiment has received considerable attention... because they imply that the pigeons’ classification behaviour is mediated by abstract, or conceptual, rules, and therefore resembles the cognitive solution accomplished by humans.”

Assuming that it is possible to perform recognition without a complex conceptualisation process; another question is concerning content-based image interpretation without first attempting to recognize the image components (objects).

In this matter, Lipson et al. [17] does mention of a recognition strategy of scene content, supported by psychophysical evidence, in which humans may holistically classify visual stimuli before recognizing the individual parts. Furthermore, it is not always possible to divide an image syntactically into smaller parts. Santini et al. [18] and Iqbal and Aggarwal [19] have proposed approaches to capture user’s semantic understanding of images at a higher level than the region representation.

Furthermore, automated linguistic indexing of pictures is a critically important area of research because of its demonstrated potential to narrow the semantic gap [20].

Notwithstanding the challenges remain in certain degree open, the progress in performing “intelligent” processing of the image content and equipping CBIR systems with capabilities for more efficient indexing, browsing, and retrieval is noticeable [ref http://www-nlpir.nist.gov/projects/trecvid/]. Moreover, there are important advances in standardization activities such as ISO/IEC MPEG-7 and its Visual Core Experiments as well as the recent Still Image Search project (JPSearch) [ref www.jpeg.org]. These standards not only are contributing to a better image representation and content analysis, but also to overcome another challenging problem for CBIR systems as is to process information placed at remote image databases.

Thus, progress in multimedia computing, more specifically in CBIR systems, is enabling more accessibility to the ever-increasing amount of multimedia content and is providing means to achieve users’ satisfaction.

1.2.2 Research Problem

Albeit the problem of narrowing the sensory data gap has been subject of extensive research, it is still unsolved. Much work has been focused on overcoming the CBIRs’ challenges namely subjectiveness, adding interpretations, feature selection, high dimensional description space, and poor generalization.

Nack [4] pointing out drawbacks on automate data retrieval said that it seems “new” problems are being solve with “old” solutions, which only furthers the crisis. He added that approaches using low-level feature extraction for automated audio-visual media understanding based on examples have been played out and failed so many times.

Though it is acceptable that some solutions are not suitable for the current challenges in multimedia computing, is also recognized that traditional methods from computer vision and pattern recognition contribute in a significant way to the solution of the present problems. An aspect to consider is the efficiency of such contributions in new computational environments, i.e. networking and computer architectures.

Some reasons for the inadequacy of proposed solutions for automatic media understanding are due to trying to find out semantics in low-level representations; using short-term learning procedures; or pretending to build general purpose systems using domain specific knowledge.

These reasons should be considered by any mechanism, approach, or method that pursues to narrow the semantic gap. In that way, concept-related indexing of image content is a promising approach to incorporate semantic interpretations into low-level descriptions. Indexing entities using a set of concept-dependent and interrelated descriptors establish a solid foundation to build concept-related (linguistic) models as part of the human-centred trends in semantic image annotation [21].

> This research work addresses the problem of using concept-related indexing of image content as a near-automatic way to perform semantic image annotation.

1.3 Objectives

The *main objective* in this research work is to provide a framework in which lexical information of visual interpretations and their components (concept-related indexes) can be used to perform content-based image annotation. Within this framework:

- The *first aim* is to propose a suitable combination of low-level image features.
- The *second aim* is to provide a mechanism to present the semantics, i.e. conceptualisation of image content, to the learner (CBIR system).
- The *third aim* is to provide high scalability. It facilitates long-term learning, in which new concepts or new images are added to the system and only part of the training set is involved in the learning update.
- The *Fourth aim* is to satisfy interoperability requirements. The framework is adjusted to the ISO/IEC standard MPEG-7.

1.4 Research Contributions

A scheme derived from the fundamental paradigm and design methodology of pattern recognition was followed in implementing the framework [22]. Several design phases starting from the formation of an MPEG-7 learning space to the construction of a robust semantic indexer were applied. Salient features of the proposed framework for concept-related indexing of image content are:

- *A suitable combination of low-level visual features.* Adhering to the *de facto* standard of image description proposed in the ISO/IEC standard MPEG-7, the interest turned out to the use of MPEG-7 visual descriptors for defining a learning feature space. A specific concern is on defining a structure in which descriptor elements may be aggregated, in a comprehensive manner, into feature vectors with a controlled dimensionality.

  Proposed structure preserves the semantics embedded in descriptor elements, avoids description overriding, and controls the vector dimensionality using the minimum number of required elements.

  Statistical and soft computing techniques are used to evaluate the quality of proposed structure regarding to its ability to discriminate examples ascribed to different semantic interpretations [23].

- *Unambiguous interpretation.* Interpretations of image content referred as image concepts are organized into a restrained lexicon (e.g. building, outdoor, animal, etc.) It is use to reduce ambiguity in mapping images into many possible interpretations. The lexicon is a built-in knowledge base consisting of basic semantic units denoted by symbols, i.e. icons or keywords. Then, the CBIR system is treated as an intelligent entity that associates image abstractions to its lexicon [24] [26].

- *Learning procedure.* An important component within the framework is the mechanism when presenting the exemplars to the learner. Statistical and soft computing techniques were evaluated in order to find out the most appropriate method to partition the learning space into semantically meaningful groups. The reason beneath this partitioning approach is presented below (See *high scalability*).

  Some of the techniques evaluated are: association rule mining, fuzzy inference systems, radial basis function networks, support vector machines, hierarchical and partitioning clustering [27] -[35].

  For the specific problem domain defined to validate the framework, the best performance was achieved by a proposed partially supervised clustering algorithm, which implements the learning procedure [36].

- *Semantic profile.* Clustering approach incorporates prior domain knowledge, through labelled data, using partially supervised learning. Resulting clusters are tagged with concepts extracted from training images. The assigned set of concepts determines the semantic profile of each cluster. The underlying relationship between features vectors is considered in defining the semantic profiles. In addition, a matching procedure is proposed to estimate distance between image-cluster, image-image, or cluster-cluster using their semantic profiles. This procedure combines low- and high-level matching.

- *Semantic separability and generalization.* The learner is presented with sets of feature vectors representing samples of images that can or cannot be associated to certain concept (or group of concepts). Separability is satisfied when descriptions are very different for images associated to different semantic profiles. On the other hand, generalization is achieved when descriptions contain very similar values for images under a similar semantic profile.

- *High scalability.* As mentioned above, it is proposed a partition of the learning space into groups relating certain semantic profile. Thus, when new data is added to the system only part of the training set is involved in the learning update, which translates in high scalability. It facilitates long-term learning, in which concepts are continually learned and refined over time, not necessarily from the interaction with one single user in a single session. Long-term learning contributes to improve the generalization capabilities.

1.5 Structure of the Document

The remainder of this document is organised as follows:

Chapter 2 contains a formal description of the image classification and annotation processes. A concise summary of the state-of-the-art along with related representative research work completed in the context of content-based image retrieval is presented.

Chapter 3 introduces the Multimedia Content Description Interface, MPEG-7, and the visual descriptors used to build the feature space.

Chapter 4 presents the proposed partially supervised fuzzy clustering algorithm. It also describes how structural data analysis uses prior domain knowledge to improve the learning space partition.

Chapter 5 describes two case studies preceding the framework design. The first case study relates to the usage of a fuzzy inference approach for building image classification. The second one implements radial basis function network (RBFN)-based approach to perform two-class and multi-class image classification.

Chapter 6 focuses on the design of the framework for concept-related indexing of image content. It describes the main components of the framework. A proposed clustering approach along with details regarding semantic profile management is presented. Afterwards, an evaluation of the framework is reported.

Chapter 7 presents an overall discussion followed by concluding remarks and introduction of further work.

Additionally, complementary explanations are organised into three appendixes.

Chapter 2

Conceptual Image Analysis

Content-based image analysis is a research tool used to determine the presence of certain concepts within an image or sets of images. *Concepts* are abstract or general ideas inferred or derived from interpretations of image content. Therefore, recognition and interpretation of image content require high-level symbolic processing [37][38].

An interpreter provides concepts, based on his knowledge and experience, after extracting qualitative and quantitative information from the image content. Consequently, the recognition problem is considered a supervised learning problem [39].

Content-based image analysis is quite often been thought in terms of *conceptual analysis*, in which a concept is chosen for determining its presence or absence. In other words, conceptual analysis implies a categorization process.

Actually, recent studies carried out by Grill-Spector and Kanwisher [40], who study how humans perceive images, show that conceptual image analysis is seemingly connected to image categorization. Some of their results suggest that: "detection does not occur prior to and independently of categorization."

Following section introduces conceptual analysis along with some useful definitions presented in [41]. Next, Sect. 2.2 describes the correspondence between conceptual analysis and multi-class classification. In Sect. 2.3 a concise summary of the state-of-the-art along with related work is presented. Sect. 2.4 describes the advances of this research with respect of the current challenges in the field. Remarks concerning potential directions for future work are presented at the end.

2.1 Image Abstraction and Interpretation

In the context of this work, the term *image* refers specifically to still pictures represented and displayed in a digital format by subdividing it into small equal-sized and shaped areas, called picture elements or pixels. Each pixel represents the brightness of certain area with a numeric value or digital number.

Image digitisation is carried out through a sampling process that divides an analogue image in a 2-D continuous space into an $N \times M$ image in a 2-D discrete space. An intersection of a row $(1, \ldots, N)$ and a column $(1, \ldots, M)$ corresponds to a pixel.

Visual primitives are extracted automatically from digital images using some measurements at pixel level. Natural (luminance, texture, etc.) and artificial (colour histograms, etc.) characteristics are combined to generate multi-modal distinguishing primitives called *image abstractions.*

Multi-modal visual primitives with a well-defined syntax and semantics are known as *Descriptors.* In order to simplify management of image abstractions, descriptors are organized into *feature vectors.*

On a higher level of abstraction, a human interpreter who examines digital imagery displayed on a computer screen provides *visual interpretations* of image content (also referred to as *image concepts*).

Interpreters make use of granules of information connected at different levels of abstraction as illustrates Fig. 2.1.

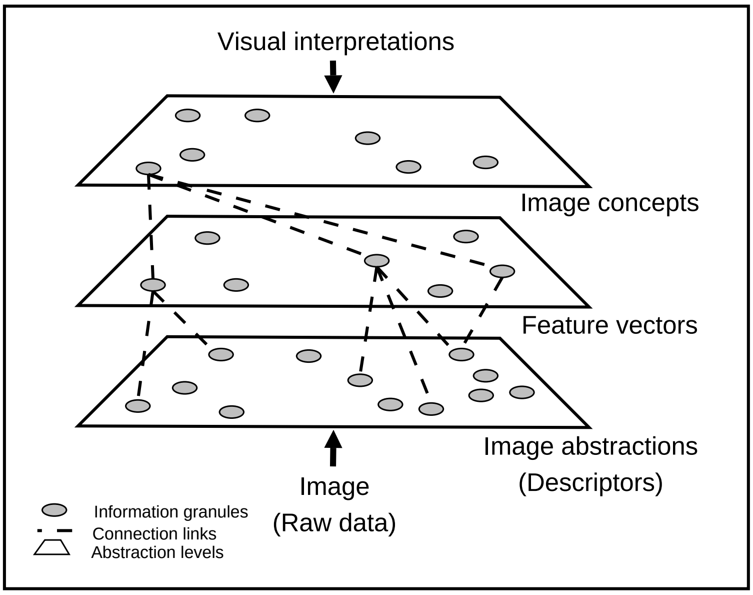
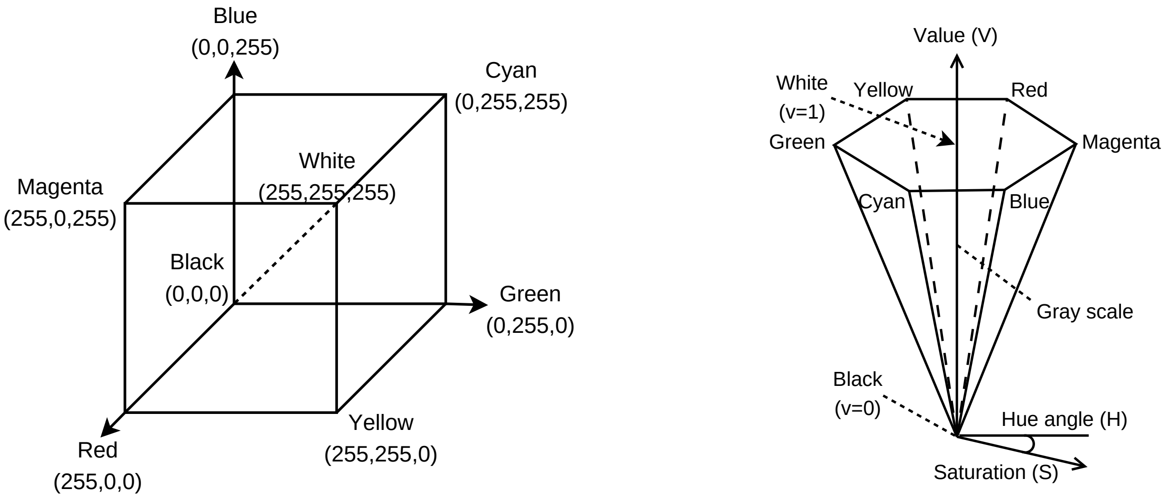


Fig. 2.1: Granules of information at different abstraction levels. *Raw data* consists of elementary units together with some general attributes such as encoding/decoding format. Primitives such as tone, shapes, and texture characterize *Image abstractions. Image concepts* connect visual interpretations to feature vectors regarding to objects or events within the scene displayed in the image

Pictures can be displayed as monochrome (black and white) images, or as colour images by combining bands representing different wavelengths. Colour images combine and display information digitally using different channels on a given colour space (e.g. RGB, HSV, YCrCb). All colour spaces are three-dimensional coordinate systems. In the case of RGB colour space, data from each channel is represented in one of the primary colours (red, green, and blue). Depending on the relative brightness of each pixel in each channel, the primary colours (or their equivalent values in another colour space) are combined in different proportions to represent a different colour. Fig. 2.2 depicts the RGB and HSV colour spaces.



(a) RGB space

(b) HSV space

Fig. 2.2: Three-dimensional coordinate RGB and HSV systems

Visual interpretation is physically limited to spatial, spectral, and radiometric resolutions. Spatial resolution is determined by the size of the smallest possible object that can be detected. Spectral resolution describes the ability of a sensor to define fine wavelength intervals. The radiometric resolution of an imaging system describes its ability to discriminate very slight differences in energy.

Besides, visual interpretation is constrained by the closeness of the colour spaces to the human system vision and the amount of images to be analysed at a time.

A shortcoming of visual interpretation is its subjectiveness, meaning that the results will vary with different beholders.

In spite of these limitations, information captured from visual interpretations is required to enhance image descriptions as a prelude to automatic classification, which enables conceptual analysis of large number of images.

Automated classification carried out by manipulating image descriptions in a computer produces more objective and generally more consistent results. However, there is a shortcoming in determining the validity and accuracy of the results. Consequently, automated classification rarely is performed as a complete replacement for manual (visual) interpretation. Instead it is used to supplement and assist the conceptual image analysis.

2.2 Classification and Conceptual Analysis

Conceptual image analysis is often approached by computing low-level features, which are processed with a classifier engine for inferring high-level information about the image.

This kind of high-level (or semantic) classification relates the problem of assigning an image to one or more specified categories based on interpretations of its content.

In designing the classifier, the attention is focused on using aggregation of classification outcomes to perform concept-related indexing of images as a prior step in the road towards semantic image annotation.

2.2.1 Semantic Classification

The classification task is addressed to grouping image features into broad classes such as city views, nature, etc. Each class is linked to an image concept. Other more specific classes, such as different type of buildings, may not be easily distinguishable. In designing a semantic classifier, problem domain knowledge is provided by an expert user. In the case of approaching image annotation by classification, the expert user would be a *professional annotator*. Formally, the classification problem is defined as follows:

Let $\mathbf{i} = (i_{11}, i_{21}, \ldots, i_{M1}, i_{12}, \ldots, i_{MN})$ be an image, $D = \{\mathbf{d}_1, \ldots, \mathbf{d}_{N_D}\}$ be a set of image descriptions extracted from $\mathbf{i}$, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_{N_X}\}$ be a set of feature vectors built with data obtained from D, and $\Omega = \{\omega_1, \ldots, \omega_{N_\Omega}\}$ be a set of classes. The multi-class classification problem is stated as: Learn a function

$$\mathrm{f} : X \mapsto \Omega \ , \tag{2.1}$$

where classes in Ω relate visual interpretations about image content, which is represented by $\mathbf{x}_i$. f can be decomposed into a number of binary (two-class) classifiers

$$\mathrm{f} \ = \ \bigvee_{1 \leq k \leq N_\Omega} \mathrm{f}_k \tag{2.2}$$

$$\mathrm{f}_k \ : \ \Re^n \mapsto \{0, 1\} \tag{2.3}$$

$$\mathrm{f}_k(\mathbf{x}_i) \ = \ \begin{cases} 1, & \mathbf{x}_i \in \omega_k \\ 0, & \mathbf{x}_i \in \mathrm{not}(\omega_k) \end{cases} \tag{2.4}$$

Subsequently, classification outcomes are aggregated as follows:

$$\mathcal{L} = \bigcup_{\omega_k \multimap \ell_j} \ell_j \ , \tag{2.5}$$

where ℓ_j is a concept-related symbol (label) associated with class ω_k.

A general semantic image classification process is depicted in Fig. 2.3. As indicated above, its input is a set of features extracted from an image; its output is a discrete label for the concept describing some visual interpretation of the image content. The classification module aggregates a number of binary (concept-related) classifiers.

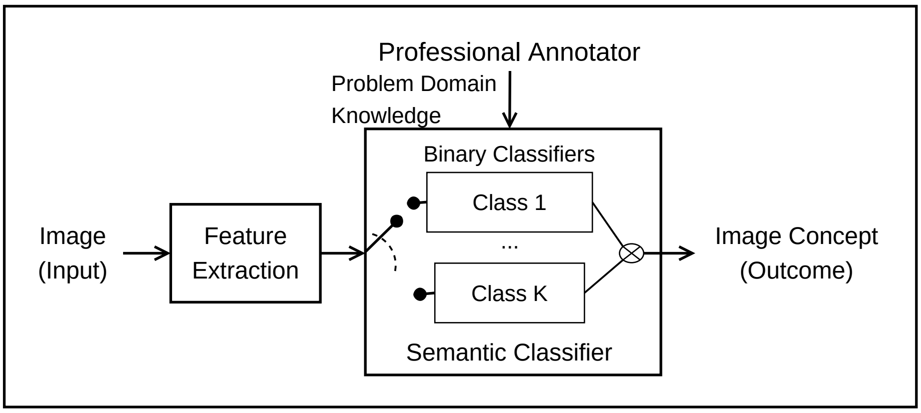
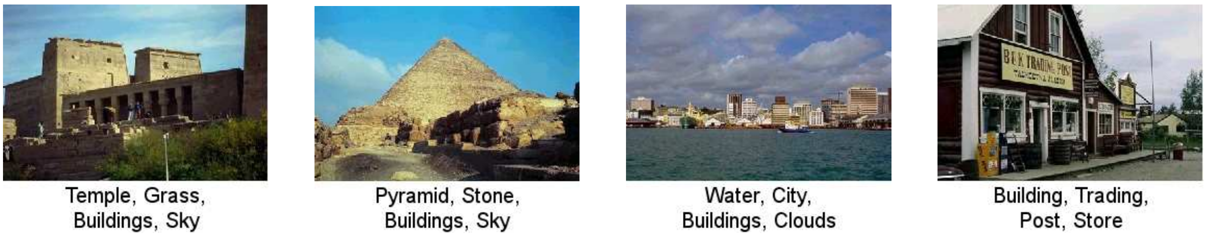


Fig. 2.3: Overview of a general semantic image classification process

Semantic classification deals with challenging problems at the learning stage such as:

- *Specific vs general labels.* General labels (hypernyms) are more appropriate for training purposes. Otherwise, it will be too difficult to generalize beyond the training set. Fig. 2.4 shows an image labelled as “Temple”, which is too specific. It could be linked to a broader class using the label “Buildings” appearing in third position within the same list.

- *Training images selection.* Exemplars are present to the classifier during learning process as a set of input-output pairs or images previously labelled. In order to avoid ambiguity that furthers misclassification, it is required to provide training images labelled appropriately regarding to the concept presented to the learner. Fig. 2.4 shows samples of pictures that can mislead the classifier if they are presented to the learner as exemplars of “building” images.



Temple, Grass, Buildings, Sky

Pyramid, Stone, Buildings, Sky

Water, City, Buildings, Clouds

Building, Trading, Post, Store

Fig. 2.4: Samples of variety of labels. Presenting these pictures to the learner as exemplars of “building” images can mislead the classifier

The applied concept-related classification is useful to describe images with a controlled vocabulary. However, it is impractical for problem domains involving large number of concepts. An alternative is to use classification outcomes as entries to a more elaborated interpretation environment, as the one provided by semantic ontology models (Mezaris et al. [42]).

Bearing this in mind, classification outcomes are further used to create a set of concept-related indexes of the corresponding image. Concept-related indexes can be used as pre-annotation sets as well as mechanisms to access the branches in an ontology map. At the end, the system will come out with a set of labels to annotate the image.

2.2.2 The Lexicon

Concept-related indexes facilitate content-based image retrieval using symbols relating visual interpretations that is what concepts are about. As mentioned above, indexes are treated as entry points to a more elaborated representation of the problem domain knowledge, i.e. a computational ontology.

Computational ontology differs from philosophical ontology as it refers to a rooted directed acyclic graph (the data structure) in which every node stands for a concept and every arc connects nodes by IS-A links. For the sake of clarification, Fig. 2.5 illustrates the usage of concept-related indexes as entries to a computational ontology. Concept *building* serves to go down and reach specific instances such as *apartment* or to go up and retrieve general instances such as *structure*.

```
index-name:  building

index-definition:  a structure that has a roof and walls
and stands more or less permanently in one place

index-entry to sub-graph:
entity
↪ object; physical_object
 ↪ artifact; artefact
  ↪ structure; construction
   ↪ building, edifice
    ↪ clubhouse; club
    ↪ chapterhouse
    ↪ center; centre
    ↪ bowling_alley
    ↪ bathhouse; bathing_machine
    ↪ aviary; bird_sanctuary; volary
    ↪ architecture
    ↪ apartment_building; apartment_house
    ↪ abattoir; butchery; shambles; slaughterhouse
```

Fig. 2.5: Index relating concept *building* and its usage as entry to a computational ontology containing such a concept [wordreference.com]

Supporting of computational ontology is beyond the scope of this research work. Eventhough, it serves to show implications on choosing certain concepts to represent visual interpretations.

Chosen concepts are organized into a *lexicon.* It is proposed the usage of a constrained lexicon. Semantic constraint refers to substitution of specific concepts, i.e. meronyms and hyponyms, by more general ones, i.e. holonyms or hypernyms. In addition, all synonyms have to be replaced by a common concept. Experiments show that constraining the lexicon the performance is increased significatively.

For instance, professional annotations of Corel images as portrayed in Fig. 2.6 involve more information: title, caption, categories, and concepts [ref fotosearch.com/corel]. These annotations can be simplified substituting them by some holonyms.

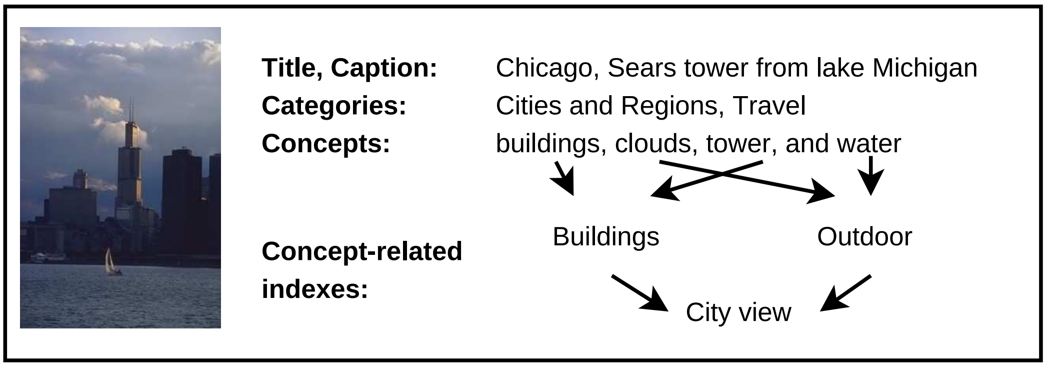

Title, Caption:
Chicago, Sears tower from lake Michigan
Categories:
Cities and Regions, Travel
Concepts:
buildings, clouds, tower, and water
Concept-related indexes:
Buildings
Outdoor
City view

Fig. 2.6: Sample of Corel annotations consisting of title, caption, categories, and concepts. Specific concepts are simplified using concept-related indexes

2.2.3 Semantic Annotation

The term *annotation* refers to the use of auxiliary symbols that are utilised to modify the interpretation of other symbols. These annotation symbols typically do not have the same kind of meaning as the symbols that they annotate [43].

On the other hand, *semantic annotation* is augmentation of data to facilitate automatic recognition of the underlying semantic structure of image content. This kind of annotation is expressed by means of symbol's structures, where each symbol is either a keyword or an icon.

In the context of conceptual image analysis, symbols are basic semantic units representing image concepts. Ambiguous interpretations are avoided using a symbol domain, which is organised into a controlled lexicon.

Based on the definition provided by Intille and Bobick [44], *image annotation* is the task of generating descriptions of still pictures that can be used for indexing and retrieval. Such descriptions are aimed to associate semantic meaning with image to improve the content-based retrieval.

For their generation, image abstractions are processed and linked to image concepts in a complex interplay among the expert, the images, and their visual interpretations [12] (See Fig. 2.1).

Image annotation requires uniform models for representing image content and facilitating the use to several users. Therefore, interoperability of image descriptions as well as durable and stable lexicons is a requirement to support sharing and reusing annotations among users.

In order to facilitate the description of the process implementing the annotation task, the notation proposed by Smeulders et al. [12] is used.

The image annotation process denoted by $\mathcal{A}$ can be summarised as

$$\{\mathcal{S}_{(0)}, \mathbf{i}, L\}\mathcal{A}\{\mathcal{S}_{(t)}, \mathcal{L}\} \tag{2.6}$$

where $\mathcal{S}_{(0)}$ is an abstract non-annotated image space defined as a tuple $\{X, L\}$, where X is a set of images features, L is a controlled lexicon, and $\mathbf{i}$ is a selected image to be annotated.

The process $\mathcal{A}$ begins creating an instance of the abstract non annotated image space

$$\mathcal{S}_{(0)} = \emptyset \ . \tag{2.7}$$

Afterwards, the annotation process $\mathcal{A}$ maps $\mathcal{S}_{(0)}$ into the annotated image space in an interactive annotation session, which is a sequence of annotation spaces

$$\mathcal{S}_{(0)} \circlearrowleft \mathcal{S}_{(1)} \circlearrowleft, \ldots, \circlearrowleft \mathcal{S}_{(t)} \tag{2.8}$$

with

$$\mathcal{S}_{(t)} \doteq \mathcal{A}(\mathcal{S}_{(t-1)}) \ . \tag{2.9}$$

The result is an annotation set $\mathcal{L} \in \mathcal{S}_{(t)}$ such that

$$\mathcal{L} = \{\ell_j \mid \ell_j \in L; 1 \leq j \leq N_{\mathcal{L}}\} \ . \tag{2.10}$$

Fig. 2.7 summarises the semantic image annotation process.

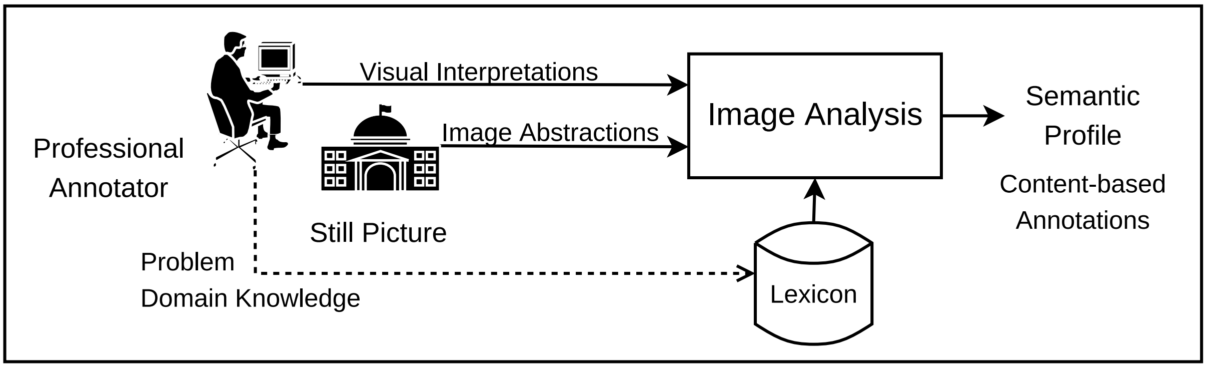


Fig. 2.7: Overview of the semantic image annotation process. The inputs are image abstractions and visual interpretations that provide the image analysis process with numerical representations and semantic concepts, respectively. The lexicon constrains set of concepts to be used during the analysis to the problem domain knowledge. The outcome is a semantic profile

This research work is based upon the hypothesis of augmentation of data extracted from image abstractions, i.e. by learning and then adding image concepts through classification, can be used as a near-automatic way to annotate images in CBIR systems. By comparing Eq. 2.5 and Eq. 2.10 it becomes noticeable that the annotation process can be approached as a multi-class classification problem.

2.3 State-of-the-Art

Many *ad hoc* methods and techniques have been proposed and developed to perform conceptual image analysis. It is evidenced in the vast amount of publications found in journals, conference proceedings, and books that are dedicated to its study. As a result, the last decades witnessed a genuine plethora of approaches to image description, formalization, and classification (Rosenfeld [53], Smeulders et al. [12], Eakins [46]). This tendency is not surprising at all given a rapidly growing interest in processes and efficient mechanisms of image annotation, categorization, and organization as well as retrieval.

This section outlines the steps in the semantic annotation process as portrayed in Fig. 2.7. It summarizes the state-of-the-art with respect to each step.

2.3.1 Image Abstractions

Image abstractions are extracted automatically from digital images using some measurements at pixel level. However, there is a recognized limitation in the usage of pixel-based representations of images to perform content-based analysis. In an attempt to resemble the way humans detect salient features from image content, extraction techniques have been oriented towards region-of-interest (ROI) approaches.

Barnard et al. [54] used an analogy between images and text to perform classification based on semantic units. Images are composed of regions and objects, and text is composed of words, or more abstractly, topics or concepts. Duygulu et al. [55] considered the problem of finding such correspondences as the translations of image regions to words, similar to the translation of text from one language to another. They investigated the effect of feature sets on the performance of linking image regions to words.

Paterno et al. [56] proposed an approach to semantic labelling of image regions. It is based on a fuzzy labelling method that measures the confidence based on the orthogonal distance of an image region's feature vector to the hyperplane constructed by support vector machines. The confidence measures assigned to a region represent the signature of the region and are used for region matching during image retrieval.

Though it is observed a tendency on applying image segmentation for a more accurate recognition of image content, there are also a steamed number of approaches tackling the problem of feature extraction without partitioning or segmenting the image. Furthermore, images are designed to convey a certain message, but this message is concealed in the whole organization of the image, and is not always possible to divide it syntactically into smaller parts.

Lipson et al. [17] proposed an approach to classify scenes without first attempting to recognize their components. The strategy is well suited for complex scenes, especially those that consist mostly of natural objects. Gu´erin-Dugu´e and Oliva [57] obtained results supported by psychological experiments showing that human subjects rapidly capture the context of the scene before recognizing its individual parts. Similarly, Iqbal and Aggarwal [19] and Santini et al. [18] presented approaches at which segmentation and detailed object representation are not required.

2.3.2 Capturing Knowledge and Interpretations

Visual interpretations of the image content provide lexical information that attached to image abstractions expand classic conceptual image analysis. It is referred to as semantic annotation. Lexical information is useful to introduce problem domain knowledge. It consists of learnable and non-learnable concepts.

A learnable concept is defined as a short program (finite state automata) that distinguishes some natural inputs from some others. According to Valiant's model, a concept has been learned if a program for recognizing it has been deduced, i.e. by some method other than the acquisition from the outside of the explicit learning program (Valiant [58]).

Shvaytser [59] pointed out that a learnable concept should be learned from a polynomial number of examples and using polynomial bounded computational resources.

Systems that learn and adapt learnable concepts represent one of the most important trends in computer vision research and may provide the only solution to the development of robust and reusable systems (Heisele et al. [39]).

In this regard, Saitta and Bergadano [60] presented an interesting comparative analysis of results from pattern recognition and theoretical machine learning regarding to the problem of learning concepts. Special attention is devoted to the learning framework proposed earlier by Valiant [58].

Qiu et al. [61] used machine learning to learn high-level concepts from examples of similar images chosen by humans. It showed how different representation schemes may affect many aspects of a machine learning system, including computational complexity and performance.

Santini et al. [18] proposed that the meaning of an image is characterized by the following properties: It is contextual, it is differential, and it is grounded in action. Subsequently, they argued that images do not have an intrinsic meaning, but that they are endowed with a meaning by placing them in the context of other images and by the user interaction. From this observation, they conclude that in an image database users should be allowed to manipulate not only the individual images, but also the relation between them. It suggests working with group of images rather than one image at time.

Nakazato and Huang [62] used ImageGrouper, a group-oriented interface for digital image search and organization, to support incremental image annotation applying relevance feedback. Bhanu and Dong [63] proposed a framework for learning concepts based on retrieval experience, which combines partially supervised clustering and probabilistic relevance feedback. Yu et al. [64] examined how prior domain knowledge might be incorporated into concept learning when only limited sample data is available.

There are also several interactive approaches that have been proposed to enable long-term learning (Yoshizawa and Schweitzer [58]) and system's adaptation (Rui et al. [66], Bhanu and Dong [63]).

2.3.3 Image Analysis Techniques

Gorkani and Picard [67] considered using global texture properties of images as a quick way to make a first pass at higher-level problem; such as annotating or retrieving a particular set of digitised vacation photos. The work has been used as reference for a large number of researches on scene classification. Szummer and Picard [68] demonstrated an improvement in classification performance when computing features on sub-blocks, classifying the sub-blocks, and then combining the results.

Vailaya et al. [69] used an approach to determine the discriminative power of features for further classification of images based on inter-class and intra-class distance distributions. Vailaya et al. [69] applied a robust scheme to identify salient image features and capturing a certain aspect of the semantic content. They classified vacation photographs into a hierarchy of high-level classes. Results show how specific low-level features can be used in constrained environments to perform conceptual image analysis.

Loui and Savakis [71][72] dedicated their research work to automatic classification of events in general consumer pictures. The difficulty of the task is increased when considering limited or no contextual information about the picture content.

In [73], Luo et al. proposed an architecture model based on belief networks that is suitable for semantic understanding of consumer pictures. Salient features of this kind of pictures are: they have unconstrained picture content and are taken under unconstrained imaging conditions; they serve the purpose of recording and communicating memories and therefore is a certain degree of agreement among observers in spite of the inherent ambiguity in terms of subjective opinion. Luo and Savakis [74] also used semantic cues (e.g. sky, grass) for improving the classification performance obtained by low-level features alone.

Wang et al. [75] presented an approach to find images containing buildings. Global and local distribution of directional edges along texture information is used to represent the image semantics.

Boutell et al. [76] introduced spatial image recomposition and simulated temporal image recomposition. The first one is designed to minimize the impact of undesirable composition produced by foreground objects. The latter is designed to minimize the effect of colour changes occurring over time. Image recomposition is applied on classification of sunset scenes.

Mojsilovic and Rogowitz [87] used a method for semantic categorization of photographic images, based on low-level image descriptors derived from perceptual experiments. The method is proposed as groundwork for better organization of (and retrieval from) large image databases.

Special attention has been paid to Support Vector Machines (SVMs) classifiers.

Barla et al. [88] investigated the potential of SVMs to solve semantic image classification problems. SVMs classifiers learn from a relatively small number of samples (Jain et al. [89]). The class (or label) of a new input is determined by a linear combination of the kernel functions evaluated on a certain subset of the examples –the support vectors and the input. The coefficients of the combination are obtained as the solution to a convex optimisation problem corresponding to the learning stage. The choice of the kernel relies on prior knowledge of the problem domain.

There are two important issues for developing SVMs classifiers. One is how to extend two-class to multi-class problems efficiently. Several methods have been proposed, such as one-against-one, one-against-all. The other issue of SVMs classifiers is to overcome the problem of overfitting in two-class classification. Huang and Liu [98] proposed the use of fuzzy support vector machines to deal with outliers or noise because of overfitting problem.

Chapelle et al. [91] presented an approach to overcome poorly generalization on image classification tasks, because of the high dimensionality of the feature space. They used a SVMs approach because its high generalization performance even when the dimension of the input space is very high.

Brunelli and Mich [92] introduced a statistical framework to analyse effectiveness of histograms in image comparison tasks. Results showed that it is possible to minimize the size of the image descriptors while maintaining good effectiveness.

Serrano et al. [93] presented an improved SVM-based approach to indoor/outdoor classification. The approach uses a low dimensional feature set in which a low computation complexity is achieved without compromising the accuracy.

Yan et al. [84] applied support vector machine ensembles to adapt binary SVMs to multi-class classification and address the high computational cost for training.

Tsai et al. [85] implemented concept-based indexing by combining SOMs and SVMs in a two-stage hybrid classifier. Prabhakar et al. [96] proposed another hybrid approach. The classification of images as either picture or graphics is performed by a combination of a rule-based tree classifier and a neural network classifier.

Dong and Yang [80] combined SVMs and kNN for hierarchical classification of web images. A threshold strategy called “HRCut” is proposed because of the difficulty in applying traditional strategies (e.g. rank-based threshold) to decide the winner category for each image.

M. R. Naphade and Basu [81] used SVMs and active learning in a very similar way than Zhang and Chen [107]. Essentially it is an extension of the method proposed by Tong and Chang [108]. Active learning enables the learner to use its own ability to respond to collect data and to influence the world it is trying to understand.

2.3.4 CBIR Systems

Several systems implementing similar or closely related processes to the one illustrated in Fig. 2.7 have been implemented.

Wang et al. [109] presented SIMPLIcity (Semantics-sensitive Integrated Matching for Picture LIbraries), an image retrieval system, which uses semantics classification methods, a wavelet-based approach for feature extraction, and integrated region matching based upon image segmentation.

Zhang and Chen [107][110][111] proposed a framework for active learning during annotation process in CBIR systems. It includes three major aspects: feature extraction, high dimen-sional indexing and system design. Among the three aspects, high dimensional indexing is important for speed performance; system design is critical for appearance performance; and feature extraction is the key to accuracy performance.

Laaksonen et al. [112][113] used PicSOM, a neural-network-based CBIR system built upon the query by pictorial example and relevance feedback principles, to combine high-level semantic concepts and the low-level features. SOM stands for Self-organizing map, which is used for unsupervised, self-organizing, and topology-preserving mapping from the image descriptor space to a 2-D lattice of artificial neural units.

Sheilholeslami et al. [114] utilized SemQuery, a hierarchy of semantic clusters, to support visual queries based on heterogeneous features. A multi-layer neural network model is used to merge the results of basic queries on individual features. The input to the neural network is the set of similarity measurements for different feature classes and the output is the overall similarity of the image.

Wang et al. [20] employed ALIP, a 2-D multi-resolution hidden Markov model, to categorize pictures. In ALIP only the statistical models for the involved concepts need to be trained or retrained.

Wang et al. [115] used SemView (a semantic-sensitive distributed image retrieval system) to support low- and high-level retrieval on distributed image databases. The system provides the users with the facility to manage large image databases in an automated, flexible, and efficient way.

Smith et al. [116] developed an approach for integrating features, models, and seman-tics in a CBIR system using MPEG-7 descriptors. Currently, this kind of approaches has been used in developing a multimedia analysis and retrieval systems such as MARVEL (ref [http://www.research.ibm.com/marvel/]).

2.3.5 Summary of Selected Experimental Studies

Tab. 2.1 summarizes experimental studies of selected references cited above. This overview is instructive in the sense it underlines the diversity of approaches taken, provides insight into a quality of classification given some collection of categories (classes) under consideration. It also shows the diversity of sources of images used in these experiments (which unfortunately does not allow to compare obtained classification rate however gives some qualitative sense as to the possible outcomes).

While the differences are quite significant, there are some commonalities when it comes to the most fundamental paradigm of classification. Classification errors are very different however any close comparison is not feasible given different semantic categories handled by each system. It is noticeable that in many cases there is some pre-processing aimed at dimensionality reduction.

2.4 On the Road to Semantic Annotation

Contributions listed above show how robust and efficient pattern recognition techniques enable automatic interpretation of image content. The primary goal of pattern recognition is supervised or unsupervised classification.

There is a special interest to evaluate intermediate methods, i.e. semi-supervised learning, to incorporate knowledge through labelled data and make use of unlabeled data as well (Pedrycz [118]). Jain et al. [97] in their detailed summary of well-known pattern recognition methods posed unsupervised classification as a clustering/categorization task based on the similarity of patterns. Semi-supervised clustering exploits problem domain knowledge to build classifier models that group patterns into groups with some semantic meaning.

The proposed framework applies unsupervised cluster in a first stage to reveal any underlying structure in the data sets without requiring prior information. Exploiting prior domain knowledge, a structural data analysis is proposed to determine proximity and overlapping between semantic groups, which leads to indexing errors.

Structural data analysis along with prior domain knowledge is used to guide a partially supervised fuzzy partitioning of the learning space. Proposed algorithm based on the work presented by Pedrycz and Waletzky [119] shows promising results in the context of conceptual image analysis.

Clustering and structural data analyses are used as exploratory data analysis to move cluster prototypes towards regions in the learning space containing relevant exemplars.

In addition, a method to determine the semantic profile of each cluster is introduced. The semantic profile facilitates indexing and is appropriated for further retrieval based on concept

Tab. 2.1: Summary of various research pursuits carried out in image classification objectives, categories of interest, pre-processing, and feature space as well as classification results

Ref	Objective	Classes	Image set	Preprocessing	Feature space	Classification results
[96]	Scene classification using support vector machines (SVMs)	Indoor/Outdoor	600 images mainly from the Web	Texture represented by gray-level co-occurrence matrices	Colour histograms and Texture	Accuracy of 93.1%
[117]	Multi-label image classification	Beach, Mountain, Sunset, Field, Fall foliage, etc.	2,400 photos from Corel and personal collection	Images divided into 49 blocks using a 7x7 grid.	Spatial colour moments	Accuracies: Single-label 79.5% Multi-label 81.8%
[91]	Histogram-based image classification using SVMs	Airplanes, Birds, Boats, Buildings, People, Fish, and Vehicles	1,400 from Corel7 and 2,670 from Corel14 database		Colour histograms	Classification error of 11% on Corel categories and 16% on more generic objects
[67]	Image classification for annotation or retrieval	Landscape City/Suburb	98 digitized photos	Images divided into 16 equal-size rectangular regions	Global texture features	Accuracy above 90%
[74]	Classification via integrating lowand mid-level features	Indoor/Outdoor	1,300 images		Colour and texture; sky and grass features	Accuracy using colour, texture, sky, and grass equal to 90.1%
[87]	Capture the semantic meaning of an image	People, Animals, Buildings, Nature, etc.	393 images for training and testing	Multidimensional scaling and hierarchical clustering	40 image-processing features	93% of accuracy achieved on the testing set
[93]	Computational efficient image classification using SVMs	Indoor/Outdoor	1,200 consumer photographs	Colour quantization, colour balance, subsampling and 4x4 tessellation	Colour in the LST color space Wavelet texture features	90.2% of accuracy
[68]	Image classification	Indoor/Outdoor	1,343 digitized photos	Region-based partition and multistage classification	Colour and global textural features	Accuracy above 90.3%
[85]	Image classification using hybrid neural networks	25 categories: Animal, Stone, Building, Snow, etc.	10,000 from Corel stock collection	Sub-sampling, Quadtree decomposition, Self-Organizing Maps (SOMs)	Colour histograms Textural wavelet decomposition	Highest accuracy is 90% in three categories
[63]	Hierarchical image classification for content-based indexing	Indoor/Outdoor City/Landscape Sunset/Forest /Mountain	6,931 images various sources	Multidimensional scaling to generate 3D feature space K-Means to identify group of images	Colour moments, Edge direction, Coherence vectors, Spatial moments	Accuracies of 90.5%, 95.3%, and 96.6% for the corresponding classification tasks
[69]	Semantic-based classification	City/Landscape	2,716 photos various sources	Features are extracted from the entire image or image subsections	Edge direction (and colour), coherence vector	Accuracy above 93.9%
[75]	Building image classification	Building/Not-Building	1,791 images from Corel	Image partition into 16 equal-size regions	Edge histogram, Gabor texture	90.6% building and 74.8% not-building

matching.

In summary, the proposed framework applies a consistent learning strategy that exploits prior knowledge and data information. It embodies a partially supervised clustering algorithm that overcomes many of the drawbacks found in applying clustering as a classification method. A learning space equipped with cluster prototypes and semantic profiles, boosts the indexing procedures of learnable concepts.

Chapter 3

An MPEG-7 Learning Space

In designing the framework for concept-related indexing of image content, the interest turned out to the use of MPEG-7 descriptors for defining a learning space. Specifically, there is a concern on the structure of the feature vectors. MPEG-7 descriptors are equipped with a well-defined syntax and semantics that facilitate representation of image abstractions. Therefore, proposed structure, in which descriptor elements will be aggregated, should be defined in a comprehensive manner and considering important issues such as vector dimensionality.

The problem of combining descriptions has been tackled using either multi-feature vec-tors or multiple feature spaces [120][121]. The approaches using multi-feature vectors break down into high-dimensional space, which drastically reduce efficiency of storage and retrieval. Harsanyi and Chang [122] introduced an unsupervised feature extraction technique applying an orthogonal subspace projection that reduces the data dimensionality. Kokare et al. [123] used tree structured wavelet decomposition for dimensionality reduction of texture features. The work presented by Wu et al. [124] reduces dimensionality preserving the local topology of the original space. Dorado and Izquierdo [125] proposed an adaptive clustering method to select representative features.

On the other hand, approaches apply on multiple feature spaces require an effective method to integrate the matching results from each space. Smith et al. [116] detailed a completed taxonomy of the different searching and matching problems on multi-feature spaces. Furthermore, Laaksonen et al. [112] proposed a framework that integrates a number of parallel tree structured self-organising maps.

The rest of this chapter is organized as follows: Sect. 3.1 introduces the Multimedia Content Description Interface, MPEG-7. Sect. 3.2 summarises some considerations in the construction of the feature space and details the components of feature vectors using the proposed structure. Then, an empirical evaluation of feature vector quality is carried out in Sect. 3.3. Sect. 3.4 summarizes the chapter.

3.1 MPEG-7

MPEG-7, formally known as Multimedia Content Description Interface is an ISO/IEC standard developed by the Moving Picture Experts Group (MPEG) for description and search of audio and video content [ref www.chiariglione.org/mpeg/]. In contrast with the early standards known as MPEG-1, MPEG-2, and MPEG-4 that are focused on coding and representation of audio-visual content, MPEG-7 moves forward and becomes more general by embracing description of multimedia content [126].

MPEG-7 has emerged as a cornerstone of the development of a wide spectrum of applications dealing with audio, speech, video, still pictures, graphics, 3D models, and alike. In a nutshell, the MPEG-7 environment delivers a comprehensible metadata description standard that is interoperable with other leading standards such as SMPTE Metadata Dictionary, Dublin Core, EBU P/Meta, and TV Anytime; refer to www.ebu.ch/trev_284-mulder.pdf. Initially, MPEG-7 was focused more on web-based applications and annotation tools (e.g. Izquierdo et al. [127], Mezaris et al. [42], Smith and Lugeon [128]). Nowadays, it is being drifted to other domains such as education, video surveillance, entertainment, medicine and biomedicine.

The ultimate objective of MPEG-7 is to provide interoperability among systems and applications used in generation, management, distribution, and consumption of audio-visual content descriptions (see Fig. 3.1). Such descriptions of streamed (live) or stored on various media help either users or applications in identifying, retrieving, or filtering essential audio-visual information, cf. [126][129].

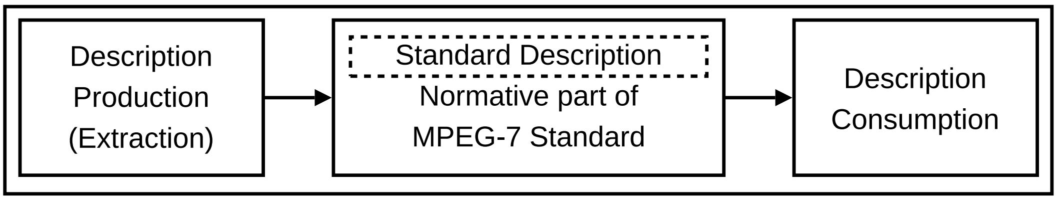

Description Production (Extraction)
Standard Description
Normative part of MPEG-7 Standard
Description Consumption

Fig. 3.1: Scope of MPEG-7

MPEG-7 specifies standardised Descriptors and Description Schemes for audio and video, as well as integrated multimedia content. Also standardised is a Description Definition Language that allows new Descriptors and Description Schemes to be defined.

MPEG-7 descriptors define syntax and semantics of features of audio-visual content. MPEG-7 allows these descriptions at different perceptual and semantic levels. At the lowest abstraction level, such descriptors may include shape, texture, and colour. At the highest abstraction level, they may include events, abstract concepts, and so forth. Tab. 3.1 presents a list of the basic MPEG-7 visual descriptors applicable to still images.

During the collaborative phase of the MPEG-7 standard, descriptors were incorporated into a common model called the eXperimentation Model or XM, which constitutes a draft of

Tab. 3.1: List of basic MPEG-7 visual descriptors used to characterize still images [1]

Colour descriptors	**Texture descriptors**	**Shape descriptors**
`Colour Layout`	`Texture Browsing`	`Region-based Shape`
`Colour Structure`	`Homogeneous Texture`	`Contour-based Shape`
`Dominant Colour`	`Edge Histogram`	
`Scalable Colour`		

the standard itself [ref www.chiariglione.org/mpeg/]. Test conditions and criteria to assess those descriptors were well defined. In the case of visual descriptors, tests were oriented to evaluate retrieval performance. It has motivated complementary studies to analyse aspects such as data quality of the extracted descriptions (Eidenberger [131][132]) or computational complexity (Ojala et al. [133][134]). Stanchev et al. [135] also evaluated the effectiveness of MPEG-7 image features on specific image data sets.

3.2 Feature Space

Proposed feature space is built using a number of selected histogram-based MPEG-7 colour and texture descriptors [130]. Treating such histograms as feature vectors

$$\mathbf{x} = [x_1, x_2, \ldots, x_p]^T \quad , \tag{3.1}$$

where $x_i \in \Re;\ 1 \leq i \leq p$, they can be organized into a feature space

$$X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{N_X}\} \ , \tag{3.2}$$

considering contributions of each descriptor element regarding to its semantics and capabilities to represent the image content. It is worth stressing that the problem of feature selection is fundamental to pattern recognition [22].

There is a vast array of methods including such prominent approaches as principal component analysis (PCA), independent component analysis (ICA) and alike [136] commonly used to select features and reduce dimensionality of feature space. A drawback of those methods in working with MPEG-7 descriptors is that they do not relate directly the original topology and consequently the interpretation of transformed features becomes less intuitive.

It is proposed a feature vector structure that combines efficiently features extracted from MPEG-7 colour and texture descriptors. The structure keeps the original topology to preserve syntax and embedded semantics on these descriptors. It selects descriptor elements avoiding not only description overriding but also controlling vector dimensionality.

3.2.1 Colour and Texture Descriptors

As aforementioned, the feature space used in this study is formed by making use of two types of visual descriptors, namely colour and texture. Then, image content is captured by combining information extracted from local and global distributions (histogram) of colour, edges and texture. For the sake of completeness, next sections include the components of the entire feature vector structure along with a brief description of their syntax and semantics.

Colour Layout Descriptor (CLD)

This descriptor captures the spatial distribution of colour. It applies the discrete cosine transform (DCT) to representative colours in an 8 × 8 grid and encodes the resulting coefficients. A picture is divided into 64 (8 × 8) blocks and their average colours are derived. The colour space adopted for CLD is YCrCb. Fig. 3.2 shows a sample of CLD in XML format.

```
<Descriptor xsi:type="ColorLayoutType">
    <YDCCoeff>6</YDCCoeff>
    <CbDCCoeff>3</CbDCCoeff>
    <CrDCCoeff>3</CbDCCoeff>
    <YACCoeff6>24 18 31 21 16 15</YACCoeff6>
    <CbACCoeff3>32 31 14</CbACCoeff3>
    <CrACCoeff3>18 15 15</CrACCoeff3>
</Descriptor>
```

Fig. 3.2: Sample of colour layout description in XML format

Colour Structure Descriptor (CSD)

This descriptor scans an image using a sliding window approach to capture localized colour distributions. Area of the window is defined by an $n \times n$-pel structuring block, which by default uses $n = 8$. A 1×1-pixel window reduces description to a standard colour histogram.

The histogram summarizes the number of times colours are reported as occurring within the window. A colour map in the Hue-Min-Max-Diff (HMMD) colour space defined by MPEG and a non-uniform quantisation determines the maximum number of bins.

Fig. 3.3(a) presents the mapping from RGB to HMMD space. The Hue channel has the same meaning as in the HSV space. If `Max` == `Min` Hue is undefined (achromatic colour); otherwise it is computed as indicated in Fig. 3.3(b).

Fig. 3.4 shows a sample of CSD in XML format.

Scalable Colour Descriptor (SCD)

This descriptor uses HSV colour space and uniform quantisation to 256 bins (16 levels in H, 4 in S, and 4 in V). In order to lower the 1024 bit/histogram representation, histograms are

```
Max   =  max(R, G, B)
Min   =  min(R, G, B)
Diff  =  Max - Min
```

(a) Min-Max-Diff

```
if (Max == R && G > B) Hue = 60*(G-B)/Diff
else if (Max == R && G < B) Hue = 360 + 60*(G-B)/Diff
else if (G == Max) Hue = 60*(2.0 + (B-R)/Diff)
else Hue = 60*(4.0 + (R-G)/Diff)
```

(b) Hue

Fig. 3.3: Computing HMMD space from RGB values

```
<Descriptor xsi:type="ColorStructureType" colorQuant="4">
   <Values>0 0 0 0 0 0 0 0 33 40 1 21 0 0 0 0 0 2 0 0 0 0 0 0 61 88
           85 60 2 1 5 80 78 41 54 19 32 17 1 92 100 72 61 16 8 6
           42 143 122 58 46 94 100 58 10 124 137 112 75 41 48 39 18
   </Values>
</Descriptor>
```

Fig. 3.4: Sample of colour structure description in XML format

encoded using a Haar Transform. The resulting scaling can be 256, 128, 64, 32, or 16. Fig. 3.5 shows a sample of SCD in XML format.

```
<Descriptor xsi:type="ScalableColorType" NumberOfCoefficients="64"
                  NumberOfBitplanesDiscarded="3">
   <Coefficients>9 0 6 0 -2 0 -2 0 1 2 -3 -1 -4 0 -1 1 0 0 0 1
                  -1 0 0 0 -1 0 -1 0 0 0 0 0 1 0 0 0 0 0 0 0 0
                  1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0
   </Coefficients>
</Descriptor>
```

Fig. 3.5: Sample of scalable colour description in XML format

Edge Histogram Descriptor (EHD)

Firstly, the image is spatially decomposed into 16 sub-images using a fixed grid with equal-size rectangles. Then, five masks are used to assign an edge category to a number of sub-blocks in which the sub-images are divided. It is an aggregation of sixteen 5-bin histograms, which captures the spatial distribution of directional edges within each sub-image. The defined types of edges are: horizontal, vertical, 45°, 135°, and non-directional edges. Fig. 3.6 shows a sample of EHD in XML format.

```
<Descriptor xsi:type="EdgeHistogramType">
   <BinCounts>0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0
               2 6 2 2 1 3 5 3 5 2 3 5 3 4 3 3 5 4 4 4
               5 3 4 5 4 5 2 4 4 1 6 1 4 5 2 4 3 5 5 3
               2 5 3 2 5 2 3 3 4 4 0 3 2 2 2 0 3 2 1 2
   </BinCounts>
</Descriptor>
```

Fig. 3.6: Sample of edge histogram description in XML format

Homogeneous Texture Descriptor (HTD)

This descriptor extracts values from a frequency layout to give a quantitative characterization of the image texture in terms of directionality, coarseness (granularity), and regularity of patterns. It uses a bank of scale and orientation sensitive filters to estimate the energy and the energy deviation. Filters are applied on individual feature channels. Feature channels are filtered applying 2-D-Gabor functions, which are modulated Gaussians. Fig. 3.7 shows a sample of HTD in XML format.

```
<Descriptor xsi:type="HomogeneousTextureType">
   <Average>184</Average>
   <StandardDeviation>102</StandardDeviation>
   <Energy>224 220 205 167 209 226 196 168 172 160
           170 162 172 148 130 164 139 147 142 114
           105 136 121 106 139 91 85 112 95 87
   </Energy>
   <EnergyDeviation>223 224 206 161 212 229 186 164
           172 157 163 153 164 142 113 152 128 141
           138 95 99 119 115 101 142 78 67 106 71 76
   </EnergyDeviation>
</Descriptor>
```

Fig. 3.7: Sample of homogeneous texture description in XML format

3.2.2 Selection of Descriptor Elements

Selection of descriptor elements becomes an important issue in defining the structure of feature vectors. Tab. 3.2 shows the different configurations of colour and textures components in the feature space. Configuration settings providing the best balance are indicated at the third column. Balance between descriptor elements is computed by

$$\text{balance}(\mathbf{d}_{colour}, \mathbf{d}_{texture}) = \arg min(\text{ abs } (\ \|\mathbf{d}_{colour}\| - \|\mathbf{d}_{texture}\| \ ) \ ) \tag{3.3}$$

where $\|\mathbf{d}_{colour}\|$ and $\|\mathbf{d}_{texture}\|$ are the number of constutient elements used by colour and texture descriptors, respectively.

Tab. 3.2: Setting feature vector dimension according to different configurations of colour and texture components in the feature space

Descriptor	Number of elements					Settings
CLD	12					12
CSD	32	64	120	184		64
SCD	16	32	64	128	256	64
EHD	80					80
HTD	32	62				62
$\|\mathbf{d}_{colour}\| = 140$, $\|\mathbf{d}_{texture}\| = 142$,						
					Balance	2
					Feature vector dimension	282

As indicated in column *Settings*, CLD is set up to the default recommended that includes six Y coefficients and three each of Cr and Cb coefficients [130]. CSD is adjusted to 64 elements, which are calculated based on approximations computed using the 184-bin descriptor. SCD is fixed to 64 by scaling the quantised representation of Haar coefficients to obtain the desired number of bits. EHD uses its default corresponding to 80 elements. HTD uses the full layer, it is to say 62 elements.

The 140-element colour and 142-element texture components produce a difference of two elements, which is an acceptable balance. Consequently, the full-size feature vector consists of 282 descriptor elements. Fig. 3.8 shows the improvement on classification accuracy when using combined description elements as feature vectors.

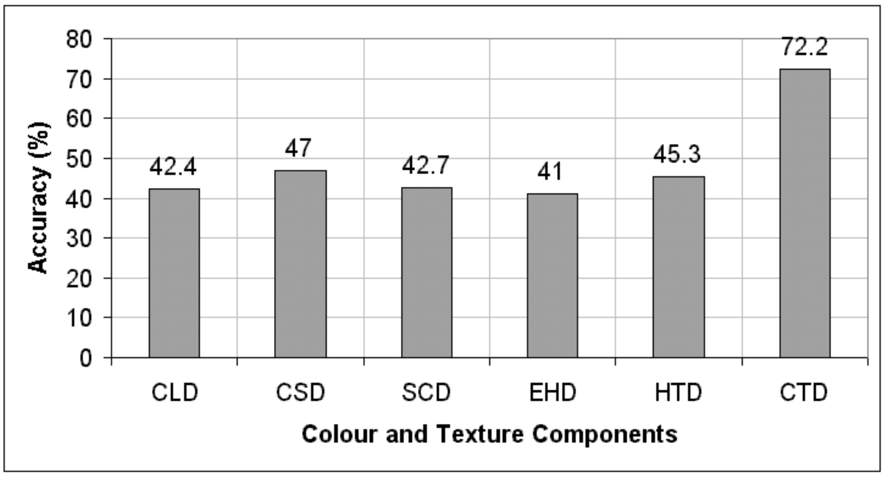


Fig. 3.8: Comparison of classification accuracy between separated and combined descriptor elements. Descriptors used to build the feature space are along the X axis (CLD=colour layout, CSD=colour structure, SC=scalable colour, EHD=edge histogram, HTD=homogeneous texture, and CTD=colour and texture). CTD aggregates colour and texture descriptors. Accuracy of the classifier is indicated at the Y axis. The number of descriptor elements is indicated above each bar

3.3 Empirical Evaluation of Feature Vector Quality

Image features work with different levels of effectiveness depending on the characteristics of the specific image data set. Quality of the proposed feature vector structure is empirically evaluated using the data collection described in Sect. 5.2.3. Briefly, analysis is conveyed using 1,000 colour images of different sizes that were grouped into five classes namely *animal*, *building*, *city view*, *landscape*, and *vegetation*. A picture was manually categorized by a single subject into a certain class if the camera is focused in an object satisfying the name of the class.

3.3.1 Univariate Analysis

It is worth to get an insight of the underlying structure of data before moving into analysing inter-class separability. The learning space is divided into 5 sub-spaces, each formed with elements from specific visual descriptors. Statistical moments (mean and standard deviation) are used to describe the distribution of the image within each class. They provide information on the location and dispersion of feature values (univariate analysis). In addition, contour plots are used to facilitate identification of feature values concentrated within certain intervals.

Vertical and horizontal axes in the contour plot correspond to the independent variables. X axis indicates the index of the descriptor element associated to the feature values. On the other hand, y axis has been divided into ten intervals to identify the approximate value of the iso-responses (contour lines). High frequency of data points within certain intervals can be observed when iso-responses look like “fingerprints”. Distinguishing features report their frequent data points at different locations along the y axis.

Fig. 3.9 to Fig. 3.13 show contour plots (left column) and mean values and standard deviations (right column) of colour and texture descriptors. Co-ordinate of each descriptor element are placed along the x axis. Contours indicate frequency of feature values in the corresponding intervals (y axis). Original feature values were transformed linearly using min-max normalization.

In general, there is a high concentration of values around certain intervals in all components. It could derive onto overlapping of classes and low differentiation between features among them. On the other hand, high variability reduces discrimination capabilities at feature level. A salient feature is not a strong representative of its class when there is a high variability intra-class. Overlapping and weak salient features have undesirable effects on classification outcomes. However, it is not convenient to anticipate low classification rates without performing an inter-class analysis. Therefore, next section is focused on that issue.

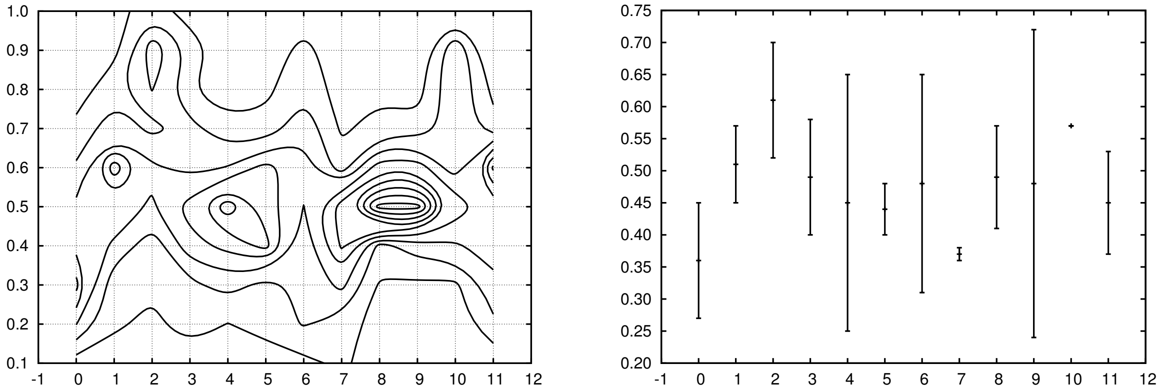


Fig. 3.9: Colour layout component. It is not observed in the contour plot (left column) a high concentration of values around a specific interval, save co-ordinates 8, 9, and 11. The mean values and standard deviations (right column) present high variability (e.g. co-ordinates 4, 6, and 9). Only few features report low variability (see co-ordinates 7, 10)

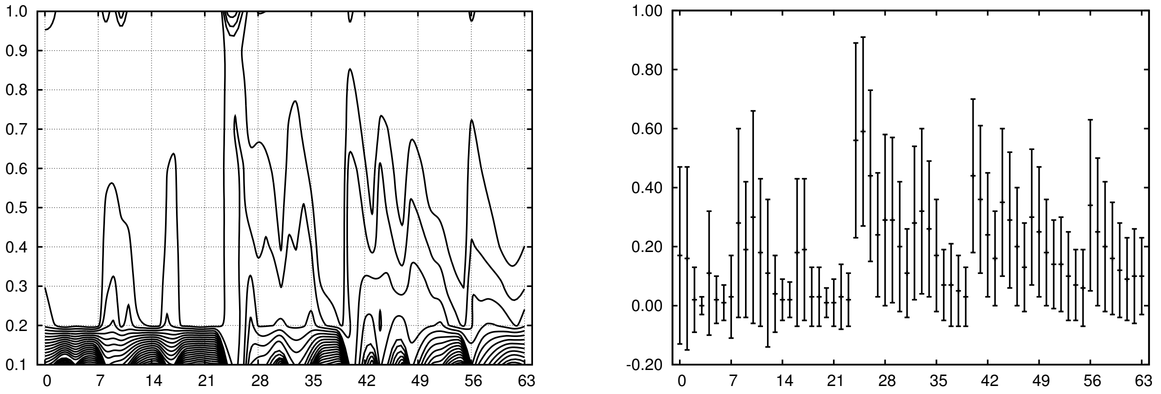


Fig. 3.10: Colour structure component. It is observed in the contour plot (left column) a high concentration of values around several intervals. The mean values and standard deviations (right column) present high variability save very few features (see co-ordinates 3, 6, 15, 20)

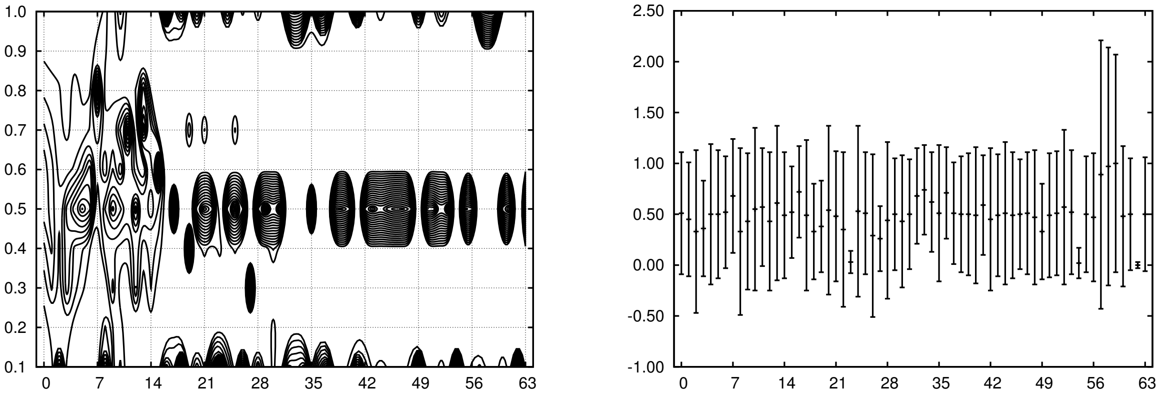


Fig. 3.11: Scalable colour component. Similarly to CSD, the contour plot (left column) shows a high concentration of values around several intervals. The mean values and standard deviations (right column) also present high variability. Only few features report low variability (see co-ordinates 24, 54, 62)

3.3.2 Inter-Class Analysis

A nonparametric *Wilcoxon rank-sum test* is applied to analyse inter-class separability of the learning space using the proposed feature vector structure. It is an alternative to the two-

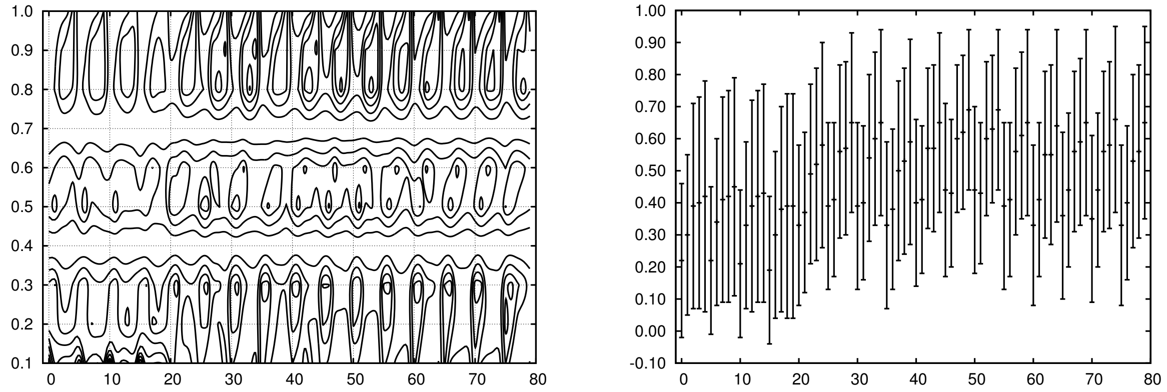


Fig. 3.12: Edge histogram component. The contour plot (left column) reports a high concentration of values around several intervals. The mean values and standard deviations (right column) present high variability

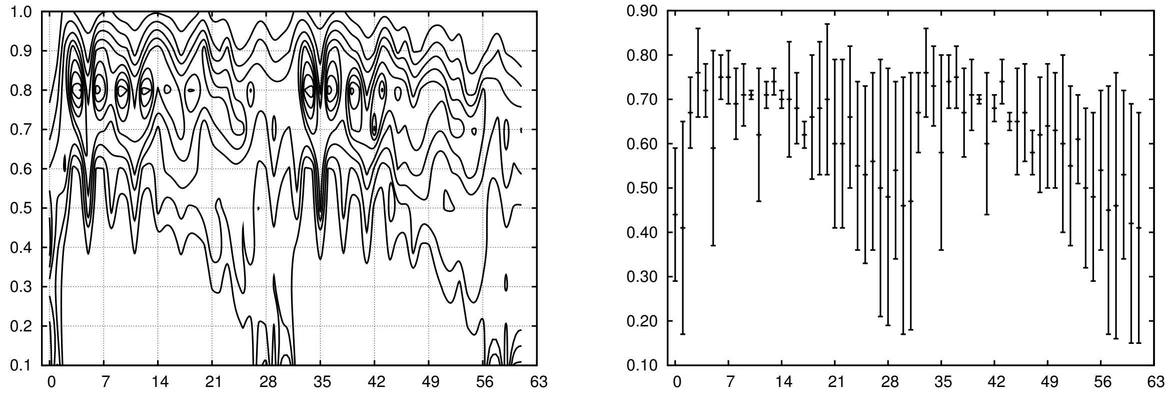


Fig. 3.13: Homogeneous texture component. It is observed in the contour plot (left column) a high concentration of values around some intervals. The mean values and standard deviations (right column) present high variability save few features (see co-ordinates 11, 13, 14, 15, 18, 41, 43, 45)

sample t-test. Some of the advantages of the Wilcoxon test are (1) validity even for data from any distribution, whether Normal or not, and (2) much less sensitivity to outliers than the two-sample t-test (Wild and Seber [137]). These properties of the Wilcoxon test are suitable to deal with MPEG-7 visual descriptors (*eg.* Homogeneous Texture is characterized by a Rayleigh distribution). A detailed description of the Wilcoxon rank-sum test is presented in App. A.

The aim of proposed feature vector structure is to retain the topology of the original descriptor elements and consequently their embedded semantics. It facilitates evaluation of features individually. The Wilcoxon test is used as filter method to quantify discrimination capabilities of feature vectors using different dimensions. Individual feature selection is carried out applying statistical hypothesis testing.

Performing the test for all individual co-ordinates ($k = 1, 2, \ldots, p$) it ends up with the set of salient features that can be seen as a family of retained indexes identifying position of elements in the image descriptions for which the null hypothesis is rejected. As is illustrated in Fig. 3.14, the following shorthand notation is used to describe these indexes:

$$\begin{aligned} \mathbf{I}_{color} &= \mathbf{I}_{CLD} \mid \mathbf{I}_{CSD} \mid \mathbf{I}_{SCD} && (3.4) \\ \mathbf{I}_{texture} &= \mathbf{I}_{EHD} \mid \mathbf{I}_{HTD} \ . && (3.5) \end{aligned}$$

$$\boxed{\begin{array}{lcc} \texttt{Original vectors} & \overbrace{1,2,\ldots}^{colour} & \overbrace{\ldots,p-2,p-1,p}^{texture} \\ \texttt{Feature selection} & \downarrow & \downarrow \\ \texttt{Retained indexes} & \mathtt{I}_{colour} & \mathtt{I}_{texture} \end{array}}$$

Fig. 3.14: Retained indexes are identified through statistical hypothesis testing of differences between means of the feature vectors ascribed to each concept. A retained index is a point in an array of co-ordinates that indicates the position of a salient feature. Retained indexes are useful as a filter method

The feature vector structure could be written down as

$$\mathbf{x} = [\mathbf{x}_{colour} | \mathbf{x}_{texture}]^T \tag{3.6}$$

to emphasize that it is a concatenation of the features concerning colour and texture. Schematically, this process of feature selection derives onto a reduction of the original learning space as is portrayed in Fig. 3.15.

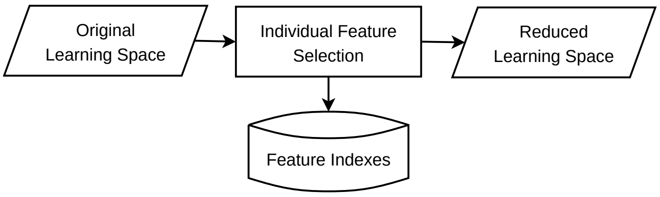


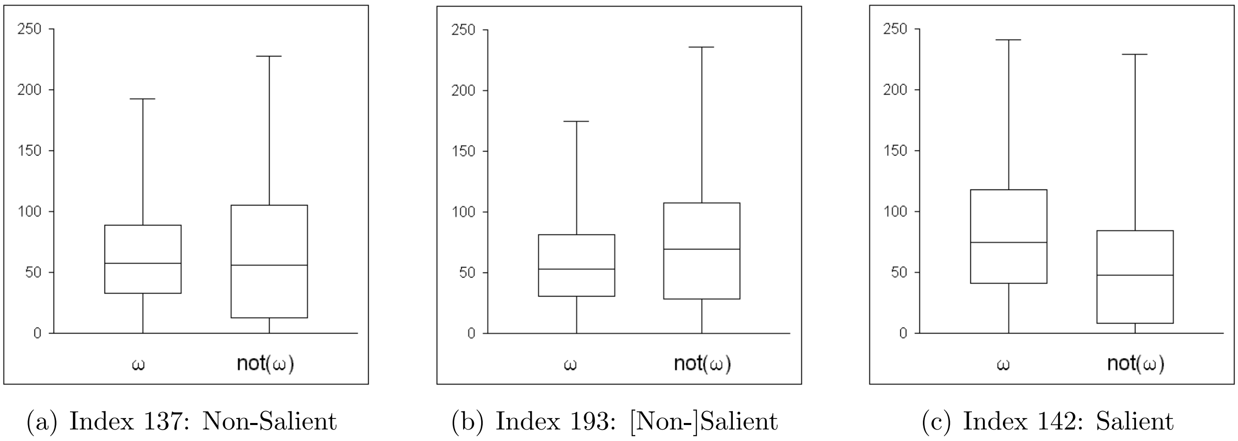
Fig. 3.15: Feature selection is carried out on features individually. Retained indexes are used to filter out non-salient features

If the P–*value* is equal to or less than a given critical value t_α (confidence level), the null hypothesis can be rejected. As shown in Fig. 3.16, the confidence level (α) impacts the feature selection and consequently the level of achieved feature reduction.

Obviously, as it is well known in statistical analysis, there are only few typical α values of such confidence levels; those being the most commonly used are 0.01, 0.05, and 0.1. Nevertheless in the evaluation some other values are also worth experimenting with so that one could gain a better insight into the selection abilities provided in this manner.

Semantic separability concerns a two-class problem, ω and not(ω), so as a matter of fact the feature selection focuses on conceptual analysis, it is to say discrimination between the occurrence or not of a concept. Furthermore, any multi-class classifier can be represented as a family of two-class classifiers.

Then, detection of salient features is carried out using two groups of observations. Each group with samples of feature vectors extracted from two sets of images. One set contains



(a) Index 137: Non-Salient (b) Index 193: [Non-]Salient (c) Index 142: Salient

Fig. 3.16: These selected examples from the colour histogram component show the effects of confidence level on feature selection. (a) Index 137 is always reported as non-salient feature; (b) Index 193 is reported as non-salient feature for $\alpha = 0.01$ and 0.05, but as salient feature for $\alpha = 0.1$ and 0.2; (c) Index 142 is always reported as salient feature

images ascribed to certain concept –class ω. The other set consists of images that are not related to the concept –class not(ω).

In this manner the design considers subsets of features that are discriminative for a specific pair of concepts. Counting how many times a given feature is recognized as being discriminative helps assess its suitability to distinguish features belonging to different classes.

Differences are predominant positive (or negative) if the alternative hypothesis, H_1, is true. Definitely, the features for which the null hypothesis has never been rejected for each class can be regarded as meaningless. The remaining ones are selected as salient features.

The number of salient features determines the percentage of retained indexes required to differentiate one class from another. Fig. 3.17 quantifies the effect of feature selection occurring for each category (concept) when varying the values of the confidence levels (values of α).

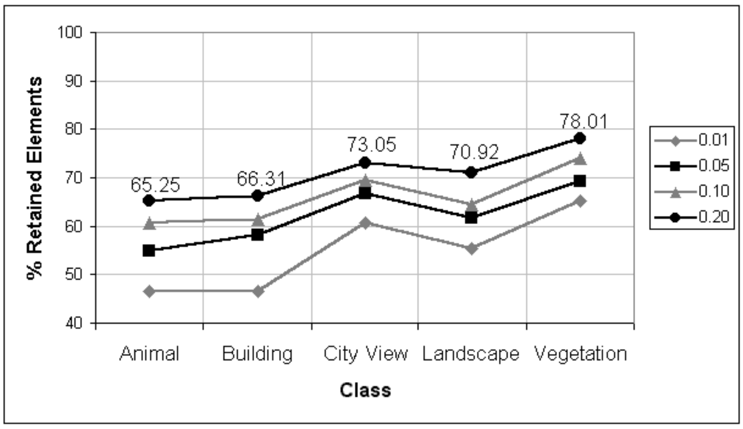


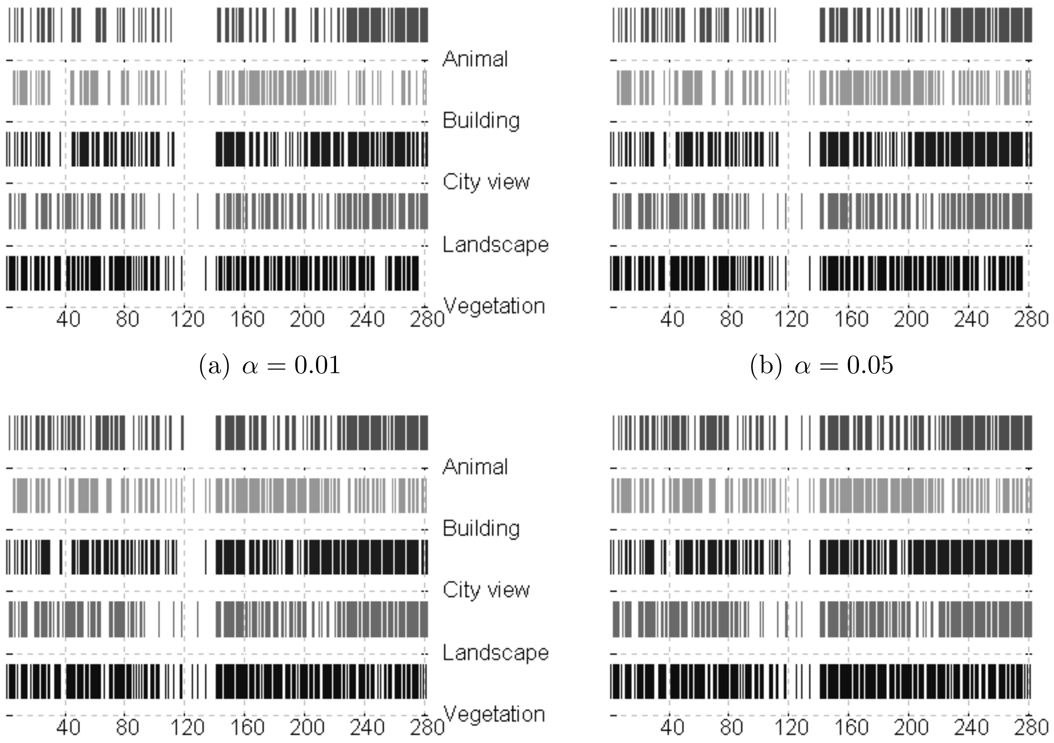
Fig. 3.17: Quantification of salient features expressed as percentage of retained indexes after varying the α value. The labels above the curves indicate percentage of features satisfying the alternative hypothesis

Tab. 3.3 shows the details of this measure of retained indexes for the α value set up to 0.2. This specific confidence level has been chosen experimentally. It represents a low probability (1-0.2)=0.8, that is, 80% confidence level. This α is used as an upper bound.

Tab. 3.3: Percentage of retained descriptor elements for different classes of images and MPEG-7 visual descriptors

	Number of elements			
$\alpha = 0.20$	$\mathbf{I}_{colour}$	$\mathbf{I}_{texture}$	$\mathbf{I}_{colour}\vert\mathbf{I}_{texture}$	% Retained elements
Original space	140	142	282	N/A
Animal	71	113	184	65.25
Building	76	111	187	66.31
City View	76	130	206	73.05
Landscape	80	120	200	70.92
Vegetation	96	124	220	78.01
Average	79.8	119.6	199.4	70.71

Fig. 3.18 depicts retained indexes formulated for the learning space at different confidence levels. Scanning the indexes by concept (horizontal direction) several non-salient features can be found. In contrast, a vertical scanning shows that practically every co-ordinate in the family of indexes contains features that satisfied the alternative hypothesis in at least one case.



(a) $\alpha = 0.01$

(b) $\alpha = 0.05$

(c) $\alpha = 0.10$

(d) $\alpha = 0.20$

Fig. 3.18: Retained indexes at different confidence levels (values of α). Horizontal axis corresponds to vector co-ordinates (282 features). Vertical axis indicates the category. Many of the co-ordinates, in the family of indexes, satisfied the alternative hypothesis

Knowing that it will be possible to partition the feature space into semantic classes, the next question is regarding to the dimensionality of the feature vectors. A first attempt could be using the retained indexes. However, experimental studies showed that there is not significant difference in the classification outcomes for reasonable percentages of training images (around 30%). Fig. 3.19 summarises selected results to illustrate this issue.

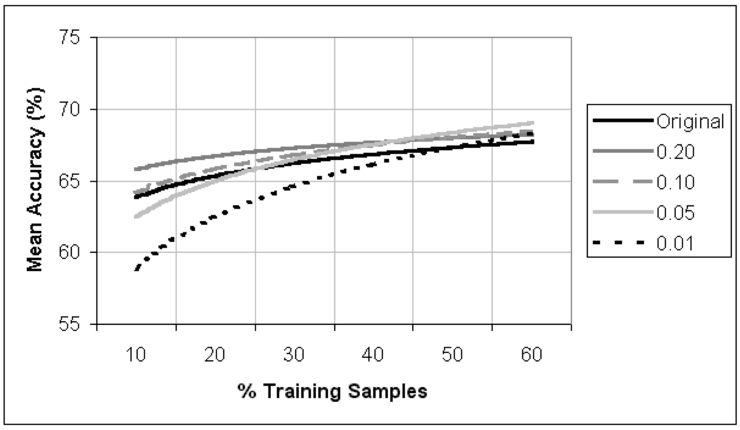


Fig. 3.19: Accuracies obtained in the multi-class problem using the original learning space is very close to the ones obtained using salient features at different confidence levels

Classification produced using salient features is quite close to the one obtained when dealing with the original feature vectors. Therefore, it is suggested to use the original vector, since the accuracy will be similar with a lower computational cost, i.e. avoiding the filtering module.

It was mentioned above the idea of avoiding transformations to preserve the semantic embedded on each feature element. However, a compulsory step is to analyse whether or not principal components could contribute to improve the partition of the feature space and subsequently the classifier performance. Following section presents considerations and results of principal component analysis with regard to the proposed feature vector structure.

3.3.3 Principal Components Analysis

The purpose of principal components analysis is to reduce the complexity of the multivariate data into the principal components space and then choose the first q principal components ($q < p$) that explain most of the variation in the original variables. The following criteria for selecting the number of principal components are suggested in the literature:

- *Scree plot.* Important components are separated from the less important ones using the Scree plot as visual reference. This plot portrays the eigenvalues λ_j against j ($j = 1, 2, \ldots, p$),. This criterion includes principal components with eigenvalues before the "elbow".

- *Component exclusion.* For the principal components calculated from the correlation matrix, the average eigenvalue is 1. This criterion excludes principal components with eigenvalues less than 1 (below the average).

- *Cumulative fraction.* This criterion includes components that explain some arbitrary amount (typically 80%) of the variance.

Fig. 3.20 and Fig. 3.21 display the Scree plots of colour and texture components used to built the feature space. The x axis contains the principal components sorted by decreasing fraction of total variance. The y axis contains the fraction of total variance. Normalized values of principal components are placed along the lower line. The small plot inside magnifies the area around the "elbow" and indicates the variance accumulated at that component. The cumulative fraction of total variance explained is also shown in the upper line. It is indicated when the cumulative fraction is 90% or greater in order to identify the corresponding principal component at that point.

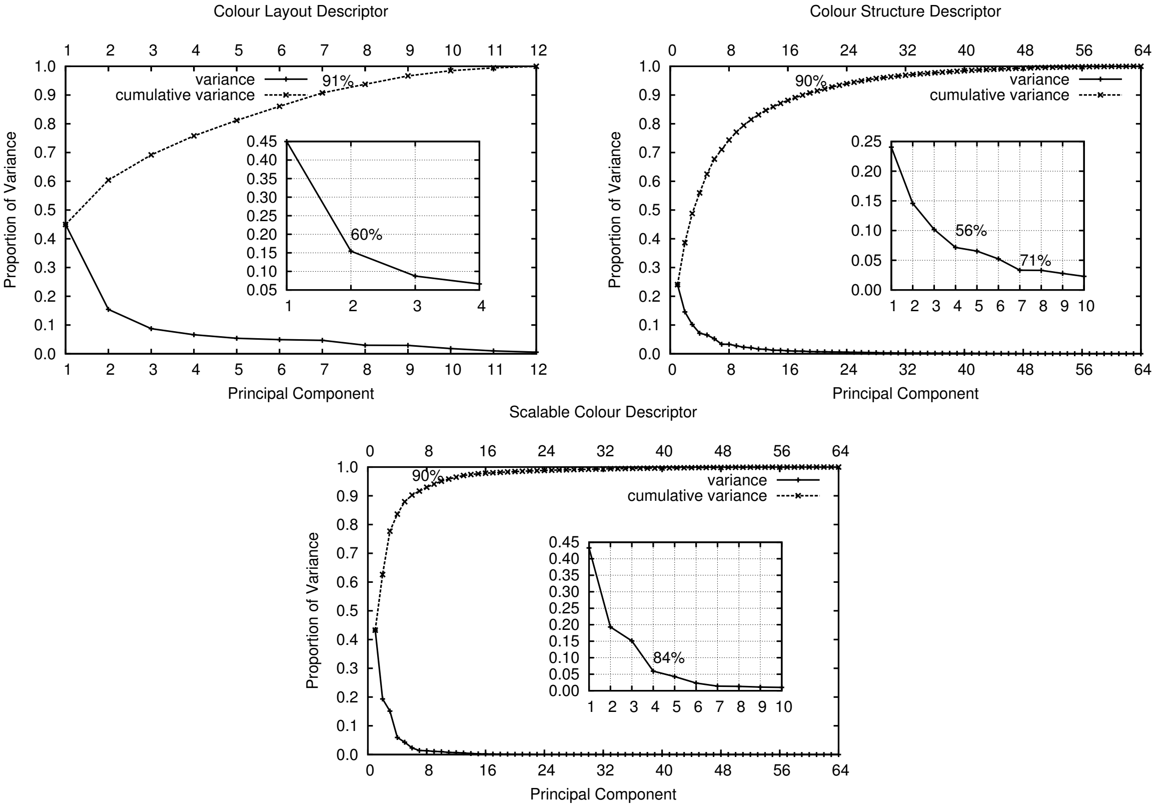


Fig. 3.20: Scree plots of colour components used to built the feature space

After combining the criteria to select the most convenient number of principal components, the Karhunen-Loeve transform is applied on the learning space. The resulting space is clustered to compare the outcomes using the original vectors and the re-defined ones. Results are summarized in Fig. 3.22.

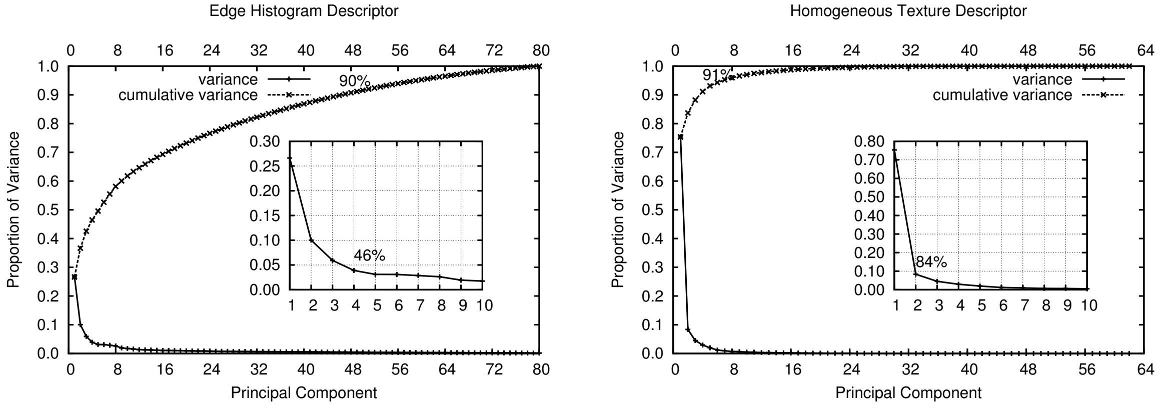


Fig. 3.21: Scree plots of texture components used to built the feature space

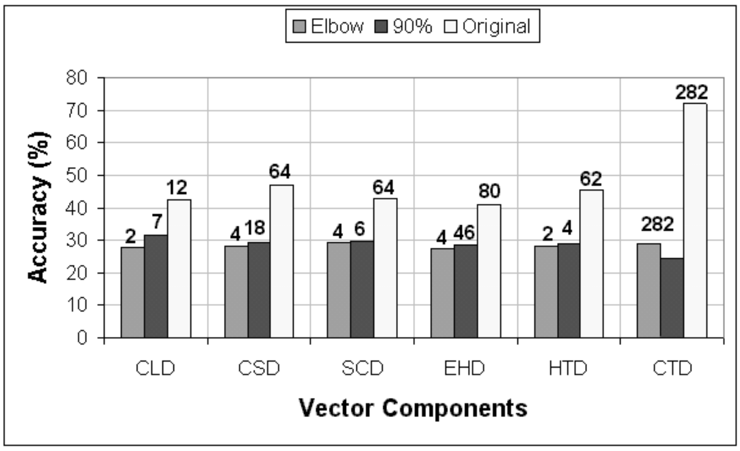


Fig. 3.22: Comparison of space partition using either principal components or original vectors. Descriptors used to build the feature space are along the X axis (CLD=colour layout, CSD=colour structure, SCD=scalable colour, EHD=edge histogram, HTD=homogeneous texture, and CTD=colour and texture). CTD aggregates the descriptors as indicated in Sect. 3.2.2. Feature vector sizes are indicated above each bar

As can be observed, partitioning of the feature space using the transformed vectors is not improved. Original vectors produce a more accurate partition in terms of expected semantic grouping. Because the framework incorporates clustering mechanisms, these results show that it is not convenient to use principal components prior to clustering. Similar conclusions have been found in another empirical studies, though in a different context (Yeung and Ruzzo [138]).

3.4 Summary

A feature vector structure for building an MPEG-7 learning space was proposed. It combines visual descriptors keeping discriminative elements and the semantics embedded in the original descriptions. Structure settings provide a suitable balance between components. Results of univariate, inter-class, and principal component analyses suggest the usage of full size vectors to accomplish a better partition of the feature space.

Chapter 4

User-Driven Fuzzy Clustering

Clustering methods are commonly used to identify natural groups of unlabelled data [139]. As classification method, clusters are associated with a set of known classes to induce a model that categorizes oncoming data.

Applying conventional clustering methods, feature vectors are either assigned to or not assigned to a defined group. Fuzzy clustering, which applies the concept of fuzzy sets (ref [140] [141]), assigns to each feature vector a membership degree, ranging between 0 and 1, to each cluster.

As mentioned by [142], three problems are found in clustering algorithms: (1) determining the optimal number of clusters to partition the data space, (2) the clusters do not match the expected groups, and (3) the cluster populations are equalized.

Since the number of classes in conceptual analysis of image content can be predetermined, the optimal number of clusters to partition the data space can be computed in function of the class set cardinality. Subsequently, validity functions (cf. [143][144]) such as *the fuzziness performance index* [145] or *the compactness and separation* [146][147] can be used. Furthermore, it can be assumed that a human interpreter has classified a small set of images per class in the database.

The second problem can be tackled using the exemplars (labelled data) to guide the clustering algorithm towards a desired partition of the descriptor space [118]. However, the nature of the problem demands an extension to deal with the subjectivity and fuzziness of the human interpretation [148].

Shape and size regularization methods have been proposed (cf. [149]) to handle the third problem, which occurs in response of the tendency presented in c-means clustering algorithms when grouping data within (hyper-) spherical or (hyper-) ellipsoid spaces based on the similarity to the cluster prototypes. Besides, Pedrycz and Waletzky [119] indicates that the objective function being selected in advance predefines the shapes that want to be found in the data set.

The rest of this chapter is organized as follows: Sect. 4.1 introduces fuzzy clustering and its usage as pre-classification step of image content. Sect. 4.2 describes a structural data analysis used to determine proximity and overlapping between classes, which leads to misclassification problems. Sect. 4.3 gives details on a proposed partially supervised clustering algorithm. Sect. 4.4 summarizes the chapter.

4.1 Unsupervised Fuzzy Partition of the Feature Space

Clustering methods help to organize low-level features into groups which interpretation may relate to some classification or description task pertaining to the image content. Thus, feature vectors are clustered according to similarities among them [150]. Such a similarity between vectors is quantified or measured using a proximity metric.

Fuzzy clustering applies a partitioning-optimisation technique based on minimization of an objective function that measures the desirability of partitions [151].

The criterion function is a scalar index that indicates the quality of the partition and has the form

$$J(X, \mathbf{V}, \mathbf{U}) = \sum_{i=1}^{N_X} \sum_{j=1}^{c} u_{ij}^m d^2(\mathbf{x}_i, \mathbf{v}_j) \ , \tag{4.1}$$

where X is a data space consisting of N_X p-dimension feature vectors to cluster, $\mathbf{V}$ is a set of c $(2 \leq c \leq N_X)$ cluster prototypes – centers, and $\mathbf{U}$ is a matrix belonging to the set of all possible fuzzy partitions defined by

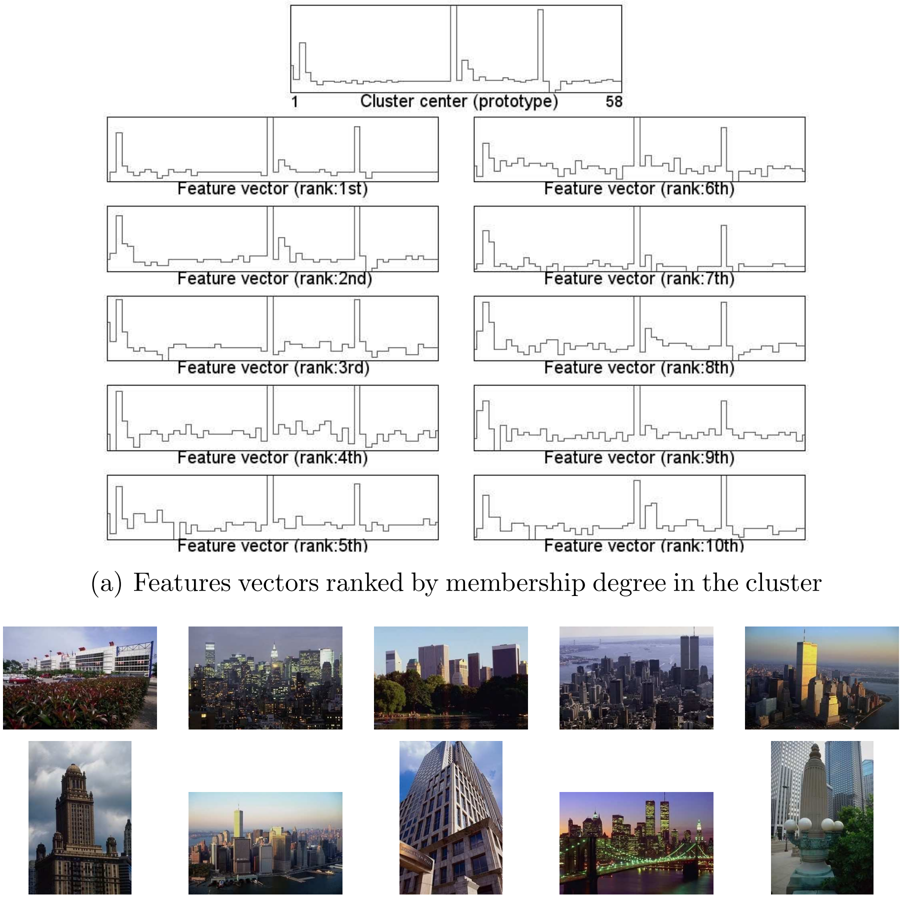
$$\Im = \left\{ \mathbf{U} \in \Re_{N_X c} | \underset{\substack{1 \leq i \leq N_X \\ 1 \leq j \leq c}}{\forall} u_{ij} \in [0,1] \, , \, \sum_{j=1}^{c} u_{ij} = 1, \ 0 < \sum_{i=1}^{N_X} u_{ij} < N_X \right\} \ , \tag{4.2}$$

where u_{ij} is the degree of membership of vector $\mathbf{x}_i$ in the cluster j, $\mathbf{v}_j$ is the p-dimension prototype of the cluster, $d^2(\cdot)$ is any distance norm expressing the similarity between any feature vector and the prototype, and m $(1 < m < \infty)$ is a fuzzification exponent which determines the degree of overlap of fuzzy clusters.

Fig. 4.1 presents clustering results after applying fuzzy c-means onto a two-class classification problem using colour descriptions. In this case feature similarity is used to ascribe images to a common semantic class.

If clusters are manually labelled as a representative of a class with an identifying string, e.g. “City view”, then the problem may appear to have been finessed. However, the adequacy of such a solution depends on human interaction, which is completely subjective.

As described in Sect. 1.1.2, it is intuitive that two objects can be similar in their visual primitives but semantically different to a human observer. This is a drawback to classify images using only low-level vision and the foundation of the critical paradigm of “bridging the semantic gap” [14].



(a) Features vectors ranked by membership degree in the cluster

(b) Images used to extract the low-level vectors

Fig. 4.1: Sample of clustering results resembling semantic grouping (e.g. Outdoor or City view)

Alternatively, if clusters are described by an equivalence class

$$[\mathbf{v}_j]_E \doteq \{\mathbf{x}_i : \mathbf{x}_i \in X, E(\mathbf{v}_j) = 1\} \ , \tag{4.3}$$

where the cluster prototype $\mathbf{v}_j$ is used to determine those feature vectors $\mathbf{x}_i$ that are part of the cluster. Then, the set of equivalence classes

$$X/E \doteq \{[\mathbf{v}_j]_E\} \tag{4.4}$$

called a quotient set forms a partition of the feature space. The clustering outcome can be used as a classification function based on the map from X onto X/E is defined by

$$\phi : X \mapsto X/E \ . \tag{4.5}$$

Information provided by this pre-classification step along with hints given by an expert user regarding to the problem domain knowledge are quite useful to design the semantic classifier and subsequently improve its accuracy.

4.2 Structural Data Analysis

As the image classifier is designed on the space formed with low-level visual primitives, it is worth to take another look at images by running cluster analysis and visualizing the relationships between the clusters and classes as well as linkages formed between the clusters themselves.

Using a Gustafson Kessel-Type of matrix, this analysis provides an interesting insight as to the geometry and complexity of classes, homogeneity of clusters and proximity between the classes. Anticipating the possible complex geometry of individual classes, usually the number of cluster is kept higher than the number of categories of images.

It is very likely that the cluster is not completely homogeneous and could comprise also some other patterns coming from remaining classes.

The first and second classes with higher percentage of patterns within the cluster are marked as *dominant classes*. Accepting the notation in which a size of dots corresponds to the percentage of the dominant class forming the cluster and a thickness of line originating from the cluster characterizes its linkages with other classes, it can succinctly portray the essential relationships between clusters, classes, and their geometry.

As illustrated in Fig. 4.2, clusters can be predominantly associated with a class with practically no linkages (associations) with other classes (e.g. $c_1 : \omega_3$). Others are very much a mixture of classes with very limited dominance of the most frequent class ($c_2 : \omega_1, \omega_2, \omega_3$). There is a weaker linkage between classes ω_1 and ω_3 than the one presented by classes ω_2 and ω_3.

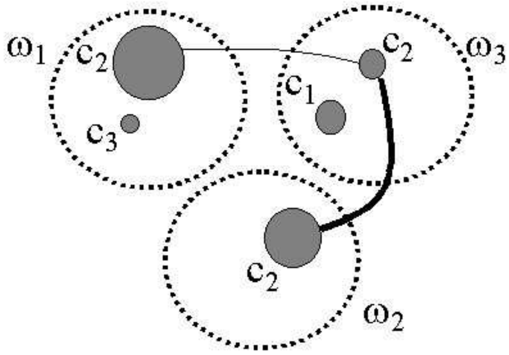


Fig. 4.2: Graphical visualization of clusters (c_1, c_2, c_3) and related classes ($\omega_1, \omega_2, \omega_3$) along with the "classification" content of the clusters

Fig. 4.3 presents cluster-class relationships after applying fuzzy c-means onto a multi-class classification problem. The feature space is grouped into five semantic classes namely *animal*, *building*, *city view*, *landscape*, and *vegetation*. The entries correspond to percentage of feature vectors from each class assigned to the cluster. Within the plot percentage are represented by the bubbles' size.

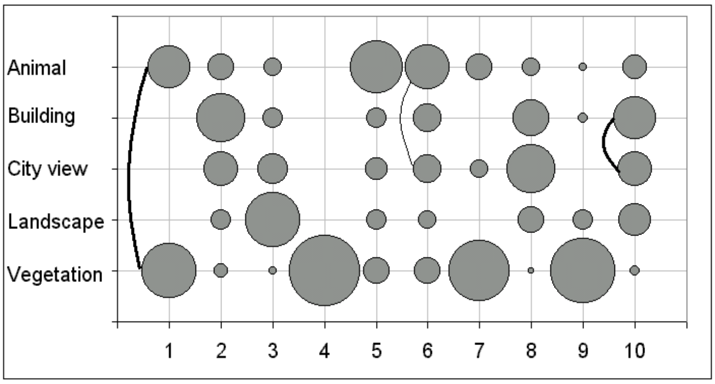


Fig. 4.3: Cluster-class dependencies (5 × 10). Connecting lines indicate class linkages. E.g. semantic relations between *animal-vegetation* (cluster 1) and *building-city view* (cluster 10) are stronger than *animal-city view* (cluster 6)

Occurrence of several dominant classes contributes to higher values of the confusion rate coming with the specific cluster. Furthermore, ranking the frequent *class pairs*, presenting higher levels of association, helps to determine classes in which their abstractions are leading to misclassification. Tab. 4.1 summarizes this observation on a basis of the cluster-class dependencies.

Tab. 4.1: Ranking of frequent class pairs presenting higher levels of association leading to misclassification

Rank	Pair of classes
1	Vegetation - Animal
2	Building - City view
3	Vegetation - Landscape
4	Landscape - City view
5	Building - Animal

Looking at these class pairs, *vegetation-animal* presents the highest confusion. This semantic overlapping can be observed at feature and perceptual levels. One sample of the latter is given in Fig. 4.4 in which an animal image satisfies also criteria of vegetation.

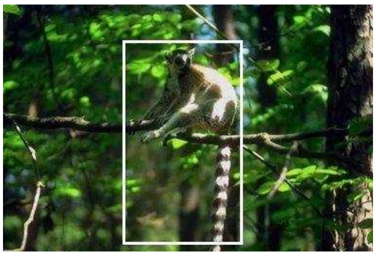


Fig. 4.4: Semantic overlapping: image categorized as *animal* with strong content of *vegetation*

4.3 Partially Supervised Clustering for Semantic Classification

Semantic classification of image content combines low and high-level numerical interpretations of visual content. The built-in knowledge of descriptions enables systems to perform more accurate partition of the feature space.

Applying standard clustering algorithms, the criterion function presented in Eq. 4.1 produces normally a partition of the feature space that is significatively different from the expected semantic groups. It can be denoted as

$$\mathbf{U} \nsim \mathbf{G}_{\Omega} \ , \tag{4.6}$$

where Ω corresponds to the set of classes and $\mathbf{G}_{\Omega}$ is the expected partition.

As aforementioned, it is possible to guide the clustering algorithm towards a desire partitioning of the feature space such that

$$\mathbf{U}_{(0)} \circlearrowleft \ldots \circlearrowleft \mathbf{U}_{(t)} \circlearrowleft \mathbf{U}_{(t+1)} \sim \mathbf{G}_{\Omega} \ . \tag{4.7}$$

On the follow, a partially supervised clustering algorithm is proposed to satisfy Eq. 4.7. The training data set is denoted by

$$X = \left\{ \underbrace{\mathbf{x}_1^1, \ \ldots \ \mathbf{x}_{n_1}^1}_{labelled\ 1} \ldots \underbrace{\mathbf{x}_1^c, \ldots, \mathbf{x}_{n_c}^c}_{labelled\ c} \ \underbrace{\mathbf{x}_1^u, \ldots, \mathbf{x}_{n_u}^u}_{unlabelled} \right\} = X^d \cup X^u \ , \tag{4.8}$$

where superscripts $1, \ldots, c$ indicate the class label for design data and u for unlabelled data. It leads to a partition matrix with the form

$$\underbrace{\mathbf{U}}_{N_X \times c} = \begin{bmatrix} \underbrace{\mathbf{U}^d}_{c \times N_d} & \underbrace{\mathbf{U}^u}_{c \times N_u} \end{bmatrix}^T \ . \tag{4.9}$$

Information concerning design data can also be provided using additional structures [119]. A binary vector to indicate whether the data is or is not labelled.

$$\begin{aligned} \mathbf{b} &= [b_i] \, , \ i = 1, 2, \ldots, N_X \\ b_i &= \begin{cases} 1, & \mathbf{x}_i \in X^d \\ 0, & \mathbf{x}_i \in X^u \end{cases} \end{aligned} \tag{4.10}$$

and a matrix containing degrees of memberships for the known data

$$\mathbf{F} = [f_{ij}] \, , \ i = 1, 2, \ldots, N_X; \ j = 1, 2, \ldots, c \ . \tag{4.11}$$

The partially supervised method, based on [142][119], defines the objective function as follows

$$J_{\Omega}(X, \mathbf{V}, \mathbf{U}) = \sum_{i=1}^{N_X} \sum_{j=1}^{c} (1 - b_i + \beta f_{ij} b_i)^m u_{ij}^m d^2(\mathbf{x}_i, \mathbf{v}_j) \ , \tag{4.12}$$

where the binary vector $\mathbf{b}$ and matrix $\mathbf{F}$ are defined by Eq. 4.10 and Eq. 4.11, respectively. β ($\beta \leq 0$) denotes a scaling factor to keep a balance between the supervised and unsupervised components within the minimization-optimisation mechanism [119]. As studied in [118], the fuzzification exponent m is set up to 2. The value of β is suggested to be proportional to the rate N_X/N_d where N_d indicates the number of labelled data.

The necessary conditions for minimization of Eq. 4.12 can be obtained using the Lagrange multipliers technique with the constraints established at Eq. 4.2.

The distance matrix is calculated as

$$d_{ij}^2 \doteq \|\mathbf{x}_i - \mathbf{v}_j\|_{\mathbf{A}}^2 = (\mathbf{x}_i - \mathbf{v}_j)^T \mathbf{A}_j (\mathbf{x}_i - \mathbf{v}_j) \ , \tag{4.13}$$

where $\mathbf{A}_j$ is the identity matrix for Euclidean distance and inverse of fuzzy variance-covariance matrix for Mahalanobis distance [119]. The latter is computed as follows

$$\mathbf{A}_j^{-1} = \left[\frac{1}{\rho_j det(\mathbf{P}_j)} \right]^{\frac{1}{n}} \mathbf{P}_j \ , \tag{4.14}$$

where typically $\rho_j = 1, \ j = 1, \ldots, c$, and

$$\mathbf{P}_j = \frac{\sum_{i=1}^{N} u_{ij}^2 (\mathbf{x}_i - \mathbf{v}_j)(\mathbf{x}_i - \mathbf{v}_j)^T}{\sum_{i=1}^{N} u_{ij}^2} \ . \tag{4.15}$$

Cluster prototypes are defined by

$$v_j = \frac{\sum_{i=1}^{N} (1 - b_i + \beta f_{ij} b_i)^2 u_{ij}^2 \mathbf{x}_i}{\sum_{i=1}^{N} (1 - b_i + \beta f_{ij} b_i)^2 u_{ij}^2} \tag{4.16}$$

and the membership degrees are computed by

$$u_{ij} = \begin{cases} f_{ij}, & b_i = 1 \\ \left[\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}} \right)^2 \right]^{-1}, & b_i = 0 \end{cases} \ . \tag{4.17}$$

The complete algorithm is summarized in Tab. 4.2

Fig. 4.5 depicts a two-class synthetic data set presented by [142]. It illustrates clustering results applying the unsupervised (standard fuzzy c-means, on the left) and partially supervised (proposed partially supervised fuzzy c-means, on the right) algorithms presented above. The contours at the bottom of each surface serve to visualise the partitions of feature space.

Tab. 4.2: Partially supervised clustering algorithm

Given	$X = X^d \cup X^u$, data space containing labelled and unlabelled data
	$N_X = N_d + N_u$, number of feature vectors
	c, number of clusters
	$\mathbf{b}$, indicator vector
	$\mathbf{F}$, known membership matrix
	δ, criterion used in the Picard Iteration
	ϵ, maximum number of epochs (optional stop condition)
Step 1	Initialise the partition matrix randomly, $\mathbf{U}_{(0)}$ including $\mathbf{F}$
Step 2	Calculate cluster prototypes using Eq. 4.16
Step 3	Compute the distance matrix applying Eq. 4.13
Step 4	Update the partition matrix using Eq. 4.17
Step 5	Compare $\mathbf{U}_{(t+1)}$ to $\mathbf{U}_{(t)}$. If $\lVert \mathbf{U}_{(t+1)} - \mathbf{U}_{(t)} \rVert < \delta$ or $t > \epsilon$ then STOP
	Otherwise return to step 2 with $\mathbf{U}_{(t)} = \mathbf{U}_{(t+1)}$

Few labelled data guide the algorithm towards the right partitions.

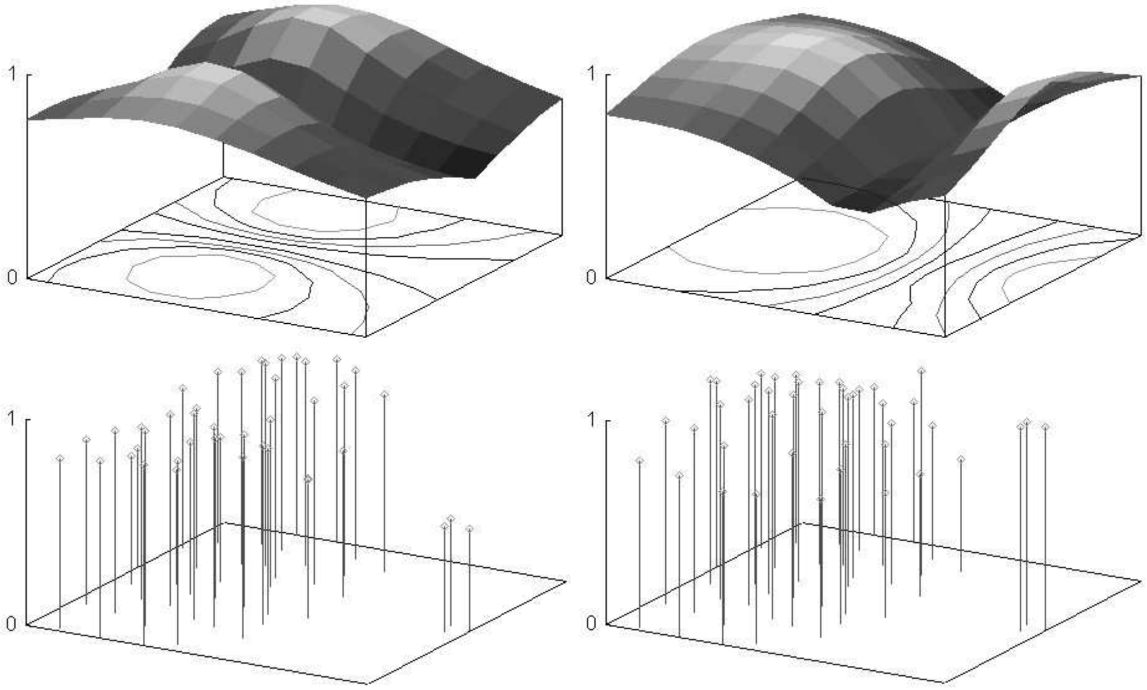
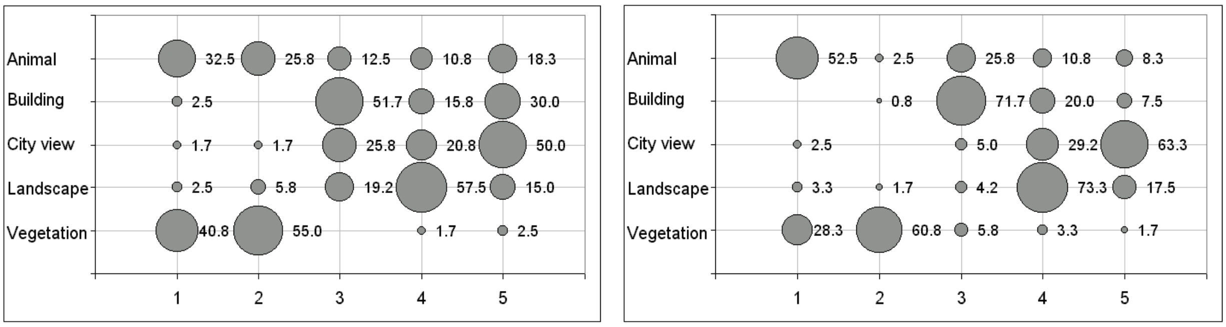


Fig. 4.5: Space partition, membership degrees, cluster prototypes, and shapes obtained by unsupervised clustering (left) do not match the expected solution as does its counterpart the partially supervised algorithm (right) using few labelled data

Is it noticeable that accordingly to the problems indicated by [142] unsupervised clustering fails in trying to present the expected groups. On the other hand, proposed algorithm not only matches the expected groups but also provides the algorithm with an appropriate objective function to cluster the populations. It is worth to stress that supervision was limited to few datapoints.

Fig. 4.6 summarises structural data analysis of unsupervised and partially supervised clustering applied on another data set. Providing the algorithm with a low percentage a labelled images, the level of matching between clusters and classes is increased in a 30%. All dominant classes within the clusters became stronger, except by cluster “4”, after introducing labelled images. Consequently, cluster-class dependencies are refined and class overlapping is reduced.



(a) Unsupervised feature space partition

(b) Partially supervised feature space partition

Fig. 4.6: Structural data analysis based on cluster-class dependencies (5 × 5). Values next to the bubbles indicate percentage of patterns within the cluster. Expected partition is improved from 49% to 64.3% by labeling 17% of training data

4.4 Summary

Information obtained through clustering algorithms (prototypes, space partition) and analysis (cluster-class dependencies, ranking of overlapped categories) provides a better insight of the image abstractions, refine the classifier design, and improve classification performance. More specific “user-driven” information is incorporated by labelled data.

Ranking of dominant classes along with information given by class pairs and the strength of their linkage is useful to perform either individual feature selection or feature weighting in order to minimize the average within-cluster dispersion and maximizes the average between-cluster dispersion.

Clustering outcomes can be used to learn support vectors and subsequently define an optimal decision hyperplane to classify new patterns. Alternatively, cluster prototypes can be used in the design of a radial basis function type of classifier. Thus, partially supervised clustering equipped with an objective function establishes a solid base to build a more accurate semantic categorization.

Chapter 6 describes a framework that integrates the proposed partially supervised algorithm to an approach for concept-related indexing of image content. Clustering is used to partition the MPEG-7 learning space presented in Chapter 3.

Chapter 5

A Framework for Concept-related Indexing of Image Content Phase One: Preceding Case Studies

Knowledge representation is one of the fundamental issues to carry out conceptual analysis of image content. This analysis is conveyed vertically and horizontally. Vertically, because it goes from low-level features to high-level concepts. Horizontally, because some results can be only achieved with a multi-modal model. Knowledge is acquired from both visual content and human interpreters.

Conceptual analysis of image content involves a semantic component, which casts the process into the supervised – or learning -from-examples paradigm. Knowledge is acquired by generalizing specific facts presented in a number of design samples (or training patterns). Knowledge acquisition is carried out by a *learning unit*. Analysis of new images is performed by an *indexing unit* that links low-level features to high-level concepts.

Two supervised approaches are studied prior to the framework design. The first approach is described in Sect. 5.1. It uses fuzzy inference rules to represent connections between granules of information at different levels of abstraction. The second approach presented in Sect. 5.2 corresponds to a radial basis function network (RBFN) that uses fuzzy-based receptive fields to map image abstractions to visual interpretations. Sect. 5.3 summarizes the chapter.

5.1 Case Study I: Fuzzy Inference Approach

The nature of the relationships that can be established between image abstractions and concepts involves uncertainty, roughly connections, and subjectivity. This characterization moves the problem into the ground of fuzzy sets theory and fuzzy systems.

One of the advantages of fuzzy-based approaches is that they are closer to human reasoning modes, than their counterparts, i.e. bivalued-logic-based approaches. The main drawback is less accuracy. However, as mentioned by Kuncheva [152], "practice has shown so far that trying to reach the accuracy of a good non-fuzzy model by a fuzzy one is likely to require more time and resources for building up the initial non-fuzzy classifier. Usually the resulting fuzzy model is not transparent enough for the end user to verify and appreciate."

The classification problem defined by Eq. 2.1 on page 11 can be re-written as follows:

$$\mathrm{f} : X \mapsto L \ , \tag{5.1}$$

where $L = \{\ell_1, \ell_2, \ldots, \ell_{n_L}\}$ is a controlled lexicon consisting of n_L symbols representing image concepts (e.g. indoor, outdoor) and X is a feature space.

In the proposed fuzzy inference approach, the classifier employs if-then rules along with an inference mechanism. A transaction $t_{i,j}$ is defined by an instance of the fuzzy classifier $\mathrm{f}(\mathbf{x}_i) = \ell_j$. Transactions are used to capture problem domain, which combines information provided by the machine and the human interpreter.

Implementation of learning and indexing units is presented in the following sections.

5.1.1 Learning Unit

Image abstractions are extracted automatically from still images, organized in feature vectors and then associated with fuzzy variables. Each variable consists of fuzzy sets, which are the inputs of the fuzzy reasoning model.

On the other hand, the learner is provided with a set of training samples tagged with a list of labels from L. Besides labels are numerical values expressing relevance of the corresponding concept with regard to each exemplar. It could be simplified assuming that the list of labels is ordered by their relevance.

The list of labels and their relevance regarding to the image content are organized either as fuzzy sets or as crisp sets (singletons). These sets are the outputs of the fuzzy reasoning model.

Fuzzy sets are used to map image abstractions from real domain ($\Re$) to fuzzy domain ($[0, 1]$). Once mapped, an image is represented by a collection of membership values (degrees of truth) to each fuzzy set. A function-theoretic form estimates the membership value (see App. C).

A membership function, denoted by $\mu_{\widetilde{A}}(\mathbf{x}_i) \in [0, 1]$, represents a possibility distribution.

The set of rules in the fuzzy reasoning model are generated applying a data mining technique on set of transactions (see App. B). Then, transactions consist of fuzzy set-concept pairs as follows:

$$t_{i,j} : (\mu(\mathbf{x}_i), \ell_j) \ , \tag{5.2}$$

where $\mu(\mathbf{x}_i)$ is a list of fuzzy values obtained from $\mathbf{x}_i$ using the membership functions associated to concept ℓ_j.

The result of the data mining process is a set of frequent fuzzy sets-concept pairs found in the transactions set, which in turn determines association rules between image abstractions and concepts.

The outcome of data mining process consists of a collection of rules of the form: "IF *antecedent* THEN *consequent*". Non-interesting rules for the inference system are filtered out to select the most suitable rules.

Fuzzy sets, concepts, and inference rules are the components of the knowledge representation provided by the framework under this approach. The complete learning procedure is summarized in Tab. 5.1.

Tab. 5.1: Learning procedure used in the fuzzy inference approach

```
learning ( sample images[ ], concepts[ ] )
{
  read settings
  for each image in sample images[ ] do
    abstractions[ ] = extract features( image )
  generate fuzzy sets
  for each concept in concepts[ ] do
  {
    for each value in abstractions[ ] do
    {
      fuzzification( value )
      write transaction
    }
    mine association rules( transactions set )
    find inference rules( )
    write rule
  }
  return rule base
}
```

5.1.2 Indexing Unit

Once the knowledge representation has been established, semantic profiles for new images can be automatically created. This is a task for the indexing unit. An overview is depicted in Fig. 5.1.

Image abstractions are extracted and mapped into a fuzzy domain. Then, the fuzzy reasoning model is applied to infer the concepts that could be used to create the semantic image profile.

The fuzzy reasoning model is implemented by an inference system that involves three basic modules: fuzzification, fuzzy inference, and defuzzification. A detailed description of the inference system is presented in App. C.

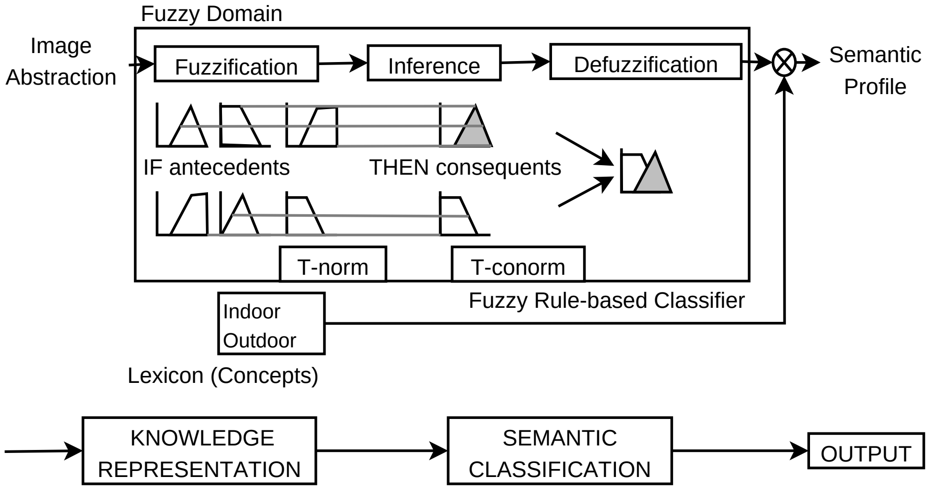


Fig. 5.1: An overview of the indexing unit under the fuzzy inference approach.

The outcome of the indexing unit is a list of concepts along with their degree of possibility to represent the image content. The list corresponds to the semantic image profile.

The complete indexing procedure is summarized in Tab. 5.2.

Tab. 5.2: Indexing procedure used in the fuzzy inference approach

```
indexing ( image, rules[ ] )
{
  abstractions[ ] = extract features( image )
  for each value in abstractions[ ] do
    fuzzification( value )
  inference( )
  defuzzification( )
  return semantic profile
}
```

5.1.3 Test Conditions

Data set used to evaluate this case study consists of over 3,000 distinct images extracted from TRECVID video repository (Smeaton et al. [153]). A single subject assigned the ground-truth for all images. Images were manually grouped into two categories namely *building* and *non-building*. Some samples of building images are shown in Fig. 5.2.

The TREC Video Retrieval Evaluation (TRECVID) benchmark provides an infrastructure for large-scale testing of retrieval technology. The TRECVID data set is comprised of approximately 170 hours of broadcast news video data from CNN Headline News, ABC World News Tonight and CSPAN. Currently, TRECVID addresses the problem of detection to 20 semantic concepts. For the task at hand, the following specification is given:

- *Building exterior* segment contains video of the exterior of a building.

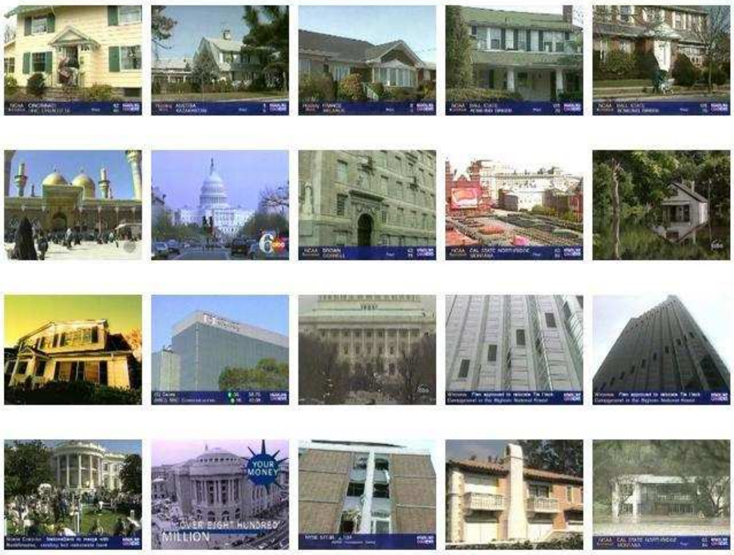


Fig. 5.2: Samples of building images extracted from TRECVID videos

5.1.4 Evaluation

The technical reference of this evaluation is summarised as follows:

Tab. 5.3: Summary of test conditions for case study I

Semantic profile	:	Single-element instance (only one index)
Lexicon	:	Building, Non-building
Approach	:	Fuzzy inference
Data set	:	TRECVID; 3,000 images
Feature vector	:	80-element Edge histogram descriptor

The fuzzy inference approach is used to index images belonging to categories *building* and *non-building*. The MPEG-7 edge histogram descriptor (See Sect. 3.2.1 on page 27) is used to represent the low-level image content. Fig. 5.3 highlights two samples of building images. In the first one, it is presented a close-up picture at which relevance of the building object is evident. The latter shows an open picture with different objects that not only introduce noise in the feature representation but also add more options to the content interpretation, i.e. street, cars.

A major aspect of edge histogram descriptor is its simplicity to represent local and global distribution of edges without expensive computation. It uses luminance mean values to detect edges, which is considered a weak approach. However, weakness of low-level description is compensated with contributions of problem domain knowledge.

According to semantics of edge histogram descriptor, five input variables are defined for the fuzzy model. Each input variable corresponds to an edge type category.

A compact fuzzy model is proposed associating a group of fuzzy sets with each input variable. Three fuzzy sets are used by default to simplify the model. Linguistic hedges are

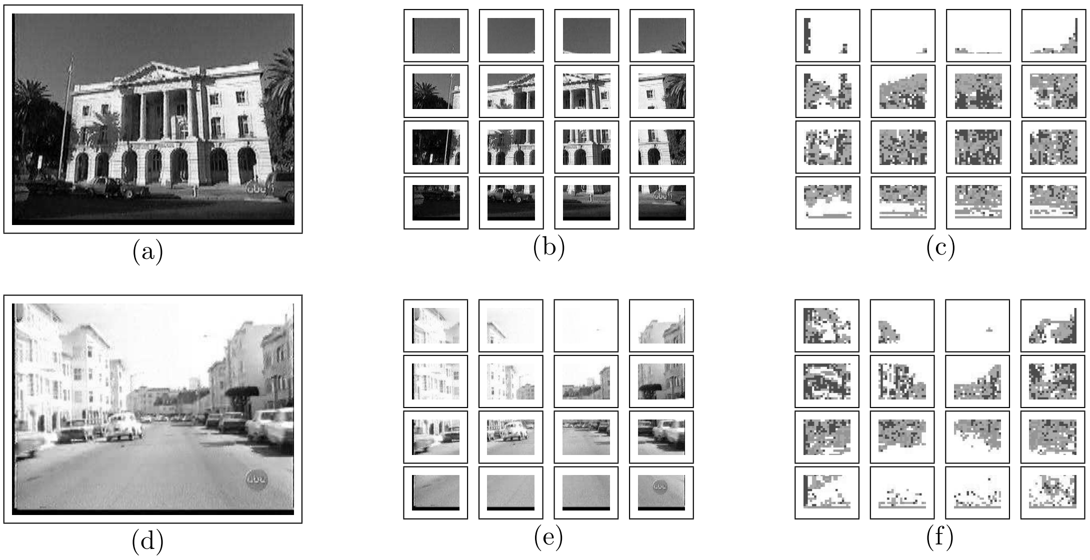


Fig. 5.3: (a) Close-up building image, (d) Open building image (b)-(e) Image decomposed into 16 sub-images, (c)-(f) Non-overlapping small square blocks

not applied on this model. Z-, Λ-, and S-shape functions are used to determine membership degrees (See App. C). Tab. 5.4 presents a summary of the fuzzy variables with their fuzzy sets and boundaries.

Tab. 5.4: Settings of fuzzy variables used to represent the edge histogram descriptor. Each fuzzy set namely Low, Medium, and High is defined by a membership function. These function have a domain between the interval determined by the corresponding boundaries b_1, b_2, b_3

Universe	Fuzzy set	b_1	b_2	b_3
Vertical Edge,				
Horizontal Edge,	Low	0.1	0.3	1.0
Diagonal 45°,	Medium	0.1	0.3	0.4
Diagonal 135°,	High	0.0	0.3	0.4
Non-directional				

The inference rules are organized into a multi-input single-output (MISO) rule space as is shown in Fig. 5.4. Each inference rule has an IF-part with five antecedents and a THEN-part with one consequent as follows:

$$\text{IF } e_1^v \text{ is } A_1^k \ldots \text{AND } e_5^v \text{ is } A_5^k \text{ THEN } y_j \ , \tag{5.3}$$

where e_i^v is an instance of an input variable (edge type), A_i^k is a linguistic term (fuzzy set name) used to transform values from a continuous to a discrete domain, and ℓ_j is a symbol labelling a semantic category. The lexicon consists of symbols *building* and *non-building.*

A sample of inference rules generated by the rule mining process is given in Tab. 5.5. Selected feature values extracted from the sub-images depicted in Fig. 5.5 and their corresponding fuzzy values are presented in Tab. 5.6. The possibility in this table indicates the score assigned by the system regarding to accuracy of the results. Both results achieved 1.0

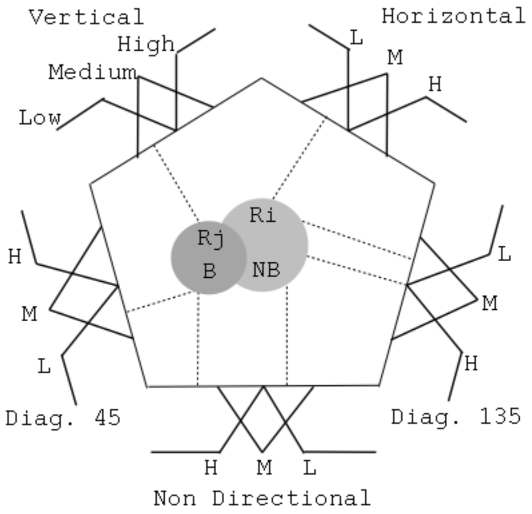
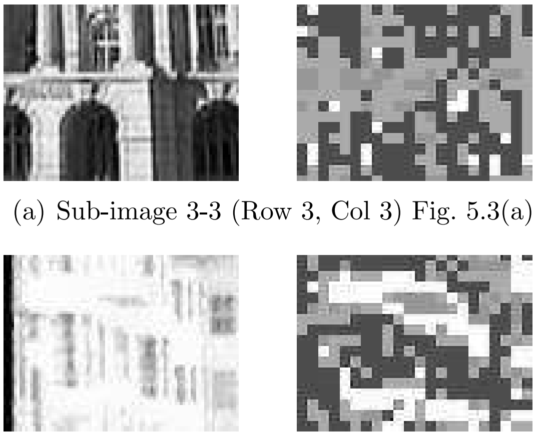


Fig. 5.4: Fuzzy variables associated with each type of edge are combined into a multi-input single-output space. R_j is an IF-THEN rule as is shown in Eq. 5.3. Fuzzy set names L, M, H stand for Low, Medium, and High

meaning that the system considers that these sub-images represent positively a building.



(a) Sub-image 3-3 (Row 3, Col 3) Fig. 5.3(a)

(b) Sub-image 2-1 (Row 2, Col 1) Fig. 5.3(d)

Fig. 5.5: Selected sub-images referenced in Tab. 5.6

Three different settings were evaluated:

- *A first run* (test 1) was performed to evaluate inference results on image descriptions limited to local analysis. As edge histogram descriptor decomposes original image into 16 sub-images, local analysis means classification of each single sub-image. This kind of analysis allows detection of parts of a building structure within the scene. Accuracy of inference is over 70%.

- *A second run* (test 2) was addressed to global analysis. It requires a minimum number of sub-images ascribed to the concept “building” before considering that the entire image contains a building.

Tab. 5.5: Example of weighted rules. V, H, D1, D2, ND, L, M, H, NB, and B stand for Vertical, Horizontal, Diagonal 45°, Diagonal 135°, Non Directional, Low, Medium, High, Non Building, and Building, respectively.

weight	**IF**					**THEN**
0.5	V is L	H is L	D1 is L	D2 is L	ND is L	NB
0.5	V is M	H is L	D1 is L	D2 is L	ND is M	NB
1.0	V is L	H is M	D1 is L	D2 is L	ND is L	NB
0.5	V is L	H is L	D1 is L	D2 is L	ND is M	NB
1.0	V is L	H is M	D1 is L	D2 is L	ND is H	NB
1.0	V is L	H is L	D1 is L	D2 is M	ND is H	B
1.0	V is L	H is L	D1 is L	D2 is L	ND is H	B
1.0	V is H	H is L	D1 is L	D2 is L	ND is M	B
1.0	V is M	H is M	D1 is L	D2 is L	ND is L	B
1.0	V is H	H is L	D1 is L	D2 is L	ND is L	B

Tab. 5.6: Feature and fuzzy values for the selected sub-images from pictures depicted at Fig. 5.3(a) and Fig. 5.3(d). Feature values (actual) appear below image IDs at 2nd and 6th column. Fuzzy values (membership degrees) are arranged below each fuzzy set's name. Two last rows indicated the expected (ground truth) and predicted (fuzzy inference) results along with the possibility of each prediction

Universe	33569	Low	Medium	High	45936	Low	Medium	High
Vertical	0.530	0.000	0.000	1.000	0.358	0.000	0.419	0.580
Horizontal	0.069	1.000	0.000	0.000	0.125	0.870	0.129	0.000
Diagonal 45°	0.123	0.882	0.117	0.000	0.068	1.000	0.000	0.000
Diagonal 135°	0.089	1.000	0.000	0.000	0.067	1.000	0.000	0.000
Non-Directional	0.166	0.668	0.331	0.000	0.108	0.956	0.043	0.000
Ground Truth	Building		Possibility		Building		Possibility	
Fuzzy Inference	Building		1.000		Building		1.000	

- *A third run* (test 3) was used to evaluate effects of varying the rule weights. These weights are real values ranging between 0 and 1. Relevance of the rule is determined to tune the knowledge representation of the problem domain.

Results are summarized in Tab. 5.7.

Tab. 5.7: Accuracy of inference process [%]

Test	Accuracy
1	70.05
2	83.95
3	86.31

Increasing the number of descriptor elements introduces a undesirable complexity on the fuzzy model because of the required number of input variables. In such cases, it is suggested to use the radial basis function network approach. There are mechanisms to transform RBFN into a fuzzy inference system (Jin and Sendhoff [154]). However, those mechanisms are not evaluated in this research work.

5.2 Case Study II: Radial Basis Function Network Approach

The semantic component indicates that some domain knowledge about the classification problem is available and can be used as part of the training procedures. The design of the classifier is addressed as a family of two-class (binary) classifiers. A binary classifier returns a Boolean variable indicating whether a given image belongs to class ω (Boolean 1) or is excluded from it (Boolean 0).

5.2.1 Learning Unit

Image abstractions (feature vectors) are grouped by concept and used as design patterns of each class. Classes are arranged on the schema one-against-all, it is to say, a concept will represent class ω at certain point and the rest of them class not(ω).

A number of *receptive fields* is formed around mean values of those groups of feature vectors by concept. In the sequel if a new feature vector is similar to the prototype the local activator (receptive field) should return 1 or a value close to it. On the contrary, if the image abstraction moves apart from the prototype, the activation level produces values close to zero.

There are numerous possibilities as to the choice of the receptive fields. Here, the receptive fields are computed as

$$u_{ij}(\mathbf{x}_k) = \left[\sum_{j=1}^{c} (\frac{d_{ik}}{d_{jk}})^{\frac{2}{m-1}} \right]^{-1} \quad , \tag{5.4}$$

where $u_i(\mathbf{x}_k)$ captures the matching level between the feature vector of new image $\mathbf{x}_k$ ($1 \leq k \leq N_X$) and the prototype of the i-th class (concept), $m > 1$ is a fuzzification exponent whose role is to adjust the shape of the receptive field, and $d^2(\cdot)$ is any distance norm expressing the low-level similarity formally defined as

$$d_{ik}^2 \doteq \|\mathbf{x}_k - \mathbf{v}_i\|_{\mathbf{A}}^2 = (\mathbf{x}_k - \mathbf{v}_i)^T \mathbf{A} (\mathbf{x}_k - \mathbf{v}_i) \ , \tag{5.5}$$

where $\mathbf{A}$ is the identity matrix for Euclidean distance or inverse of variance-covariance matrix for Mahalanobis distance and $\mathbf{v}_i$ ($1 \leq i \leq c$) is the prototype. It suggest the neuron model presented in Fig. 5.6.

As presented in Sect. 3.2.2, image abstractions consist of an aggregation of visual descriptors. Contributions of each descriptor to classify images vary with the classes, e.g. colour is suitable to characterize images belonging to the class *animal*. Then, weights are assigned to each component of the feature vector using a radial basis function network (RBFN) that captures information provided by professional annotators when define training patterns for the classes. The hidden layer with RBF activation functions

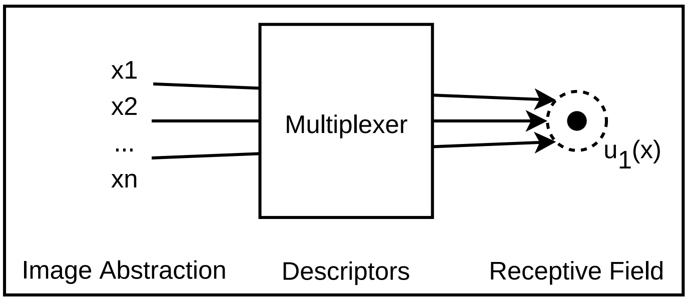


Fig. 5.6: Hidden neuron model. The output depends on the distance of the image abstraction (input vector) from the prototype. The multiplexer implements the support of multiple descriptors because the receptive fields are descriptor-dependent

$$u_1, u_2, \ldots, u_{N_D} \quad , \tag{5.6}$$

where N_D indicates the number of descriptors, is aggregated in an output layer with linear activation function

$$y = w_1 u_1 + w_2 u_2 + \ldots + w_{N_D} u_{N_D} \quad , \tag{5.7}$$

where $w_j (0 \leq j \leq N_D)$ are the weights assigned to determine the relevance of the descriptor regarding to the class ω. The learning procedure for all the training patterns at the same time can be written in matrix form as

$$\begin{bmatrix} u_1(\mathbf{x}_1) & \ldots & u_{N_D}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ u_1(\mathbf{x}_k) & \ldots & u_{N_D}(\mathbf{x}_k) \end{bmatrix} [w_1 \ldots w_{N_D}]^T = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \tag{5.8}$$

Defining the activation matrix by

$$\Phi = \begin{bmatrix} u_1(\mathbf{x}_1) & \ldots & u_{N_D}(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ u_1(\mathbf{x}_k) & \ldots & u_{N_D}(\mathbf{x}_k) \end{bmatrix} \tag{5.9}$$

Eq. 5.8 can be re-written as

$$\Phi \begin{bmatrix} w_1 \\ \vdots \\ w_{N_D} \end{bmatrix} = \begin{bmatrix} y_1 \\ \vdots \\ y_k \end{bmatrix} \tag{5.10}$$

Thus, the optimisation of the linear mapping is realized in a supervised mode by means of the standard pseudo-inverse method. The weights are computed by

$$[w_1, w_2, \ldots, w_{N_D}]^T = \Phi^{\ddagger} [y_1, \ldots, y_k]^T \tag{5.11}$$

The complete learning procedure is summarized in Tab. 5.8.

Tab. 5.8: Learning procedure used in the RBFN approach

```
learning ( sample images[ ], concepts[ ], descriptors [ ] )
{
  abstractions[ ] = extract features( sample images[ ] )
  groups[ ] = group ( abstractions[ ], concepts[ ] )
  for each group in groups[ ] do
  {
   mean[ ] = mean( )
   stdev[ ] = standard deviation( )
   for each descriptor in descriptors[ ] do
   {
    compute receptive fields( group, descriptor )
    define activation matrix
    weights[ ] = pseudo-inverse matrix( )
   }
  }
  return RBFN model
}
```

5.2.2 Indexing Unit

The aggregation procedure of multiple descriptors is possible because the topology of the original descriptor elements is kept in the feature vectors (See Fig. 5.6 and Sect. 3.2.2).

RBFN architecture acts as an inference engine using the neuron model. In this way, semantic profiles are created by attaching class labels to visual abstractions.

The multi-class classification relies on results obtained through the collection of two-class classifiers as is illustrated schematically in Fig. 5.7. The outcome of multi-class classifiers is combined into a single set of concepts or semantic profile.

The complete indexing procedure is depicted in Fig. 5.8 and summarized in Tab. 5.9.

Tab. 5.9: Indexing procedure used in the RBFN approach

```
indexing ( image[ ], RBFN model )
{
  abstractions[ ] = extract features( image )
  for each concept in rbfn model do
  {
   get activation levels
   remove concepts ( threshold )
  }
  return semantic profile
}
```

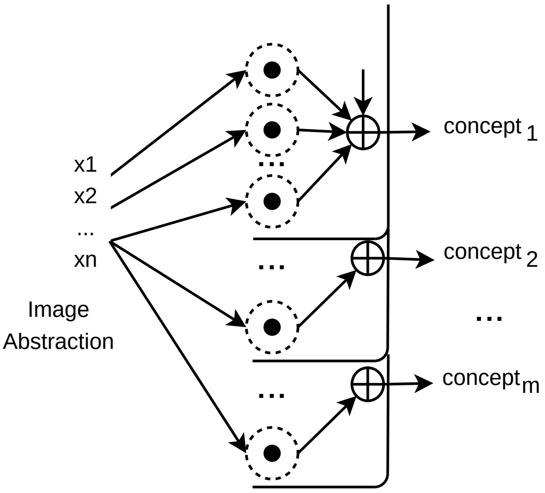


Fig. 5.7: RBFN and linear activation levels. Binary classifiers are aggregated to create the semantic profile

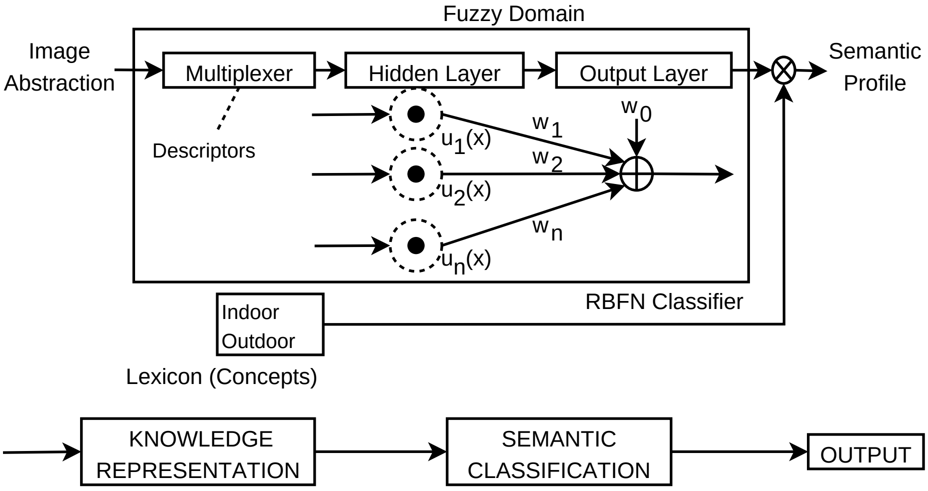


Fig. 5.8: An overview of the indexing unit under the RBFN approach.

5.2.3 Test Conditions

Experimental material used to evaluated this case study consists of 1,000 colour images of different size collected from Corel stock gallery (ref [corel.com]) and download from the Web (ref [freefoto.com]). Images were manually grouped into five categories namely *animal*, *building*, *city view*, *landscape*, and *vegetation*.

To underline the diversity of the graphic material, Fig. 5.9 includes some sample images used in the experimental studies. It becomes noticeable that while the categories under discussion seem to be quite distinct, in virtue of the very nature of visual information, it is very likely that some images could be correctly ascribed to semantic profiles containing simultaneously several categories, say an *animal* being a part of some *landscape*.

The subjective criterion to assign an image to certain category is the match between the object on which the camera is primarily focused and the category name. It is assumed that this criterion helps to determine the relevant object in the scene. Then, relevance is decided

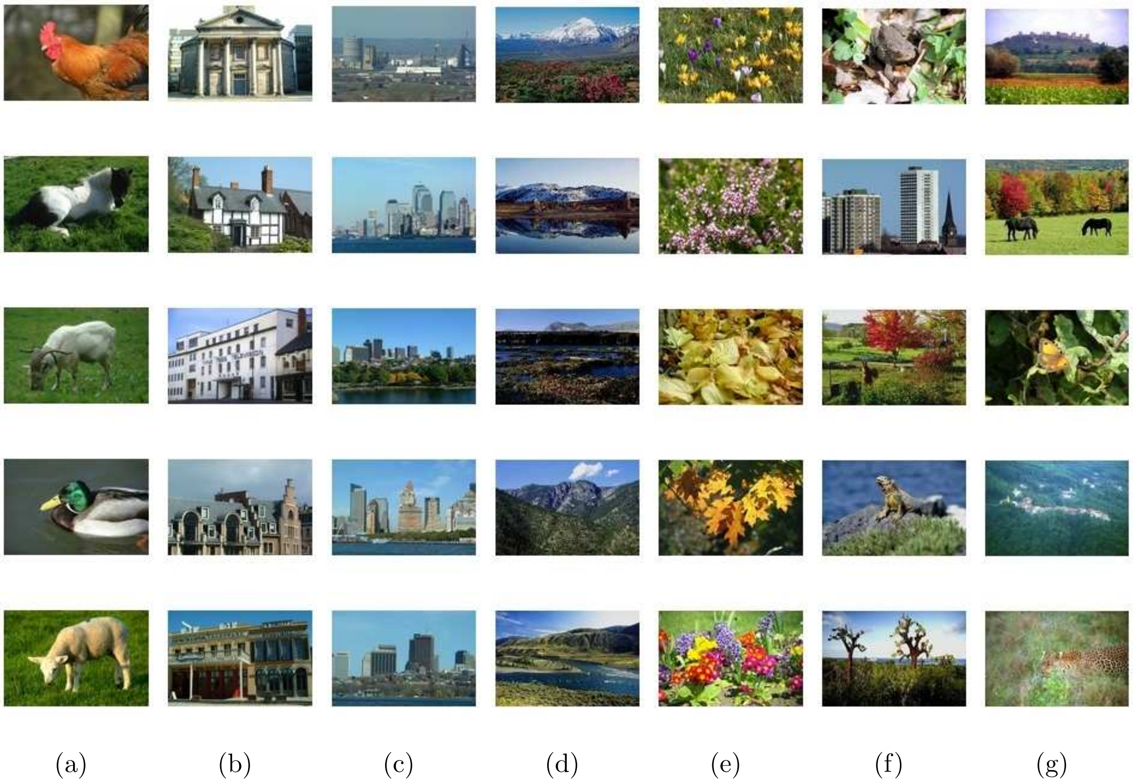


Fig. 5.9: Sample images used in the study; (a) Animal, (b) Building, (c) City view, (d) Landscape, (e) Vegetation; note that combinations of several categories occurring in the same image are considered; this becomes evident in case of images shown in (f) and (g)

looking at region occupied by the object and its spatial location, i.e. centred and in the foreground. It is also supposed that these characteristics obey an interest of the photographer of capturing such an element in the scene. The following criteria was used to assign images into categories:

- *Animal*, a scene falls into this category when an animal of a visible size appears in the picture. One shortcoming in this type of images is the mimetic properties of some species.

- *Building*, a scene is tagged with this concept when a building structure of a visible size appears in the picture. Two shortcomings are identified: the different types of buildings, i.e. castles, residential houses, warehouses, religious facilities, etc.; and the different distances used in the close-ups.

- *City view*, a scene containing panoramic views of cities, specifically buildings, is assigned to this category. The different angles and small size of the man-made structures, among others, are examples of shortcomings to categorize these scenes.

- *Landscape*, a scene is placed in this category when the picture depicts scenery with natural elements such as mountains, forest, seashores, etc. No man-made objects are expected to be in this type of images.

- *Vegetation*, a scene belongs to this category when nature, i.e. plants, appears in the picture.

5.2.4 Evaluation

The technical reference of this evaluation is summarised as follows:

Tab. 5.10: Summary of test conditions for case study II

Semantic profile	:	Single-element instance (only one index)
Lexicon	:	Animal, Building, City View, Landscape, Vegetation
Approach	:	Radial basis function network
Data set	:	Corel Stock and FreeFoto; 1,000 images
Feature vector	:	282-element Colour and texture descriptor

The RBFN approach is used to index images belonging to the five categories described above. 282-element feature vectors are used to represent the low-level image content (See Sect. 3.2.2 on page 28).

The fuzzification factor (exponent) used in the development of the receptive fields (See Eq. 5.4 on page 56) requires some attention. A suite of experiments was completed to quantify the effect of its value on the performance of the RBFN classifier. Fig. 5.10 shows that changes in the values of the fuzzification factor (m) produce a little impact on the performance.

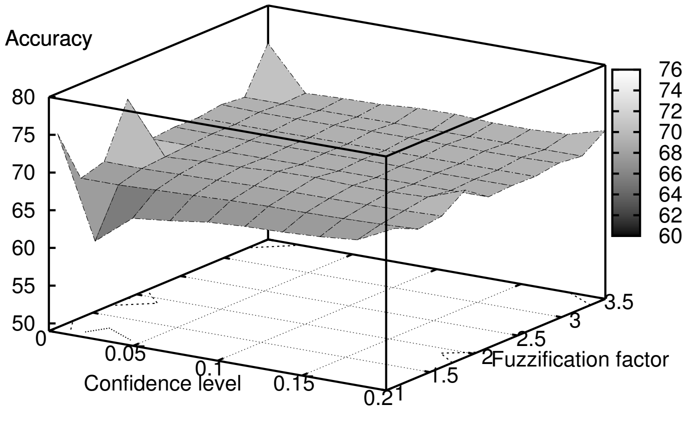


Fig. 5.10: Effects of the fuzzification factor on the RBFN classifier performance

Practically, to avoid unnecessary and quite tedious optimisation (as it cannot be completed with the use of gradient-based techniques), it is advisable to fix the value of the coefficient as 2.0. This value is the one typically used in most applications of the FCM clustering algorithm [141].

Given the multi-class problem, the performance achieved for the training and testing data sets on the MPEG-7 learning space are shown in Fig. 5.11 along with the details collected in a tabular fashion at Tab. 5.11. Accuracies obtained with concatenated descriptions are higher in all cases.

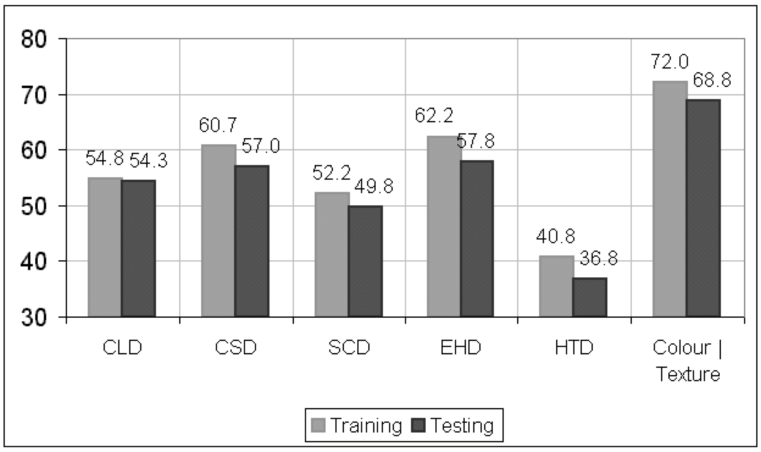


Fig. 5.11: Performance obtained when using an RBFN approach with either separated or concatenated descriptor elements as feature vectors

Tab. 5.11: Confusion matrix when using an RBFN approach with concatenated descriptor elements as feature vectors

	Training (%)					Testing (%)				
Category	A	B	C	L	V	A	B	C	L	V
(A) Animal	**59.17**	1.67	8.33	0.00	17.50	**41.25**	2.50	6.25	0.00	11.25
(B) Building	5.00	**78.33**	14.17	0.00	1.67	15.00	**75.00**	12.50	0.00	1.25
(C) City View	8.33	5.00	**51.67**	2.50	0.83	8.75	8.75	**50.00**	2.50	0.00
(L) Landscape	4.17	11.67	24.17	**92.50**	1.67	10.00	12.50	25.00	**95.00**	5.00
(V) Vegetation	23.33	3.33	1.67	5.00	**78.33**	25.00	1.25	6.25	2.50	**82.50**
Accuracy	72.00					68.75				

The confusion matrices deliver a detailed quantification of the distribution of the classification error. These matrices confirm the expected semantic conflicts in separating certain categories (See Sect. 4.2 on page 42). In particular, the categories come in the following pairs: *animal-vegetation*, *building-city view*, and *city view-building/landscape*. Overlapping between groups is solved choosing the highest level of matching as categorization outcome. This final stage step in the RBFN approach is presented in Fig. 5.12.

Differences between RBFN performance for the two-class and multi-class problems are presented in Fig. 5.13. The two-class problem reports acceptable results, i.e. above 82%. In contrast, performance in the multi-class problem is reduced dramatically to a range between 60% and 75%.

While the results reported so far are primarily concerned with the quantitative aspects of the approaches, it is equally important to discuss its qualitative facet. In particular, it is of

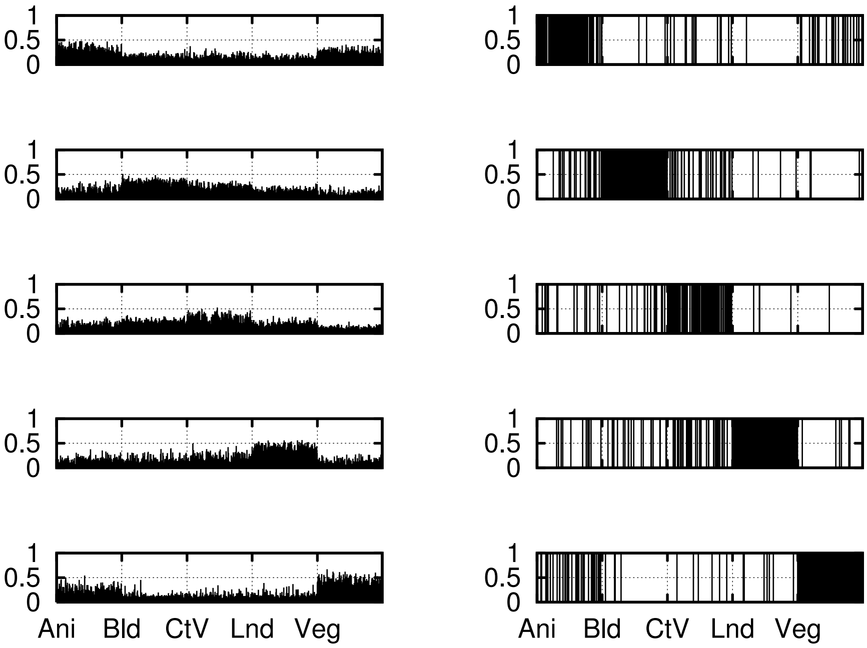


Fig. 5.12: Final stage in the RBFN approach to solve the overlapping between categories

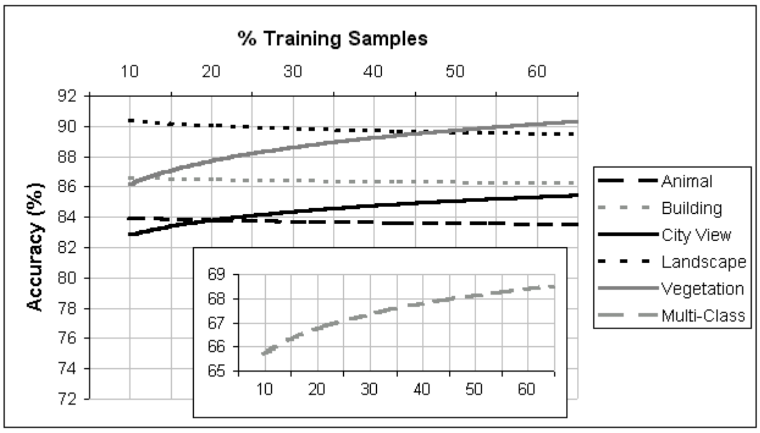


Fig. 5.13: Performance of the RBFN approach for the two- and multi-class problems

interest to learn what images are tagged with a wrong semantic profile, why, and what could help improve the procedures.

The following remarks are derived of observations obtained from outcomes of approaches presented above and corresponding images as the ones presented in Fig. 5.14 and 5.15. As can be observed, images do not appear in ascending nor descending order. It is because the classifier does establishes a ranking with the images. The score is obtained from the binary classifier that obtained the higher outcome.

Animal images are correctly categorized when there is a clear contrast between the background and the foreground. Otherwise, the salient features for these images are poorly discriminated and the semantic profiles are mainly instantiated as *vegetation*.

The prediction of building images becomes satisfactory with feature vectors containing strong texture orientation. One shortcoming in the manual labelling was the differentiation between *city view* and *building* based on the close-up. Some of these building images were placed into the *city view* group.

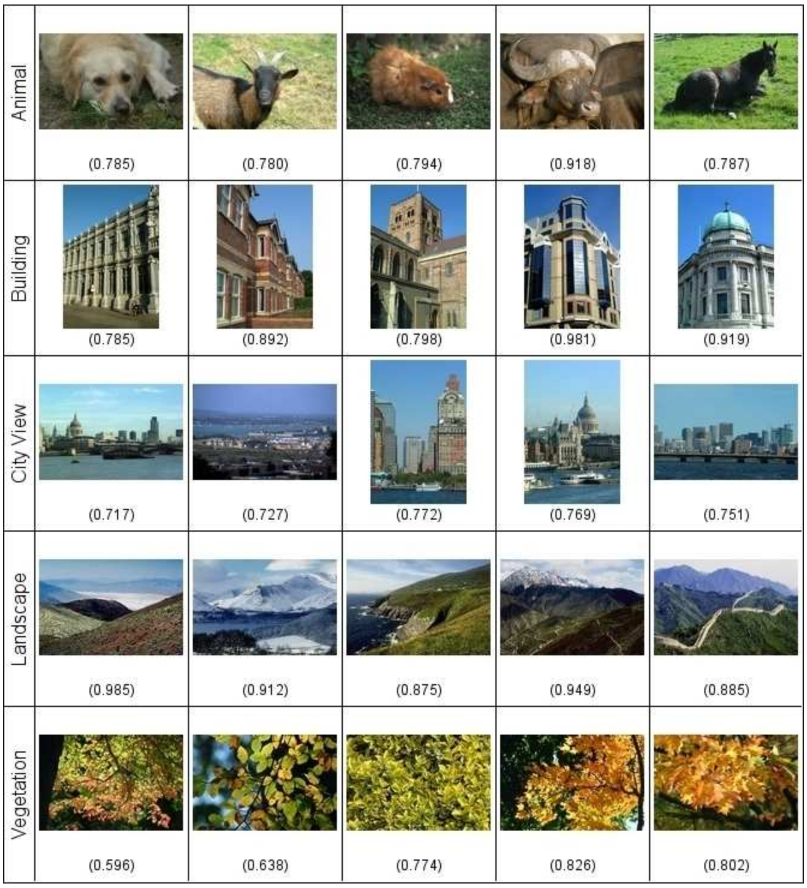


Fig. 5.14: Samples of semantic profiles correctly assigned. Classification score is indicated below each image

The texture and colour descriptors allow certain discrimination between *landscape* and *city view* images. However, there are images that even to the human observer are semantically similar.

The same drawback happens when looking at the group of images containing concepts *landscape* and *vegetation*. The close-up criterion was a certain shortcoming with other groups. For instance, most of the images belonging to group *vegetation* with objects filling almost the whole area of the picture were wrongly categorized.

5.3 Summary

Two classification approaches have been evaluated according to availability of training images and data size, namely fuzzy inference and radial basis function network-type of classifier.

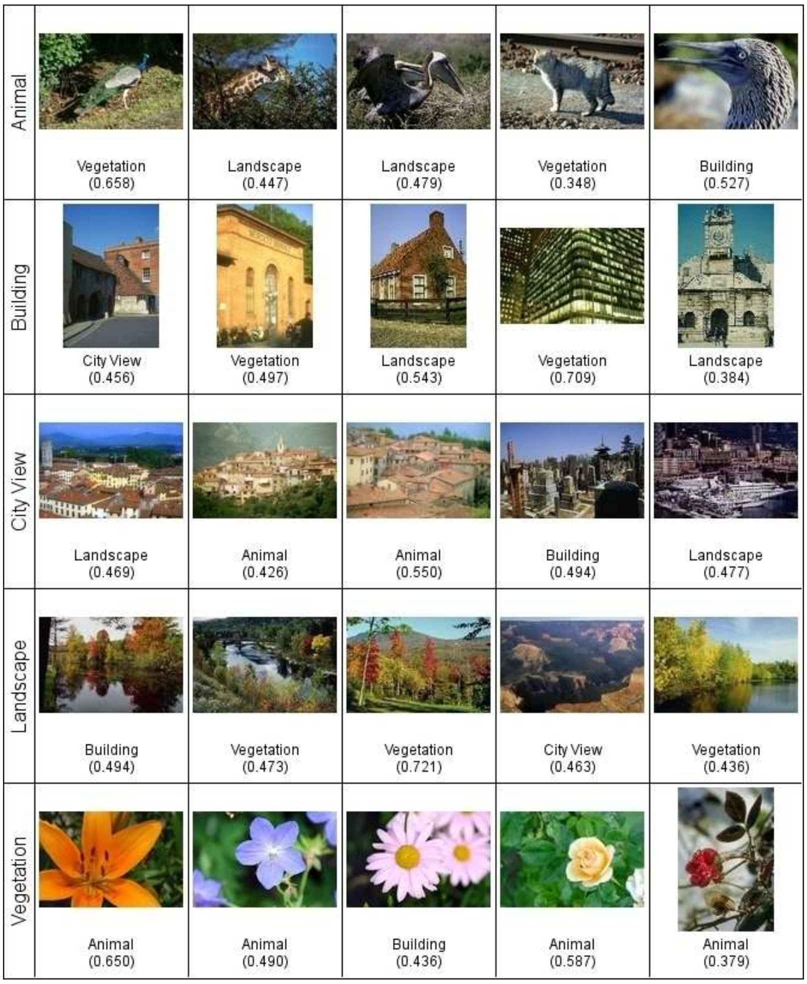


Fig. 5.15: Samples of semantic profiles wrongly assigned. Classification score and assigned concept are indicated below each image

There is a suitable groundwork to apply the fuzzy inference approach when the number of variables is low and the availability of exemplars is high (above 30% of the current data). Otherwise, as presented in the case study I, it becomes too computational expensive and less effective to build the classifier.

On the other hand, known classes and availability of exemplars for each of them, establishes a convenient groundwork for RBFN-type classifiers. However, a shortcoming appears when there is high variability intra-class. It could be overcome by defining specialised receptive fields that serve as activation function of certain groups of images within the same class. It derives in deeper levels of granularity, i.e. adding more neurons or layers, and relates the issue of partitioning the feature space.

Considering these limitations along with full supervised learning mode in which the above approaches operate, a partially supervised clustering approach is introduced to the framework. Details on the design of the framework are presented in the following chapter.

Chapter 6

A Framework for Concept-related Indexing of Image Content Phase Two: Design and Evaluation

The proposed framework takes advantage of fuzzy sets theory (Klir and Folger [155]) to induce a classifier model for concept-related indexing of image content. The framework creates a built-in knowledge base to store interpretations of the image content.

Sets of concept-related indexes (named *semantic profiles*) are proposed to connect image abstractions to interpretations. Semantic profiles are inexact in nature similar to human interpretations.

Each time new labelled images are added to the database, new concepts can emerge from the constant evaluation of the semantic profiles, leading to continuously changing of knowledge base. The updating task is carried out by a *learning unit*. Accordingly, new unlabelled images activates an *indexing unit* in order to create a semantic profile for them.

A framework overview is given in Sect. 6.1. Afterwards, a fuzzy clustering approach used to build the knowledge base and generate semantic profiles for new images is described. A salient feature of this approach is its capability to approximate human-like reasoning. Sect. 6.2 is devoted to explain the semantic profiles. It also presents the proposed matching procedure to perform comparison image-image, image-cluster, and cluster-cluster combining low- and high-level information.

Second part of the chapter reports experimental studies carried out to evaluate performance of the framework. Sect. 6.3.1 describes test conditions. Finally, Sect. 6.4 summarizes the chapter.

6.1 Framework Overview

The proposed framework automatically assigns concepts, from a pre-defined lexicon, to images combining granules of information at different abstraction levels (ref to 2.1, Fig. 2.1). It creates a semantic profile to describe the image content. Such a description appertains to the machine interpretation of the low-level features, which relies on the experience accumulated by the system. The framework maintains a built-in knowledge base to store interpretations of the image content. The framework consists of two main units: learning and indexing.

6.1.1 Learning Unit

The *learning unit* consists of three sequential modules: feature extraction, knowledge representation, and semantic profile generation. Fig. 6.1 illustrates the data flow of this unit.

The inputs are labelled and unlabelled training images and visual interpretations with regard to the content of the labelled images. A professional annotator provides interpretations. The feature extraction module generates compliant image abstractions with the MPEG-7 standard. Low- and high-level information is combined into a learning space.

Problem domain knowledge is introduced through a controlled lexicon, which contains the concepts related to the image content. The knowledge representation module uses a partially

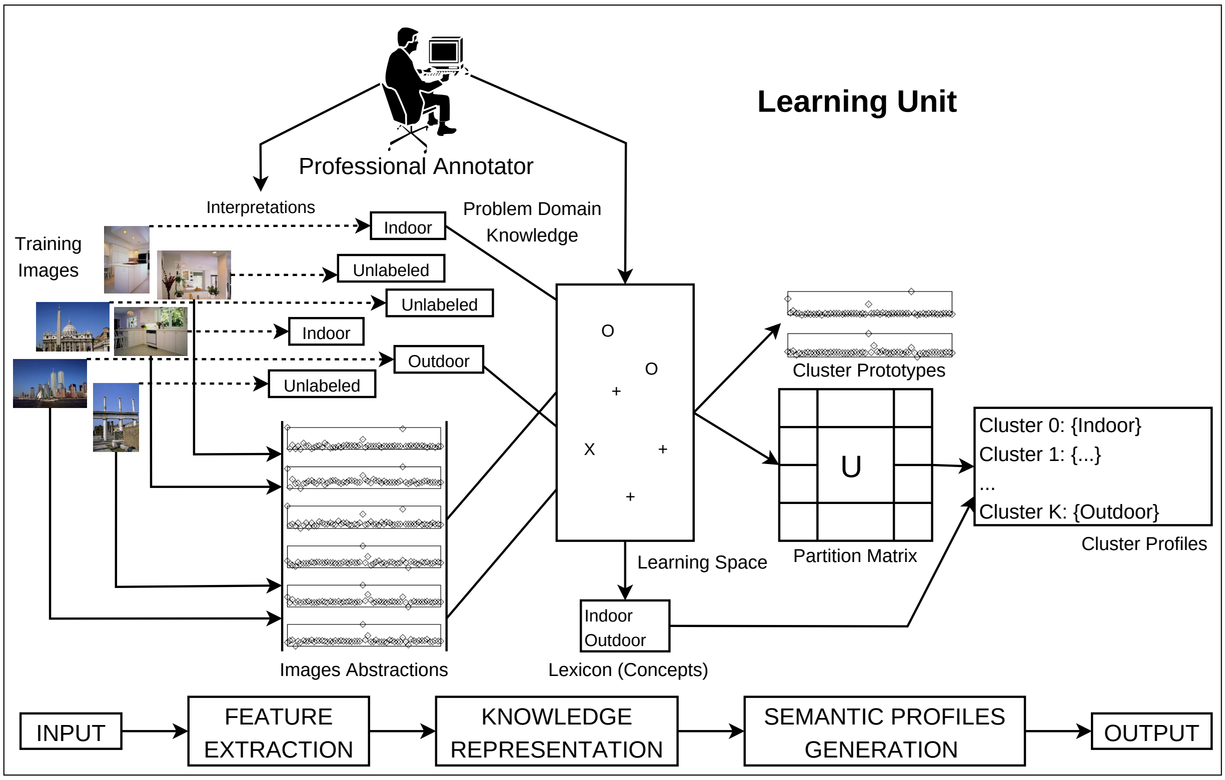


Fig. 6.1: Overview of the learning unit. The first module extracts a set of low-level features in compliance with the MPEG-7 syntax and builds a learning feature space. The second module relays on human knowledge to produce suitable links between low-level features and high-level concepts on a limited training set. The lexicon is used to control the signification space. The third module generates semantic cluster profiles

supervised clustering algorithm to create a built-in knowledge base.

Clustering is also applied to generate a suitable partition of the feature space. Based on the clustering outcomes summarised in a partition matrix, each cluster is labelled with a number of concepts in order to create semantic cluster profiles. These profiles constitute the outcome of the learning unit.

The complete learning procedure is summarized in Tab. 6.1. Some details are added to previous description of the procedure:

`Line 6`, unsupervised clustering is used to create an initial partition of the learning space, though there are available labelled images. The reason is to present the professional annotator an insight of underlying relationships that could affect the desired semantic grouping.

`Line 9`, a preliminar semantic profiling is generated prior structural data analysis. `Line 12-13` compute cluster-class dependencies and identify class pairs (overlapping) based on labelled images.

`Line 14` captures information from professional annotator to set up the partially supervised clustering algorithm (e.g. refined labels). Another important setting parameter is the number of elements to be used in instantiating the profiles. It is to say, single- or multiple-element instance. In case of the latter, it is suggested to indicate the expected number of elements to refine the profiling. `Line 17` relates the iterative component of the annotation process.

Tab. 6.1: Learning procedure used in the clustering approach

```
1    learning ( sample images[ ], concepts[ ] )
2    {
3      read interpretations
4      for each image in sample images[ ] do
5        abstractions[ ] = extract features( image )
6      clusters[ ] = generate unsupervised learning space
7      for each cluster in clusters[ ] do
9        compute semantic profile ( clusters[ ], concepts[ ] )
10     for each concept in concepts[ ] do
11     {
12       compute cluster-class dependencies ( )
13       display structural data analysis ( )
14       settings = capture annotator’s feedback ( )
15     }
16     clusters[ ] = generate semi-supervised learning space
17     if refine space then go to Line 10
18       compute semantic profile ( clusters[ ], concepts[ ] )
19     return semantic cluster profiles
20   }
```

6.1.2 Indexing Unit

The *indexing unit* automatically assigns concepts to new images using the knowledge acquired by the learning unit. The feature extraction module extracts low-level features of unlabelled images. Afterwards, the knowledge representation module maps image abstractions into the corresponding domain, i.e. fuzzy domain. The semantic classification module relies on the built-in knowledge base to generate the semantic profile for the image. Therefore, the indexing unit consists of three sequential modules: feature extraction, knowledge representation, and semantic classification. The knowledge representation module provides the classifier with a lexicon to control the signification space. Fig. 6.2 portrays components and data flow of the indexing unit.

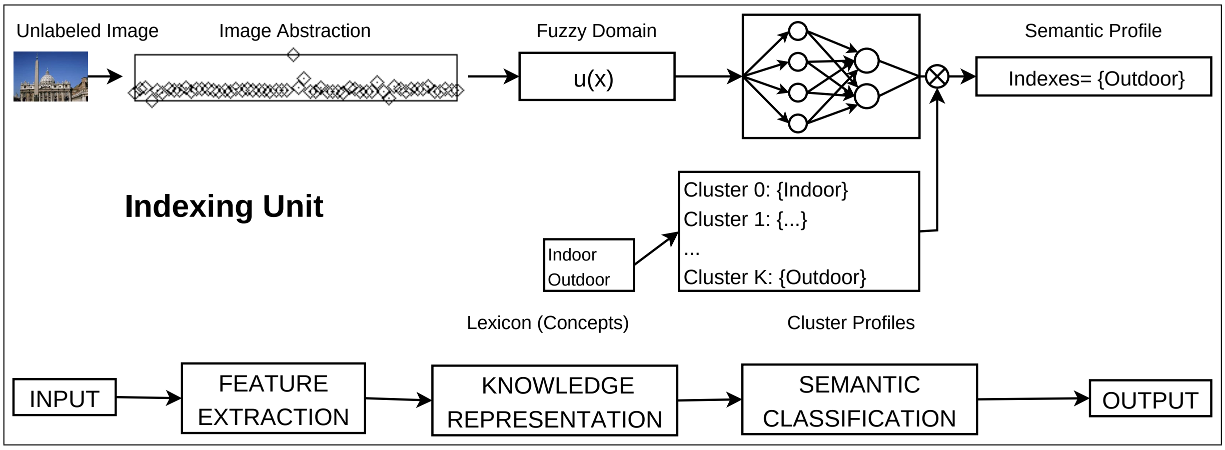


Fig. 6.2: Overview of the indexing unit. The first module extracts a set of low-level features in compliance with the MPEG-7 syntax and builds a learning feature space. The second module provides the classifier with a lexicon to control the signification space. It also guides the clustering algorithm towards a semantic grouping of incoming images. The third module generates semantic image profiles

The inputs are unlabelled images. The feature extraction module generates compliant image abstractions with the MPEG-7 standard. The knowledge representation module uses information on learning space partition to find a group for the incoming image. Clustering-based classification defines degrees of membership to each cluster and create the sematic profile for the new image.

The complete learning procedure is summarized in Tab. 6.2.

Tab. 6.2: Indexing procedure used in the clustering approach

```
indexing ( image, cluster prototypes[ ] )
{
   extract feature vector ( image )
  for each prototype in clusters prototypes[ ] do
   compute membership ( )
  create semantic profile ( )
  return semantic image profile
}
```

6.2 Semantic Profiles

6.2.1 Concept Profiling

A semantic profile consists of one or more concept-related indexes. These profiles can be defined by a *single-element instance* (one and only one index) or a *multiple-element instance* (two or more indexes).

Initially, learning space is partitioned into clusters approximating semantic groups. Then, occurrences of concepts attached to training images are summarized by cluster. A weight is assigned to each concept considering the membership degrees of the images within the cluster.

In case of having multiple-element instances in the semantic images profiles, position of the concept is also involved in computing the weight. In order to simplify this step, membership degrees are divided by the concept's position.

$$w_{jk} = \sum_{1 \leq i \leq N_X} \frac{\mu_k(\mathbf{x}_i)}{\mathrm{pos}(\ell_j, \mathbf{x}_i)} \ , \tag{6.1}$$

where w_{jk} is the weight of concept ℓ_j in cluster k, μ is the membership degree of feature vector $\mathbf{x}_i$ to cluster k, and $\mathrm{pos}(\cdot)$ returns the position of a concept within the list of concepts tagged to $\mathbf{x}_i$.

Subsequently, weights are normalized and concepts sorted in descending order by weight. The preliminar profile of cluster k is as follows:

$$\mathcal{L}_k = \left\{ \ell_j \mid \underset{\substack{r<j \\ s>j}}{\forall} w_{rk} \leq w_{jk} \leq w_{sk} \right\} \ . \tag{6.2}$$

In order to eliminate low-ranked concepts a threshold criterion (δ) is applied.

$$\underset{\ell_j \in \mathcal{L}_k}{\forall} w_{jk} \geq \delta \ . \tag{6.3}$$

Tab. 6.3 summarizes the semantic profiling procedure.

Tab. 6.3: Semantic profiling procedure

Given	$X = X^d \cup X^u$, learning space that contains labelled and unlabelled data
	$N = N_d + N_u$, number of feature vectors
	$L = \{\ell_1, \ell_2, \ldots, \ell_{N_L}\}$ is a lexicon
	c, number of partitions
	δ, threshold criterion
Step 1	Partition the learning space
Step 2	Rank images within each partition
Step 3	Compute concepts' weights using Eq. 6.1
Step 4	Normalize weights and create cluster profile (Eq. 6.2 and Eq. 6.3)
Step 5	Return semantic cluster profiles $\mathcal{L}_k$ $(1 \leq k \leq c)$

Tab. 6.4 presents a sample of semantic profiles obtained after applying clustering approach onto a descriptor space that contains feature extracted from images that can be grouped into five classes.

Note that concept *landscape* in cluster “3” occurs in the 22.3% of the images and its relevance is high. Knowing, that number of images of each class is the same, it indicates that most of the best-ranked images belong to that class.

In contrast, concept *city view* in cluster “0” occurs in almost same percentage of images but its relevance is lower. It indicates that images ascribed to that class have low membership degrees (ranking) within the cluster.

Profile of cluster “4” consists of concepts *building* and *city view*, which reports an overlapping between this class pair. Membership degrees are low as indicator of either high intra-class variability or low feature discrimination.

Tab. 6.4: Sample of cluster profile. Test conditions are: five classes and equal number of clusters, and threshold criterion $\delta = 0.8$. Percentage of ascribed images to the cluster is indicated. The number besides each concept corresponds to its relevance within the cluster

ID	% images	Cluster Profile
0	21.9	city view (0.462)
1	15.1	animal (0.541)
2	20.0	vegetation (0.702)
3	22.3	landscape (0.729)
4	20.7	building (0.369), city view (0.283)

Once the semantic profile is assigned to each cluster, the next step is to define a procedure to estimate distances combining low- and high-level matching.

6.2.2 Proposed Matching Procedure for Performance Evaluation

As the framework embeds classification approaches, several interesting classification scenarios may arise. Assuming ω as the true class or expected concept, the classifiers may produce the following outcome

- Image is ascribed to only one class. There are two options: the class is the same as ω –no classification error, the class is different –misclassification.
- Unknown image. The classifier does not identified a class (that is it has produced a classification decision of not(ω)).
- Image is ascribed to several classes; say ω_r, ω_s, etc. ($1 \leq r, s \leq c$) This is described as lack of specificity of classification. Under these circumstances, two situations may occur. First, ω is within the set of these classes. The result is correct but not specific. Second,

ω is not in these classes being identified by the classifiers. In this case the classification result is neither correct nor specific.

However, the proposed concept-related indexing cannot be reduced to typical classification outcomes, though it incorporates classification mechanisms. Furthermore, performance evaluation of the index assignments relates the issue of semantic matching constraints.

Matching constraints take into account the fact that, firstly, the incoming image may match multiple entries in the lexicon and, secondly, the matching that do occur may be imperfect or incomplete.

In contrast to typical procedures used to determine performance of a classifier, the framework requires estimation of matching (or mismatching) between the expected and assigned semantic profiles.

Expected and assigned semantic profiles can differ not only in their content, but also in their size (number of concepts). Bringing methods from text matching, two procedures were evaluated:

- *Changes-based matching.* This procedure is based on the Levenshtein distance ([156]). It counts the number of deletions and insertions require on the assigned profile to be identical to the expected profile. Concept's position is not considered by this procedure. The level of matching is reported as the number of required changes divided by the number of concepts on both profiles.

- *Position-based matching.* Search for the occurrence of the concept starting at first position and considers the position of it within the list. As aforementioned, position is assumed as indicator of relevance. It is to say; a concept at the first position is more relevant for interpreting the image content than another close to the back of the list (semantic profile).

In addition, exact matching is applied at concept level. It is to say, any prefix/suffix variation or case sensitive change is assumed as mismatch. The procedures are illustrated in the following example:

Given the expected profile {*sky, water, building, park*} and assigned profile {*clouds, park, tree, building, mountain*}:

- Changes-based matching detected three deletions (*clouds*, *tree*, *mountain*) and two insertions (*sky* and *water*). The final result is $\frac{5}{(4+5)} = 0.56$.

- Position-based matching penalizes not only the absence of concepts, but also the position. The result is 0.68 meaning high mismatching (Tab. 6.5).

Tab. 6.5: A sample of position-based matching procedure

	1	2	3	4	5	Minimum
Profiles	clouds	park	tree	building	mountain	distance
1 sky	1.00	1.00	1.00	1.00	1.00	1.00
2 water	1.00	1.00	1.00	1.00	1.00	1.00
3 building	1.00	1.00	1.00	0.17	1.00	0.17
4 park	1.00	0.40	1.00	1.00	1.00	0.40
Min. distance	1.00	0.40	1.00	0.17	1.00	0.68

Experimental results demonstrated that proposed position-based procedure reports more accurate levels of matching between expected and assigned profiles than the changes-based matching procedure.

Incorporating low-level matching enhances position-based matching. The complete matching procedure is presented in Tab. 6.6.

Tab. 6.6: Proposed matching procedure to compare semantic profiles

Given	$\mathbf{x}_i$ $(1 \leq i \leq N_{\mathbb{X}})$ a feature vector associated with the i-th image
	$\mathbf{v}_j$ $(1 \leq j \leq c)$ a cluster prototype
	L a lexicon
	$\mathcal{L}_i = \{l_r \mid l_r \in L \;\; (1 \leq r \leq n_r)\}$ the semantic profile of $\mathbf{x}_i$
	$\mathcal{L}_j = \{l_s \mid l_s \in L \;\; (1 \leq s \leq n_s)\}$ the semantic profile of cluster c_j
Step 1	Compute low-level matching using Eq. 6.4
Step 2	Compute high-level matching using Eq. 6.5
	where $d_{r,s}$ is calculated applying Eq. 6.6
Step 3	Compute distance $d_{i,j}$ using Eq. 6.7

$$d^2_{low}(\mathbf{x}_i, \mathbf{v}_j) \doteq \|\mathbf{x}_i - \mathbf{v}_j\|^2_{\mathbf{A}} \tag{6.4}$$

$$d_{high}(\mathbf{x}_i, \mathbf{v}_j) = \frac{\sum_r^{n_r}\left(\min_{1\leq s\leq N_s}(d_{r,s})\right) + \sum_s^{n_s}\left(\min_{1\leq r\leq N_r}(d_{r,s})\right)}{n_r + n_s} \tag{6.5}$$

$$d_{r,s} = \begin{cases} 1 & l_r \neq l_s \\ \begin{cases} 0 & r = s \\ (r-s)/(n_r+1) & r > s \\ (s-r)/(n_s+1) & r < s \end{cases} & \text{otherwise} \end{cases} \tag{6.6}$$

$$d_{ij} = d_{low} * d_{high} \tag{6.7}$$

The proposed matching procedure satisfies metric axioms:

- Level of matchings are non-negative.

- Level of matching is zero only if the features and profiles are identical (including position of concepts within the annotation list).
- Level of matching between two objects (i.e. cluster prototype or image) are symmetric.

Using the matching procedure as metric and adding it to the partitioned space, the result is a metric space. Then, the framework is provided with a learning space that can be formally summarised as

$$< X, \mathbf{V}, L, \mathcal{L}, d > \quad , \tag{6.8}$$

where X is a set of p dimensional feature vectors, $\mathbf{V}$ is a set of cluster prototypes, L is a controlled lexicon, $\mathcal{L}$ is the subset of concepts used to create the semantic cluster profiles, and d is the metric for computing low- and high-level matching.

6.3 Evaluation

6.3.1 Test Conditions

To carry out the assessment of the framework environment in a comprehensive and meaningful manner, the test conditions has been carefully organized. In particular a number of design scenarios that help to quantify some interesting combinations of the framework's settings have been included. Experimental data consisting in two data sets collected from large image collections: Corel stock gallery, and Freefoto.com.

The training and testing data sets used were randomly generated. Percentage of training samples is ranged from 10% to 30% of all images and the remaining is allocated for testing.

The first data set is described in Sect. 5.2.3. The second data set is enlarged by adding more images from Corel stock gallery. In order to limited the context of the image annotations, experiments were tailored to images containing the concept "building" or "buildings". Over 3,000 were selected from the Corel categories listed in Tab. 6.7.

Tab. 6.7: Corel categories containing "building" images

Agriculture	Nature
Architecture	Plants
Art	Religion
Cities and Regions	Travel
Government	Urban
Historic and Vintage	

In contrast to the first data set, this one consists of around 750 concepts. Tab. 6.8 shows some samples of concepts.

Tab. 6.8: Samples of some concepts appearing in the third dataset

Apartment	Auditorium	Cityscape	Downtown	Entrance
Farms	Gardens	Hill	Houses	Inn
Jungle	Korean	Landmark	Mall	Mosque
Newspaper	Ocean	Pantheon	Racks	Railroad
Seashore	Shop	Table	Tavern	Temples
Tower	Umbrella	University	Valley	Vegetation
Walkway	etc.			

6.3.2 Hierarchical Clustering

Hierarchical clustering methods were evaluated to compare against the partitioning method applied in this study. Specifically, the clustering using representatives (CURE) algorithm proposed by Guha et al. [157] was used because its capabilities to deal with outliers and work on large-scale and multidimensional data sets.

CURE selects representative points for each cluster that are generated by selecting well scattered points from the cluster and then shrinking them toward the center of the cluster by a specified fraction. This enables CURE to adjust the geometry of clusters having non-spherical shapes and wide variances in size. Using random sampling, CURE handles large data sets efficiently.

A main drawback found when applying hierarchical clustering is the difficulty to obtain a desirable partition. Fig. 6.3 illustrates this situation depicting a dendrogram in which semantically meaningful groups are merged at different stages of the process. Merging is represented by horizontal lines connecting to vertical lines. The numbers at the bottom are used as image IDs.

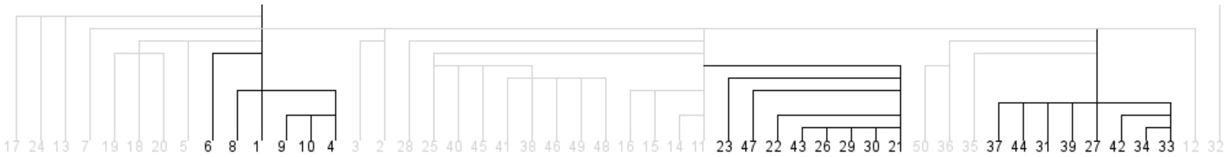


Fig. 6.3: Dendrogram based on clustering outcome produced by CURE

Fig. 6.4 presents three highlighted groups appering in the dendrogram. These groups consist to numbers 6, 8,. . . , 4; 23, 47,. . . , 21; and 37, 44, . . . , 33; respectively. If the partition were taken when picture number 6 is added to the cluster containing animal images, the cluster of city views will be negatively affected by adding images 29, 40,. . . ,11. In contrast, the cluster of landscape images will get image number 35, which is landscape as well, being affected positively. This illustrative case shows the drawback mentioned above.

CURE outperforms in certain data sets. However, it was not the case with the data sets concerning to this study. The proposed clustering approach achieves more optimal grouping in terms of semantic categorization.

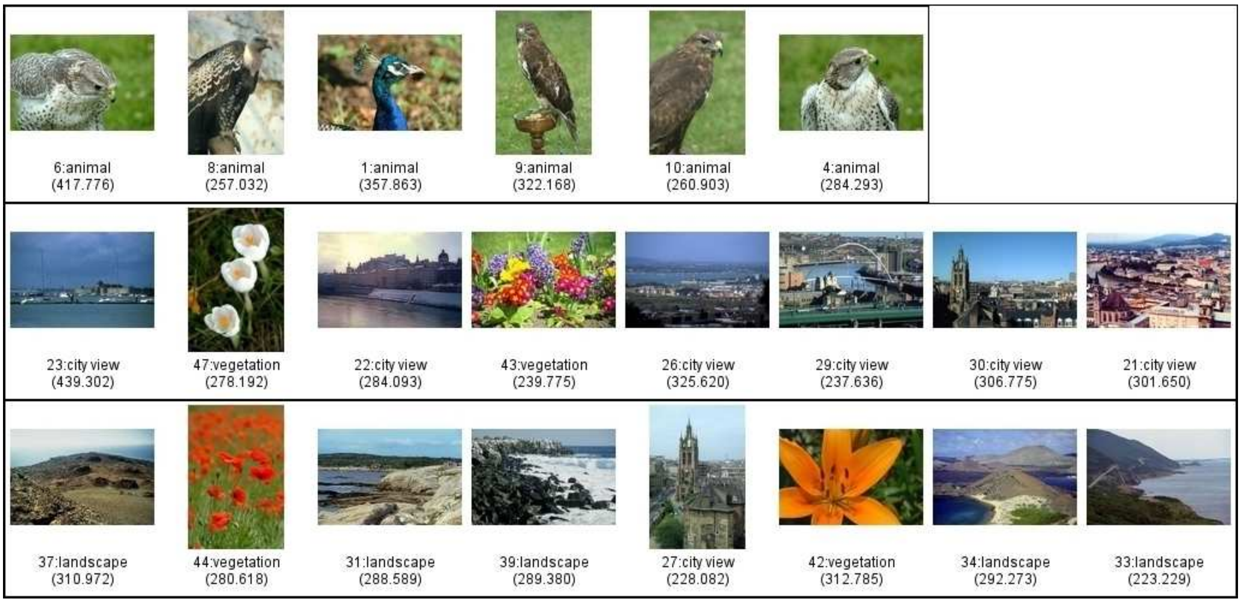


Fig. 6.4: Samples of images within selected clusters generated using CURE algorithm. Depicted clusters are labelled as *animal* (1st row), *city view* (2nd row), and *landscape* (3rd row). They are represented in the dendrogram by numbers 6, 8,. . . , 4; 23, 47,. . . , 21; and 37, 44, . . . , 33; respectively

6.3.3 Single-element instance

Tab. 6.9: Summary of test conditions used to evaluate the framework

Semantic profile	:	Single-element instance (only one index)
Lexicon	:	Animal, Building, City View, Landscape, Vegetation
Approach	:	Fuzzy clustering
Data set	:	Corel Stock and FreeFoto; 1,000 images
Feature vector	:	282-element Colour and texture descriptor

Semantic profiles obtained when partitioning the feature space into 10, 15, and 20 clusters in reported in Tab. 6.10, Tab. 6.10, and Tab. 6.10, respectively. At the bottom of tables, reported levels of matching are computed applying Eq. 6.4, Eq. 6.5, and Eq. 6.6.

Tab. 6.10: Experiment si5-c10-t0.8, cluster profiles obtained with single-element instances and 5 concepts (si5), 10 clusters (c10), and a threshold criterion δ = 0.8. Percentage of ascribed images to the cluster is indicated. The number besides the concepts corresponds to their relevance within the cluster

ID	% images	Cluster profile	ID	% images	Cluster profile
0	9.5	building (0.776)	6	10.0	vegetation (0.879)
1	10.0	animal (0.715)	7	12.2	landscape (0.848)
2	11.7	city view (0.766)	8	11.2	animal (0.544),
3	8.5	building (0.484),			vegetation (0.251139)
		animal (0.240225)	9	10.5	city view (0.363),
4	9.7	landscape (0.740)			building (0.331053)
5	6.7	vegetation (0.872)			

Low-level matching: 68.3
High-level matching: 80.0
Proposed matching: 78.8

Tab. 6.11: Experiment si5-c15-t0.8, cluster profiles obtained with single-element instances and 5 concepts (si5), 15 clusters (c15), and a threshold criterion $\delta = 0.8$. Percentage of ascribed images to the cluster is indicated. The number besides the concepts corresponds to their relevance within the cluster

ID	% images	Cluster profile	ID	% images	Cluster profile
0	7.0	city view (0.817)	8	7.8	vegetation (0.901)
1	2.8	vegetation (0.964)	9	8.4	landscape (0.694)
2	5.6	landscape (0.467),	10	5.3	building (0.876)
		city view (0.301484)	11	8.8	city view (0.349),
3	6.5	city view (0.635)			building (0.32393)
4	8.1	animal (0.686)	12	3.3	vegetation (0.669)
5	6.7	vegetation (0.666)	13	5.8	building (0.912)
6	10.3	landscape (0.871)	14	8.2	animal (0.743)
7	5.4	animal (0.432),			
		city view (0.270503)			

Low-level matching: 93.5
High-level matching: 73.7
Proposed matching: 72.5

Tab. 6.12: Experiment si5-c20-t0.8, cluster profiles obtained with single-element instances and 5 concepts (si5), 20 clusters (c20), and a threshold criterion $\delta = 0.8$. Percentage of ascribed images to the cluster is indicated. The number besides the concepts corresponds to their relevance within the cluster

ID	% images	Cluster profile	ID	% images	Cluster profile
0	5.7	city view (0.781)	9	1.8	vegetation (1.000)
1	3.9	landscape (0.433),	10	7.9	landscape (0.695)
		city view (0.334668)	11	7.8	landscape (0.890)
2	6.5	animal (0.780)	12	1.1	vegetation (1.000)
3	4.7	animal (0.735)	13	7.3	landscape (0.652)
4	4.5	building (0.466),	14	3.6	vegetation (0.734)
		city view (0.314465)	15	5.5	animal (0.690)
5	4.3	building (0.933)	16	4.4	building (0.464)
6	3.5	building (0.816)	17	6.3	vegetation (0.902)
7	3.6	building (0.896)	18	5.7	city view (0.784)
8	4.5	animal (0.407),	19	7.4	vegetation (0.814)
		city view (0.291893)			

Low-level matching: 94.0
High-level matching: 75.4
Proposed matching: 74.6

Levels of matching between expected and assigned profiles are above 70%, which represents an improvement in comparison with unsupervised results that are about 45%. It means an approximated increment of 55%. Results are alike to the ones obtained with the fully supervised RBFN approach presented in the previous chapter, though with less than half of the training images. Clustering outcomes are highly consistent with the perceptual and semantic grouping as can be observed in the following figures.

Fig. 6.5 shows samples of best-ranked images in three clusters whose profiles lead to correct classification and consequently resemble the expected semantic grouping.

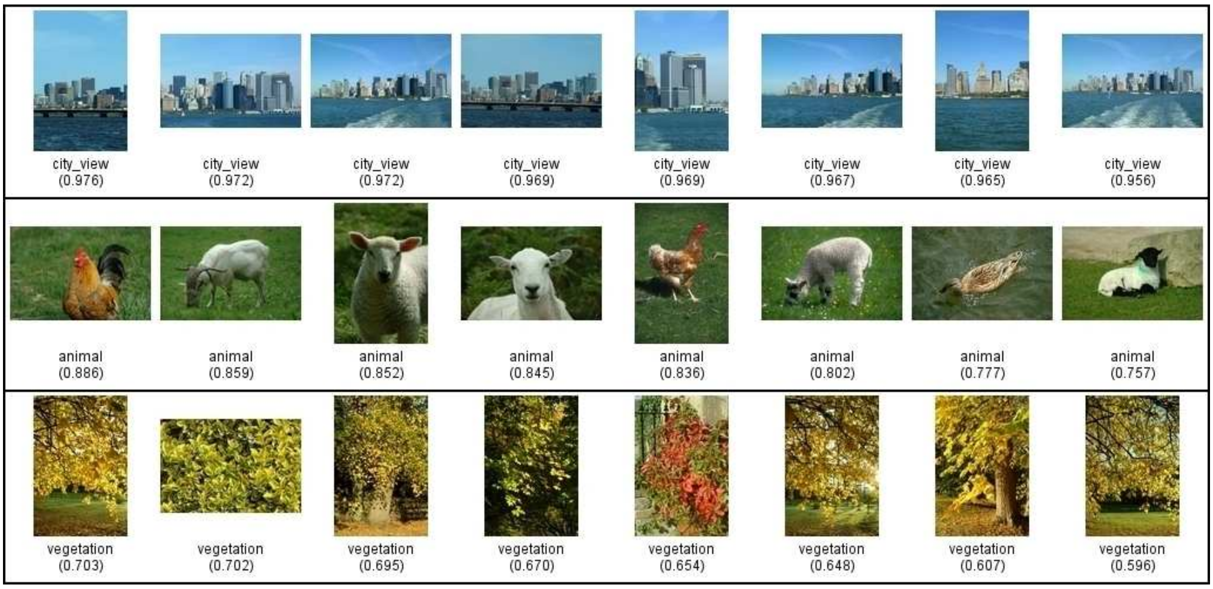


Fig. 6.5: Samples of clusters satisfying the expected profile. The cluster profiles are as follows: city view (first row), animal (second row), and vegetation (third row). Membership degree is indicated below each image

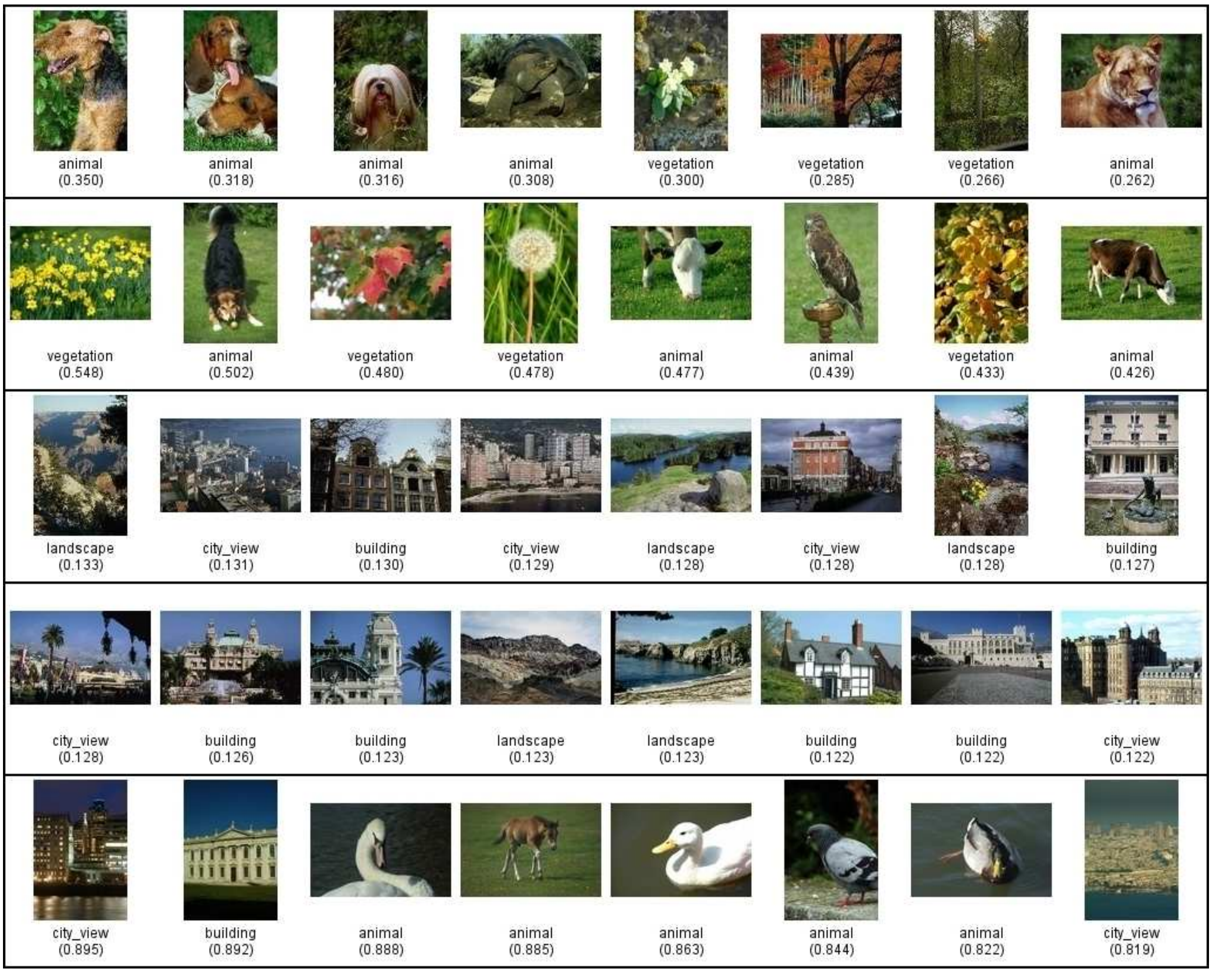


Fig. 6.6: Samples of misleading clusters. First row: animal (0.462); second row: vegetation (0.578); third row: building (0.286), city view (0.286); fourth row: building (0.341), landscape (0.340); and fifth row: animal (0.478), city view (0.320). Membership degrees appear below each image

In contrast, Fig. 6.6 depicts samples of clusters leading to misclassification. Low membership degrees reveal the high variability of feature vectors in relation to the cluster prototypes. Consequently, re-training is highly recommended.

Fig. 6.7 shows clusters containing some images that introduce "noise" into the cluster prototype, though the semantic profiles is refined using the threshold condition, in which irrelevant concepts are discarded.

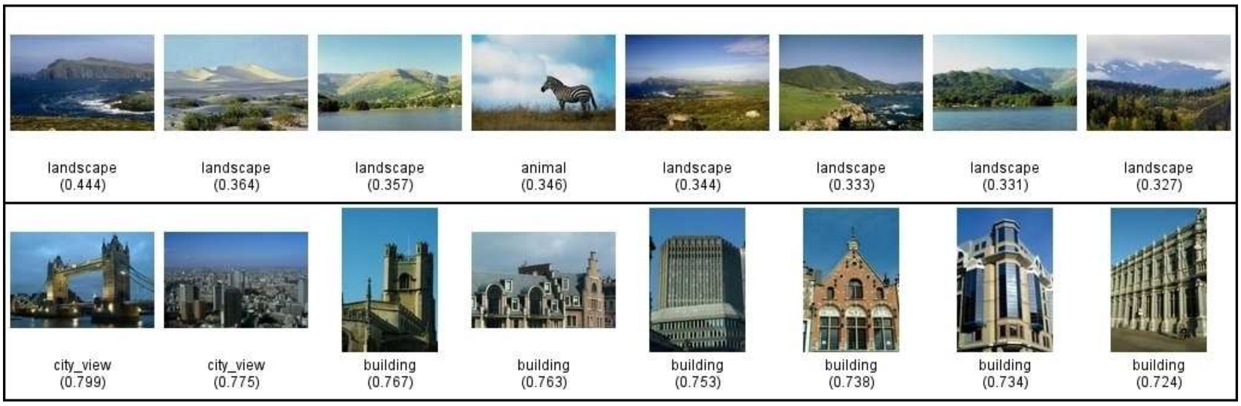


Fig. 6.7: Samples of clusters satisfying partially the expected profile. There is an animal image (first row, 4th picture) in a cluster ascribed to landscape. Two of the best-ranked images in a cluster of building images contain scenes of city views (second row, 1st and 2nd pictures). Membership is indicated below each image

Fig. 6.8 shows some outlier images whose low-level features are dissimilar to the rest of the images and consequently appear isolated in clusters with few elements (one or two vectors).

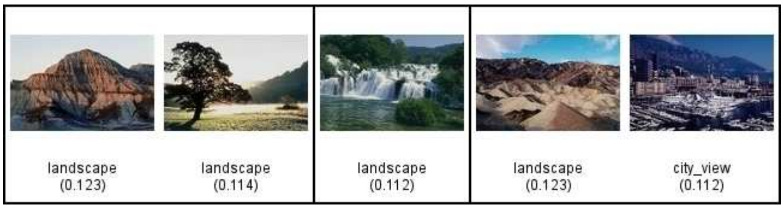


Fig. 6.8: Samples of outlier images

6.3.4 Multiple-element instance: Enlarging the Lexicon

A shortcoming found in the preceding case studies was their limited capability to deal with images containing multiple annotations. They are highly constrained to single-element instances.

In the following experimental results, the learner was presented with a list of concepts tagging each image. In average four concepts annotate an image. The total number of concepts is over 700. Influence of colour and texture component on clustering results was evaluated. The technical reference of this evaluation is summarised in Tab. 6.13.

Tab. 6.13: Summary of test conditions used to evaluate the framework

Semantic profile	:	Multiple-element instance (two or more indexes)
Lexicon	:	Over 700 concepts
Approach	:	Fuzzy clustering
Data set	:	Corel stock; 3,000 images
Feature vector	:	282-element Colour and texture descriptor

List of clusters in Tab. 6.14 group images presenting high-level of colour matching. Strong influence of colour can be observed in Fig. 6.9 showing some best-ranked images within these clusters.

Tab. 6.14: Experiment mi700-c50-t0.4, cluster profiles obtained with multiple-element instances and over 700 concepts (mi700), 50 clusters (c50), and a threshold criterion $\delta = 0.4$. Percentage of ascribed images to the cluster is indicated. The number besides the concepts corresponds to their relevance within the cluster. Images within these clusters present a high-level of matching in the colour component

ID	% Images	Cluster profile
4	0.27	building (0.156), road (0.117), path (0.117)
13	0.41	flowers (0.217), trees (0.123)
30	1.63	buildings (0.129), building (0.106), trees (0.102)
39	1.29	people (0.153), building (0.112), tree (0.079)

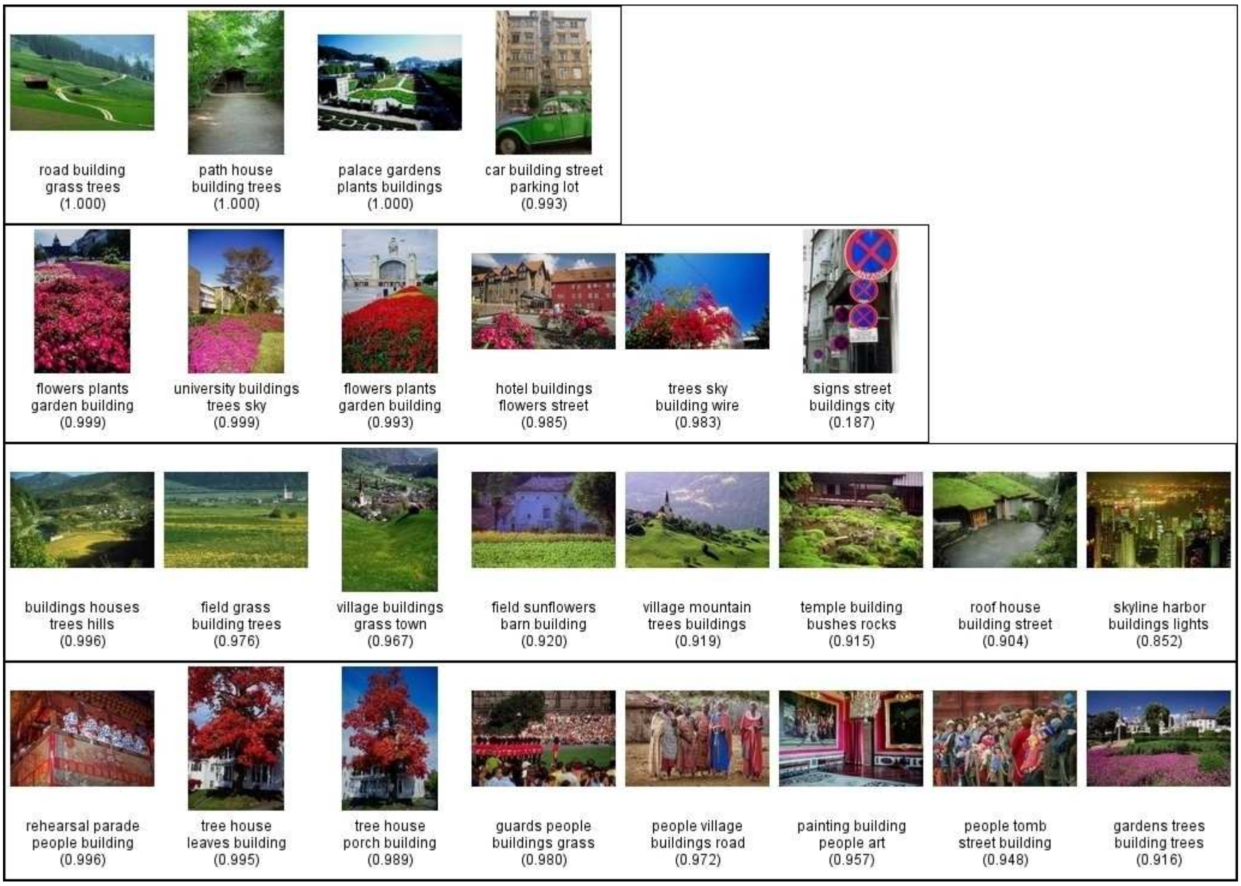

road building grass trees (1.000)
path house building trees (1.000)
palace gardens plants buildings (1.000)
car building street parking lot (0.993)
flowers plants garden building (0.999)
university buildings trees sky (0.999)
flowers plants garden building (0.993)
hotel buildings flowers street (0.985)
trees sky building wire (0.983)
signs street buildings city (0.187)
buildings houses trees hills (0.996)
field grass building trees (0.976)
village buildings grass town (0.967)
field sunflowers barn building (0.920)
village mountain trees buildings (0.919)
temple building bushes rocks (0.915)
roof house building street (0.904)
skyline harbor buildings lights (0.852)
rehearsal parade people building (0.996)
tree house leaves building (0.995)
tree house porch building (0.989)
guards people buildings grass (0.980)
people village buildings road (0.972)
painting building people art (0.957)
people tomb street building (0.948)
gardens trees building trees (0.916)

Fig. 6.9: Samples of clusters with strong influence of colour

Cluster profile shows in Tab. 6.15 correspond to images with strong texture similarity. It can be observed in the best-ranked images within this cluster depicted in Fig. 6.10.

Tab. 6.15: Experiment mi700-c50-t0.4. Percentage of ascribed images to the cluster is indicated. The number besides the concepts corresponds to their relevance within the cluster. Images within this cluster present strong texture similarity

ID	% Images	Cluster profile
25	2.38	building (0.152), buildings (0.074), people (0.054), street (0.046), water (0.045)

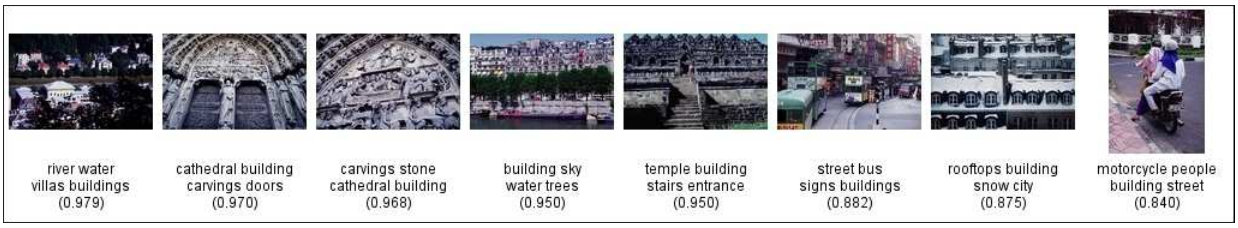


Fig. 6.10: Samples of cluster containing images with strong texture similarity

As presented in Tab. 6.16, relevance of the concepts depends not only on the number of images but also on the levels of matching with regard to the cluster prototype. The clusters in this case contain few images, but because of the low membership degrees, concepts' weights are low as well. Fig. 6.11 shows the images assigned to these clusters.

Tab. 6.16: Experiment mi700-c50-t0.4. Concept's relevance does not depend on the number of images, but in the levels of matching

ID	% Images	Cluster profile
16	0.07	mill (0.480)
18	0.20	sky (0.211), gate (0.166)
27	0.14	facade (0.266)

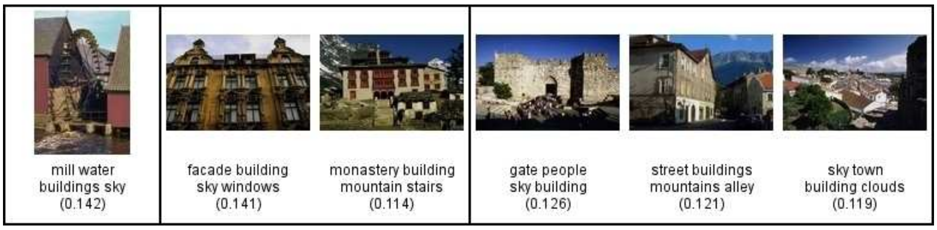


Fig. 6.11: Samples of clusters with very few images

Samples of clusters containing images satisfying conceptual and perceptual similarity are presented in Tab. 6.17 and Fig. 6.17. The order of image profiles has an effect on the high-level matching.

Tab. 6.17: Experiment mi700-c50-t0.4. Percentage of ascribed images to the cluster is indicated. The number besides the concepts corresponds to their relevance within the cluster. Concept's relevance depends on the order of annotations

ID	% Images	Cluster profile
29	2.18	buildings (0.181), building (0.112), sky (0.090)
42	3.06	building (0.155), street (0.100), buildings (0.084)

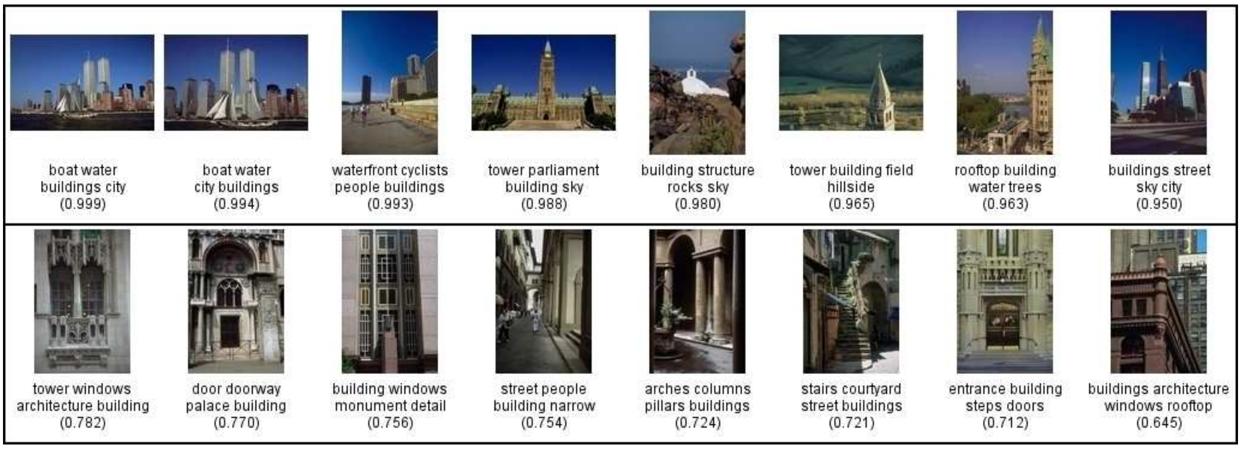

boat water buildings city (0.999)
boat water city buildings (0.994)
waterfront cyclists people buildings (0.993)
tower parliament building sky (0.988)
building structure rocks sky (0.980)
tower building field hillside (0.965)
rooftop building water trees (0.963)
buildings street sky city (0.950)
tower windows architecture building (0.782)
door doorway palace building (0.770)
building windows monument detail (0.756)
street people building narrow (0.754)
arches columns pillars buildings (0.724)
stairs courtyard street buildings (0.721)
entrance building steps doors (0.712)
buildings architecture windows rooftop (0.645)

Fig. 6.12: Samples of clusters containing images satisfying conceptual and perceptual similarity. The first two images (1st row) are similar, however the order of their profiles has a negative effect on the semantic cluster profile

Tab. 6.18: Experiment mi700-c50-t0.4. Cluster profiles containing images with high saturation and darkness

ID	% Images	Cluster profile
2	0.75	building (0.161), buildings (0.155)
21	2.52	buildings (0.147), building (0.104), sky (0.051), street (0.049), elevators (0.034)
49	1.36	buildings (0.160), building (0.122), sky (0.072)

Tab. 6.18 and Fig. 6.13 show samples of how contribution of descriptor elements facilitates separation of images in different clusters, though their commonalities in terms of saturation and brightness.

Results in Tab. 6.19 and Fig. 6.14 correspond to a variety of images of costumes, paintings, soldiers, people, etc. Beyond any low-level feature similarity, these images have a common concept "building".

Tab. 6.19: Experiment mi700-c50-t0.4. Cluster profiles containing a variety of images

ID	% Images	Cluster profile
6	1.84	building (0.124), buildings (0.068), lights (0.065), palace (0.064), people (0.061)
24	3.26	building (0.146), people (0.132), buildings (0.066), street (0.053)
28	3.67	building (0.109), trees (0.102), buildings (0.080), water (0.049), village (0.043)
33	3.54	people (0.119), building (0.117), street (0.101)

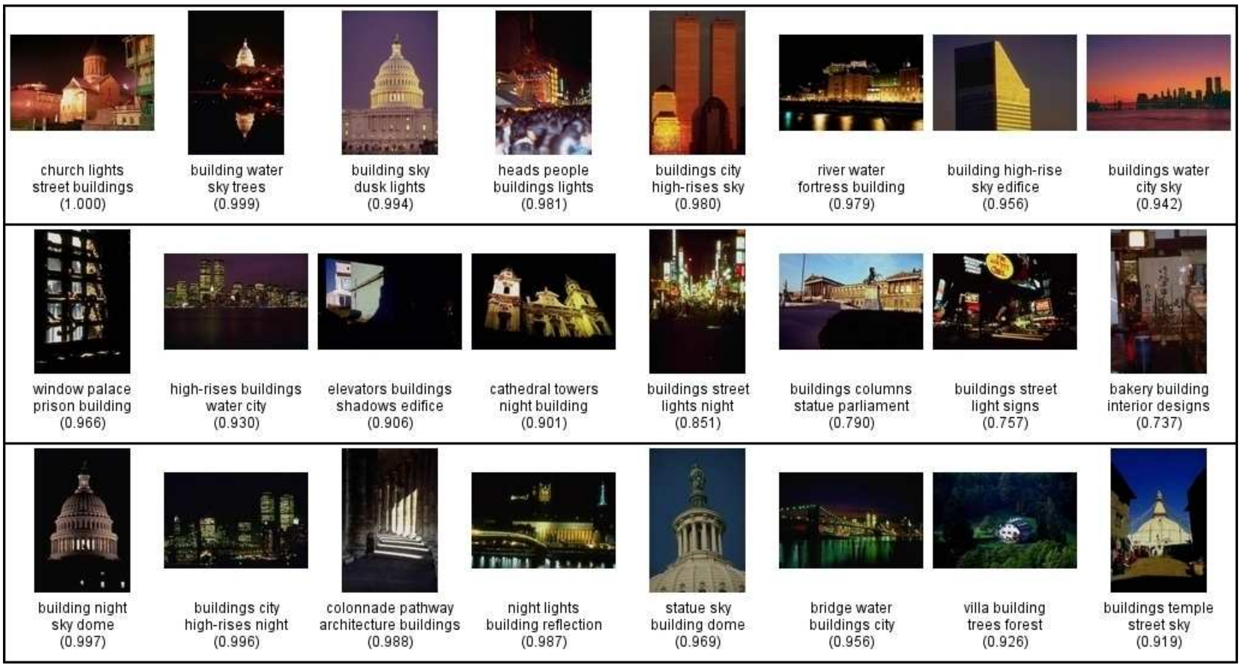

church lights street buildings (1.000)
building water sky trees (0.999)
building sky dusk lights (0.994)
heads people buildings lights (0.981)
buildings city high-rises sky (0.980)
river water fortress building (0.979)
building high-rise sky edifice (0.956)
buildings water city sky (0.942)
window palace prison building (0.966)
high-rises buildings water city (0.930)
elevators buildings shadows edifice (0.906)
cathedral towers night building (0.901)
buildings street lights night (0.851)
buildings columns statue parliament (0.790)
buildings street light signs (0.757)
bakery building interior designs (0.737)
building night sky dome (0.997)
buildings city high-rises night (0.996)
colonnade pathway architecture buildings (0.988)
night lights building reflection (0.987)
statue sky building dome (0.969)
bridge water buildings city (0.956)
villa building trees forest (0.926)
buildings temple street sky (0.919)

Fig. 6.13: Samples of clusters containing images with high saturation and darkness

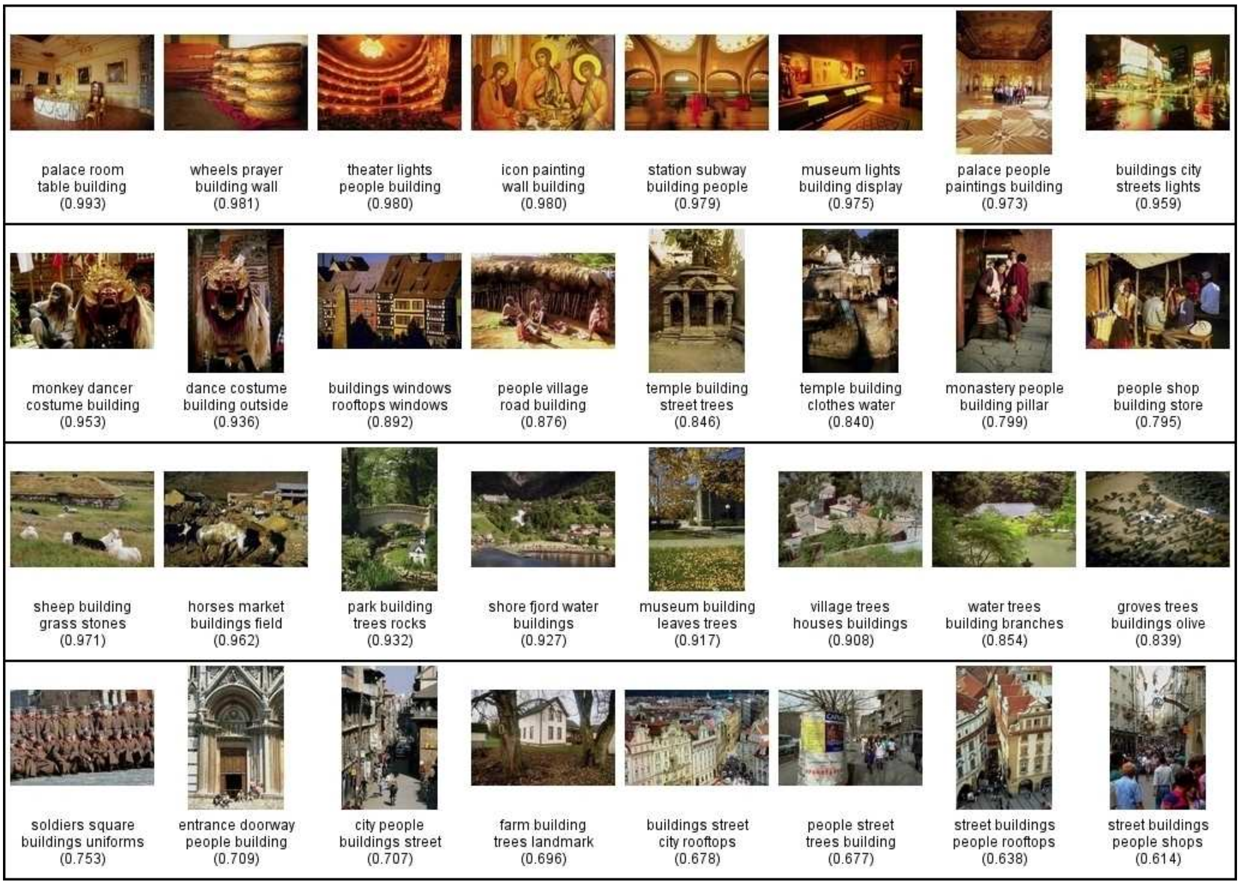

palace room table building (0.993)
wheels prayer building wall (0.981)
theater lights people building (0.980)
icon painting wall building (0.980)
station subway building people (0.979)
museum lights building display (0.975)
palace people paintings building (0.973)
buildings city streets lights (0.959)
monkey dancer costume building (0.953)
dance costume building outside (0.936)
buildings windows rooftops windows (0.892)
people village road building (0.876)
temple building street trees (0.846)
temple building clothes water (0.840)
monastery people building pillar (0.799)
people shop building store (0.795)
sheep building grass stones (0.971)
horses market buildings field (0.962)
park building trees rocks (0.932)
shore fjord water buildings (0.927)
museum building leaves trees (0.917)
village trees houses buildings (0.908)
water trees building branches (0.854)
groves trees buildings olive (0.839)
soldiers square buildings uniforms (0.753)
entrance doorway people building (0.709)
city people buildings street (0.707)
farm building trees landmark (0.696)
buildings street city rooftops (0.678)
people street trees building (0.677)
street buildings people rooftops (0.638)
street buildings people shops (0.614)

Fig. 6.14: Samples of clusters with a variety of images

6.4 Summary of Experimental Studies

Selected experimental results are listed in a tabular way in Tab. 6.20 to summarise considerations with regard of proposed framework. Resubstitution and holdout methods were applied in selecting training and testing sets.

Tab. 6.20: Run 1-11: multiple-element instance, run 12-24: single-element instance. Profile size indicates the number of concepts in each profile. Low-, high-level, and proposed matching are computed using Eq. 6.4, Eq. 6.5, and Eq. 6.6, respectively

Run	Clusters	Threshold	Average profile size	Low-level matching (%)	High-level matching (%)	Proposed matching (%)
1	5	0.5	8.60	74.67	26.18	50.77
2	10	0.5	8.10	55.99	27.26	59.65
3	15	0.5	7.80	54.56	27.66	60.72
4	20	0.5	7.00	50.51	28.92	64.44
5	25	0.5	6.80	50.50	29.05	64.41
6	30	0.5	6.53	48.94	28.97	65.47
7	50	0.2	1.04	45.31	16.70	62.62
8	50	0.3	2.02	45.31	25.06	66.43
9	50	0.4	3.46	45.31	29.74	68.63
10	50	0.5	5.46	45.31	29.47	68.61
11	50	0.4	1.28	45.31	35.85	70.94
12	5	0.8	1.20	89.80	54.38	60.64
13	10	0.8	1.30	68.26	65.10	78.75
14	15	0.8	1.20	93.52	69.23	72.46
15	20	0.1	1.00	93.95	72.20	74.65
⋮	⋮	⋮	⋮	⋮	⋮	⋮
20	20	0.6	1.00	93.95	72.20	74.65
21	20	0.7	1.05	93.95	72.18	74.62
22	20	0.8	1.15	93.95	72.11	74.57
23	20	0.9	1.50	93.95	67.35	70.03
24	20	1.0	3.25	93.95	43.35	47.03

Run 1 to 6: Number of clusters have a direct effect on the quality of the partition. Looking at the combined metric it is noticeable that it derives in better semantic grouping.

Run 7 to 10: The threshold criterion determines the number of concepts to be included in the semantic profiles. It can be noted in the variation of profile size's average, which modifies the high-level matching and consequently the combined result presented at the last column.

Run 11: Replacing some concepts by their synonyms, hypernyms, or holonyms; simplifies the lexicon and contributes to increase the high-level and combined matching.

Run 12-14: The effects of refined partitions is demonstrated with a constrained lexicon. Run 15-24: The impact of threshold criterion is shown with a constrained lexicon.

The expected number of the concepts assigned to a certain profile is observed as a strong factor that influences the quality of the annotation. It is also noted that simplifying the lexicon increases the levels of matching.

Chapter 7

Conclusions and Discussion

Accuracy, timeliness, and ease of use are important components in determining end-user computing satisfaction (Doll and Gholamreza [158]). The first two refer to the quality of the responses independently of the complexity of the content and the efficiency of the system in terms of request/response delay.

Ease of use component is moving systems towards intelligent architectures that incorporate reasoning and learning capabilities. In this regard, image understanding is a frontrunner within multimedia computing technologies in the use of images to interface people and systems.

One of the challenges in image understanding is the sensory data gap. It expresses the difference between human perception and interpretation of multimedia content and the interpretation derived from an automatic analysis of the machine.

Following conclusions and discussion stem from the research work leading to the proposed framework for concept-related indexing of image content. Automated linguistic indexing of images is becoming a critically important area of research because of its demonstrated potential to narrow the semantic gap.

7.1 Conclusions

Visual interpretations of the image content provide lexical information that attached to image abstractions expand classic conceptual image analysis. Lexical information is useful to introduce problem domain knowledge. This sort of information can be incorporated through exemplars or inference rules.

Inference rules are suitable to compensate the absence of available exemplars or guide the learner to the correct prototype of a learnable concept. On the other hand, learning from examples is convenient since it can combine labelled and unlabelled data under partially supervised learning modes.

The nature of the relationships that can be established between image abstractions and concepts involves uncertainty, roughly connections, and subjectivity. This characterization moves the problem into the ground of fuzzy sets theory and fuzzy systems.

An important component within the proposed framework is the mechanism when presenting the exemplars to the learner. After evaluating statistical and soft computing techniques, it was found that partially supervised clustering is an appropriate method to reveal underlying structures in low-level representations, to create prototypes representing learnable concepts, and to partition the description space into groups that facilitates incremental learning.

This learning mode reduce the burden of collecting samples randomly as well as improves the quality of the chosen ones taking into account low- and high-level similarity (or dissimilarity by learning from negative exemplars or misclassified patterns). The learning strategy is also a practical way to introduce system's adaptation and can be extended onto relevance feedback.

Partitioning and hierarchical clustering algorithms were evaluated. Partially supervised fuzzy clustering equipped with an objective function establishes a more solid base to build a more accurate semantic categorization than its counterparts. Prior domain knowledge, structural data analysis, and users' feedback serve to overcome identified shortcomings in partitioning clustering.

These shortcomings refer to determining the optimal number of clusters to partition the data space, mismatching between clusters and semantic groups, and equalizing cluster populations.

A main drawback found in resembling semantic groups with clustering is the tendency presented in fuzzy c-means algorithms when grouping data within (hyper-) spherical or ellipsoid spaces based on the similarity to the cluster prototypes. An option is in adapting the mean-based decision to proximity-based approaches.

Cluster compactness can be achieved by shrinking data towards the mean or medoid. It improves the cohesiveness intra-cluster. Furthermore, cluster representatives can be used to determine boundaries inter-clusters and define an optimal decision hyperplane to differentiate them. These representatives are a kind of support vectors.

Information obtained through clustering algorithms (prototypes, space partition) and analysis (cluster-class dependencies, ranking of overlapped categories) provides a better insight of the image abstractions, refine the classifier design, and improve classification performance.

A shortcoming found in semantic classification approaches is their limited capability to deal with images containing multiple annotations. They are highly constrained to single-element instances. It is to say, only one concept by image. When dealing with multiple-element instances (two or more concepts by image) it is suggested to simplify the lexical information by substituting some concepts by their holonyms. It becomes an important design step in the construction of a robust semantic indexer. Compact lexical information reduces ambiguity in mapping images into many possible interpretations.

7.2 Discussion and Further Work

The power of new computational environments found in computer architectures and networks provides a convenient platform to build more robust solutions for the current challenges in multimedia computing.

Besides, computer vision, machine learning and pattern recognition methods contribute in a significant way to the attempt of conciliating human and machine interpretations.

However, some questions arise regarding to the computational efficiency and generalization capabilities of proposed recognisers. It attaches directly some of the components of end-user computing satisfaction and turns out the attention towards learnable and non-learnable visual concepts.

It has been pointed out that a learnable concept should be learned from a polynomial number of examples and using polynomial bounded computational resources.

Examples can be chosen through random sampling of large-scale image databases, which does not guarantee quality or good representation of the concept. For instance, there could be substantial variations in the exemplars in the random sample. Another drawback is on the definition of “good example” itself, which involves subjectivity.

This sort of search could imply the need to traverse the entire database deriving in expensive computational procedures. On top of that, availability of exemplars is a concern. Consequently, selection of suitable examples becomes a critical design step of a pattern recogniser.

In this way, active learning and relevance feedback have been incorporated successfully to strategies addressing the problem of learning concepts. Furthermore, partially supervised learning approaches are taking advantage of unlabelled data.

Accepting these approximations to provide multimedia systems with learnable concepts, it is required a procedure to determine how much knowledge a machine has accumulated with regard to a concept (reckoning partial learning) and mechanisms to identify the exemplars that contributed most to the learning achievements.

Another issue in treating a multimedia system, as an intelligent entity is the representation of acquired knowledge used by the learning algorithm. Adhering to the MPEG-7 standard, description schemes should be used to describe meta-information of learnable concepts.

It could contribute in the definition of a common “language” (lexical information) and uniform models for representing image content. It would provide interoperability of image descriptions as well as durable and stable lexicons that are a requirement to support sharing and reusing metadata among users.

The recognition process has been oriented to broad classes such as city views, nature, etc. Each class is linked to an image concept. More specific classes may not be easily distinguishable. Even though, progress in computational ontology opens the doors to more elaborated representations of image content. Aforementioned description schemes can be equipped with specific metadata structures linking linguistic indexes to ontology maps.

In summary, four issues presented in this section are: incremental concept learning, collection and organization of exemplars, the usage of existing standard metadata structures, i.e. MPEG-7 description schemes, to facilitate interoperability of knowledge representations, and the growing need of research on computational ontologies to exploit capabilities of lexical information. These issues could be considered in future studies such as:

- What can be learned from situations where different interpretations of data are possible, i.e. the ambiguity of perception (Whaite and Ferrie [159]).
- Non-learnable visual concepts may become learnable if something is known (or assumed) about the probability of the examples (Shvaytser [59]). It is required models to deal with non-learnable concepts in the sense of tracking them to find out the way these concepts may become learnable.
- Choosing good representatives of a class that accurately characterize it, is an intuitive strategy when using exemplar-based methods. Studies as the one reported by Jacobs et al. [160] turn out the attention towards selection of atypical points that can be used to describe a class. For instance, Guha et al. [157] and Lam et al. [161] proposed strategies that attempts to select good prototypes and maintains a balance of different kinds of prototypes.
- It is required a common and mathematically sound framework for the problem of image parsing, in which conceptualisation of visual patterns and their components (vocabularies) can be expressed in a standard-like way. As reported by Zhu [162], it could lead to richer and more advanced classes of vision models.

These considerations seem to be appealing when an emerging standard is taking place. The Still Image Search –JPSearch standard, destined to become ISO/IEC NP 24800, “aims to specify additional metadata and related functionalities to support the implementation of a flexible and efficient still image search.”

Bibliography

[1] A. Dasu and S. Panchanathan. A survey of media processing approaches. *IEEE Trans. Circuits and Systems for Video Technology*, 12(8):633–645, Aug 2002.

[2] R. Mohan, J.R. Smith, and C.-S. Li. Adapting multimedia internet content for universal access. *IEEE Trans. Multimedia*, 1(1):104–114, Mar 1999.

[3] J. Hunter. Enhancing the semantic interoperability of multimedia through a core ontology. *IEEE Trans. Circuits and Systems for Video Technology*, 13(1):49–58, Jan 2003.

[4] F. Nack. Aesthetics of contradiction. *IEEE Multimedia*, 10(1):11–13, Jan-Mar 2003.

[5] T. Kanade. Immersion into visual media: new applications of image understanding. *IEEE Expert*, 11 (1):73–80, Feb 1996.

[6] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, and G. Taubin. The QBIC project: Querying images by content, using color, texture, and shape. In *Proc. Storage and Retrieval for Image and Video Databases*, volume 1908 of *SPIE Proceedings*, pages 173–187, San Jose, CA, USA, 1993. ISBN 0819411418.

[7] J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. *ACM Multimedia*, pages 87–98, 1996.

[8] R. Lienhart, W. Effelsberg, and R. Jain. VisualGREP: A systematic method to compare and retrieve video sequences. In *Proc. Storage and Retrieval for Image and Video Databases*, pages 271–283, 1998.

[9] D. Ponceleon, S. Srinivasan, A. Amir, D. Petkovic, D. Zivkovic, and D. Diklic. Key to effective video retrieval: Effective cataloging and browsing. In *Proc. of the 6th ACM Int'l Conf. on Multimedia*, pages 99–108, N.Y., Sep 1998. ACM Press.

[10] D. Zhong and S.-F. Chang. An integrated approach for content-based video object segmentation and retrieval. *IEEE Trans. Circuits and Systems for Video Technology*, 9(8):1259–1268, 1999.

[11] J.-R. Ohm, F. Bunjamin, W. Liebsch, B. Makai, K. Muller, A. Smolic, and D. Zier. A multi-feature description scheme for image and video database retrieval. In *Proc. IEEE Multimedia Signal Processing Workshop*, pages 123–128, Copenhagen, 1999.

[12] 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 Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[13] J.C. Simon. Recent progress to formal approach of pattern recognition and scene analysis. *Pattern Recognition*, 7:117–124, 1975.

[14] C. Dorai and S. Venkatesh. Bridging the semantic gap with computational media aesthetics. *IEEE Multimedia*, 10(2):15–17, Apr–Jun 2003.

[15] R.E. Bellman. *Adaptive Control Processes*. Princeton University Press, 1961.

[16] L. Huber. *Visual Categorization in Pigeons*. Department of Psychology, Tufts University and Compar- ative Cognition Press, http://www.pigeon.psy.tufts.edu/avc/toc.htm, Sep 2001.

[17] P. Lipson, E. Grimson, and P. Sinha. Configuration based scene classification and image indexing. In *Proc. IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 1007–1013, June 1997.

[18] S. Santini, A. Gupta, and R. Jain. Emergent semantics through interaction in image databases. *IEEE Trans. Knowledge and Data Engineering*, 13(3):337–351, May/June 2001.

[19] Q. Iqbal and J.K. Aggarwal. Combining structure, color and texture for image retrieval: A performance evaluation. In *Proc. 16th Int'l Conf. on Pattern Recognition*, volume 2, pages 438–443, Aug 2002.

[20] J.Z. Wang, J.L., and S.C. Lin. Evaluation strategies for automatic linguistic indexing of pictures. In *Proc. IEEE Int'l Conf. on Image Processing*, volume 3, pages 617–620, Sep 2003.

[21] E. Izquierdo and A. Dorado. Climbing the semantic ladder: Towards semantic semi-automatic image annotation using MPEG-7 descriptor schemas. In *Proc. IEEE Int'l Workshop on Computer Architecture for Machine Perception*, Italy, Jul 2005. (to appear).

[22] R.O. Duda, P.E. Hart, , and D.G. Stork. *Pattern Classification*. J. Wiley, New York, NY, USA, 2nd edition, 2001.

[23] E. Mrowka, A. Dorado, W. Pedrycz, and E. Izquierdo. Dimensionality reduction for content-based image classification. In *Proc. 8th IEEE Int'l Conf. on Information Visualisation*, pages 435–438, London, UK, Jul 2004.

[24] A. Dorado and E. Izquierdo. An approach for supervised semantic annotation. In E. Izquierdo, editor, *Proc. Digital Media Processing for Interactive Services, Proc. 4th European Workshop on Image Analysis for Multimedia Interactive Services*, pages 117–121, London, UK, Apr 2003. World Scientific. ISBN 981- 238-355-7.

[25] A. Dorado and E. Izquierdo. Semi-automatic image annotation using frequent keyword mining. In *Proc. 7th Int'l Conf. on Information Visualization*, volume 1, pages 532–535, London, UK, Jul 2003. IEEE Computer Society.

[26] A. Dorado and E. Izquierdo. Semantic labeling of images combining color, texture and keywords. In *Proc. IEEE Int'l Conf. on Image Processing*, volume 2, pages 433–436, Barcelona, Spain, Sep 2003. [27] A. Dorado and E. Izquierdo. Knowledge representation of low level features for semantic video analysis.
In *Proc. of the ECML/PKDD 2003 Workshop on Multimedia Discovery and Mining*, pages 72–83, Cavtat-Dubrovnik, Croatia, Sep 2003. ISBN 953-6690-32-2.

[28] V. Zeljkovic, A. Dorado, and E. Izquierdo. A modified shading model method for building detection. In *Proc. of 5th European workshop on Image Analysis for Multimedia Interactive Services*, page 95, Portugal, Apr 2004.

[29] A. Dorado and E. Izquierdo. *Image and Video Retrieval: Third Int'l Conf., CIVR 2004 Proceedings*, volume 3115 of *Lecture Notes in Computer Science*, chapter Exploiting Problem Domain Knowledge for Accurate Building Image Classification, pages 199–206. Springer-Verlag GmbH, Dublin, Ireland, Jul 2004.

[30] A. Dorado, D. Djordjevic, W. Pedrycz, and E. Izquierdo. Supervised semantic scene classification based on low-level clustering and relevance feedback. In *Proc. European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology*, London, UK, Nov 2004.

[31] V. Zeljkovic, A. Dorado, Z. Trpovski, and E. Izquierdo. Classification of building images in video sequences. *Electronics Letters*, 40(3):169–170, Feb 2004.

[32] A. Dorado, J. Calic, and E. Izquierdo. A rule-based video annotation system. *IEEE Trans. Circuits and Systems for Video Technology*, 14(5):622–633, May 2004. Special Issue on Audio and Video Analysis for Multimedia Interactive Services.

[33] V. Zeljkovic, A. Dorado, and E. Izquierdo. Combining a fuzzy rule-based classifier and illumination invariance for improved building detection. *IEEE Trans. Circuits and Systems for Video Technology*, 14(11):1277–1280, Nov 2004.

[34] D. Djordjevic, A. Dorado, W. Pedrycz, and E. Izquierdo. Concept-oriented sample images selection. In *Proc. 6th European workshop on Image Analysis for Multimedia Interactive Services*, Montreux, Switzerland, Apr 2005.

[35] A. Dorado, D. Djordjevic, W. Pedrycz, and E. Izquierdo. Efficient image selection for concept learning. *IEE Proceedings Visual, Image & Signal Processing*, 2005. (to appear).

[36] A. Dorado, W. Pedrycz, and E. Izquierdo. User-driven fuzzy clustering: On the road to semantic classification. In *Proc. The Tenth Int'l Conf. on Rough Sets, Fuzzy Sets, Data Mining, and Granular Computing*. Springer-Verlag, Aug/Sep 2005. (to appear).

[37] D. Crevier and R. Lepage. Knowledge-based image understanding systems - a survey. *Computer Vision and Image Understanding*, 67(2):161–185, 1997.

[38] A. Rares, M.J.T. Reinders, and E.A. Hendriks. Image interpretation systems: State-of-the-art in image interpretation. Technical Report MCCWS 2.1.1.3.C, Information and Communication Theory Group, Telematica Instituut, Aug 1999.

[39] B. Heisele, A. Verri, and T. Poggio. Learning and vision machines. *Proceedings of the IEEE*, 90(7): 1164–1177, July 2002.

[40] K. Grill-Spector and N. Kanwisher. Visual recognition: as soon as you see it, you know what it is. *Psychological Science*, 16(2):152–160, Feb 2005.

[41] Fundamentals of remote sensing: CCRS remote sensing tutorial. www.ccrs.nrcan.gc.ca/ccrs/.

[42] V. Mezaris, H. Doulaverakis, R. Medina Beltran de Otalora, S. Herrmann, I. Kompatsiaris, and M. G. Strintzis. *Image and Video Retrieval: Third International Conference, CIVR 2004 Proceedings*, volume 3115 of *Lecture Notes in Computer Science*, chapter A Test-Bed for Region-Based Image Retrieval Using

Multiple Segmentation Algorithms and the MPEG-7 eXperimentation Model: The Schema Reference System, pages 592–600. Springer-Verlag GmbH, Dublin, Ireland, Jul 2004.

[43] M. Stefik. *Introduction to Knowledge Systems*. Morgan Kaufmann Publishers Inc., 1995.

[44] S.S. Intille and A.F. Bobick. Closed-world tracking. In *Proc. IEEE Fifth Int'l Conf. on Computer Vision*, pages 672–678, 1995.

[45] M. Mrak, C. K. Abhayaratne, E. Izquierdo, "On the influence of motion vector precision limiting in scalable video coding", *7th International Conference on Signal Processing* (ICSP 2004). Beijing, China, 31 August - 4 September 2004, Volume 2, Pages 1143-1146

[46] E. Tsomko, H-J. Kim, E. Izquierdo, "Linear Gaussian blur evolution for detection of blurry images", *IET Image Processing*, Volume 4, Issue 4, Pages 302-312

[47] M. Mrak, N. Sprljan, E. Izquierdo, "Motion estimation in temporal subbands for quality scalable motion coding", *IET Electronics Letters,* Volume 41, Issue 19, Pages 1050-1051

[48] A. Pinheiro, E. Izquierdo, M. Ghanhari, "Shape Matching using a Curvature Based Polygonal Approximation in Scale-Space", *IEEE Proceedings International Conference on Image Processing (ICIP 2000).* Vancouver, BC, 10-13 September 2000, Volume 2, Pages 538-541

[49] Y. Wang, E. Izquierdo, "High-Capacity Data Hiding in MPEG-2 Compressed Video", *9th International Workshop on Systems, Signals and Image Processing (IWSSIP 2002),* World Scientific, Manchester, England, 7-8 November 2002, Pages 212-218

[50] P. Borges, J. Mayer, E. Izquierdo, "Robust and Transparent Color Modulation for Text Data Hiding", *IEEE Transactions on Multimedia,* Volume 10, Issue 8, Pages 1479-1489

[51] E. Izquierdo, et al., "Advanced Content-Based Semantic Scene Analysis and Information Retrieval: The Schema Project", *4th European Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2003*), World Scientific Publishing, London, England, 9-11 April 2003, Pages 519-528

[52] T. Zgaljic, N. Sprljan, E. Izquierdo, "Bitstream Syntax Description based Adaptation of Scalable Video", *2nd European Workshop on the Integration of Knowledge, Semantics and Digital Media Technology (EWIMT 2005).* London, England, 30 November - 1 December 2005

[53] A. Rosenfeld. Image analysis: Problems, progress and prospects. *Pattern Recognition*, 17(1):3–12, 1984.

[54] K. Barnard, P. Duygulu, and D. Forsyth. *Exploiting Text and Image Feature Co-occurrence Statistics in Large Datasets*. Lecture Notes in Computer Science. Springer-Verlag, 2005.

[55] P. Duygulu, O.C. Ozcanh, and N. Papernick. Comparison of feature sets using multimedia translation. In *Proc. 18th Int'l Symposium on Computer and Information Sciences*, Nov 2003.

[56] M.C.S. Paterno, F.S. Lim, and W.K. Leow. Fuzzy semantic labeling for image retrieval. In *Proc. IEEE Int'l Conf. on Multimedia and Expo*, volume 2, pages 767–770, June 2004.

[57] A. Gu´erin-Dugu´e and A. Oliva. Classification of scene photographs from local orientations features. *Pattern Recognition Letters*, 21:1135–1140, 2000.

[58] L.G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[59] H. Shvaytser. Learnable and nonlearnable visual concepts. *IEEE Trans. Pattern Analysis Machine Intelligence*, 12(5):459–466, May 1990.

[60] L. Saitta and F. Bergadano. Pattern recognition and Valiant's learning framework. *IEEE Trans. Pattern Analysis Machine Intelligence*, 15(2):145–155, 1993.

[61] G. Qiu, X. Feng, and J. Fang. Compressing histogram representations for automatic colour photo categorization. *Pattern Recognition*, 37:2177–2193, 2004.

[62] M. Nakazato and T.S. Huang. Extending image retrieval with group-oriented interface. In *Proc. IEEE Int'l Conf. on Multimedia and Expo*, volume 1, pages 201–204, Aug 2002.

[63] B. Bhanu and A. Dong. Concepts learning with fuzzy clustering and relevance feedback. *Engineering Applications of Artificial Intelligence*, 15:123–138, 2002.

[64] T. Yu, T. Jan, J. Debenham, and S. Simoff. Incorporating prior domain knowledge in machine learn- ing: A review. In *Proc. Int'l Conf. on Advances in Intelligent Systems - Theory and Applications*, Luxembourg, Nov 2004.

[65] T. Yoshizawa and H. Schweitzer. Long-term learning of semantic grouping from relevance-feedback. In *Proc. 6th ACM SIGMM Int'l Workshop Multimedia Information*, pages 165–172, 2004.

[66] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans. Circuits and Systems for Video Technology*, 8(5):644–655, Sep 1998.

[67] M. M. Gorkani and R. W. Picard. Texture orientation for sorting photos at a glance. In *Proc. IEEE 12th Int'l Conf. on Pattern Recognition*, volume 1, pages 459–464, Oct 1994.

[68] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *Proc. IEEE Int'l Workshop on Content-Based Access of Image and Video Database*, pages 42–51, Jan 1998.

[69] A. Vailaya, A. Jain, and H. J. Zhang. On image classification: City vs. landscape. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries*, pages 3–8, Jun 1998.

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

[71] A. Loui and A. Savakis. Automated event clustering and quality screening of consumer pictures for digital albuming. *IEEE Trans. Multimedia*, 5:390–402, Sep 2003.

[72] A. Loui and A. Savakis. Automatic image event segmentation and quality screening for albuming applications. In *Proc. IEEE Int'l Conf. on Multimedia and Expo*, volume 2, pages 1125–1128, Aug 2000.

[73] J. Luo, A. Savakis, S.P. Etz, and A. Singhal. On the application of Bayes networks to semantic understanding of consumer photographs. In *Proc. IEEE Int'l Conf. on Image Processing*, volume 3, pages 512–515, Sep 2000.

[74] J. Luo and A. Savakis. Indoor vs outdoor classification of consumer photographs using low-level and semantic features. In *Proc. IEEE Int'l Conf. on Image Processing*, volume 2, pages 745–748, Oct 2001.

[75] Y-N. Wang, L-B. Chen, and B-G. Hu. Semantic extraction of the building images using support vector machines. In *Proc. IEEE Int'l Conf. on Machine Learning and Cybernetics*, volume 3, pages 1608–1613, Beijing, Nov 2002.

[76] M. Boutell, J. Luo, and R.T. Gray. Sunset scene classification using simulated image recomposition. In *Proc. IEEE Int'l Conf. on Multimedia and Expo*, volume 1, pages 37–40, Jul 2003.

[77] T. Zgaljic, N. Sprljan, E. Izquierdo, "Bit-stream allocation methods for scalable video coding supporting wireless communications", *Signal Processing: Image Communication (Special edition on Mobile Video)*, Volume 22, Issue 3, Pages 298-316

[78] E. Izquierdo, M. Ernst, "Motion/Disparity analysis for 3D-Video-Conference Applications", 1995 *International Workshop on Stereoscopy and 3-Dimensional Imaging (IWS3DI 1995).* Santorini, Greece, September 1995

[79] E. Izquierdo, M. Ghanbari, "Key Components for an Advanced Segmentation System", *IEEE Transactions on Multimedia,* Volume 4, Issue 1, Pages 97-113

[80] J. Calic, E. Izquierdo, "Temporal Segmentation of MPEG Video Streams", EURASIP Journal on Applied Signal Processing, Issue 6, Pages 561-565, 2002/6/26

[81] S. Kay, E. Izquierdo, "Robust Content Based Image Watermarking", *3rd Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2001)*, Tampere, Finland, 16-17 May 2001, Pages 53-56

[82] J. Calic, S. Sav, E. Izquierdo, S. Marlow, N. Murphy, N. O'Connor, "Temporal Video Segmentation For Real-Time Key Frame Extraction", *27th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2002*). Orlando, FL, 13-17 May 2002, Volume 4, Pages 3632-3635

[83] J. Calic, E. Izquierdo, "A Multiresolution Technique for Video Indexing and Retrieval", *2002 International Conference on Image Processing (ICIP 2002*), Rochester, NY, September 22-25, 2002, Volume 1, Pages 952—955

[84] J. Ramos, N. Guil, J. González, E. Zapata, E. Izquierdo, "Logotype detection to support semantic-based video annotation", *Journal Signal Processing: Image Communication,* Volume 22, Issue 7-8, Pages 669-679

[85] J. Calic, E. Izquierdo, "Towards Real-Time Shot Detection in the MPEG Compressed Domain", *3rd Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS* 2001), Tampere, Finland, 16-17 May 2001, Pages 1-5

[86] M. Hanke, E. Izquierdo, R. März, "On Asymptotics in Case of Linear Index-2 Differential-Algebraic Equations", *SIAM Journal on Numerical Analysis* 1998, Volume 35, Issue 4, Pages 1326-1346.

[87] A. Mojsilovic and B. Rogowitz. Capturing image semantics with low-level descriptors. In *Proc. IEEE Int'l Conf. on Image Processing*, volume 1, pages 18–21, 2001.

[88] A. Barla, F. Odone, and A. Verri. Old fashioned state-of-the-art image classification. In *Proc. 12th IEEE Int'l Conf. on Image Analysis and Processing*, pages 566–571, Sep 2003.

[89] A.K. Jain, P.W. Duin, and J. Mao. Statistical pattern recognition: a review. *IEEE Trans. Pattern Analysis Machine Intelligence*, 22(1):4–37, 2000.

[90] H.-P. Huang and Y.-H. Liu. Fuzzy support vector machines for pattern recognition and data

mining.*Int'l Journal of Fuzzy Systems*, 4(3):826–835, Sep 2002.

[91] O. Chapelle, P. Haffner, and V. N. Vapnik. Support vector machines for histogram -based image classification. *IEEE Trans. Neural Networks*, 10(5):1055–1064, 1999.

[92] R. Brunelli and O. Mich. On the use of histograms for image retrieval. In *Proc. IEEE Int'l Conf. on Multimedia Computing and Systems*, volume 2, pages 143–147, June 1999.

[93] N. Serrano, A. Savakis, and A. Luo. A computationally efficient approach to indoor/outdoor scene classification. In *Proc. IEEE 16th Int'l Conf. on Pattern Recognition*, volume 4, pages 146–149, 2002.

[94] R. Yan, Y. Liu, R. Jin, and A. Hauptmann. On predicting rare classes with SVM ensembles in scene classification. In *Proc. IEEE Int'l Conf. on Acoustics, Speech, and Signal Processing*, volume 3, pages 21–24, Apr 2003.

[85] C-F. Tsai, K. McGarry, and J. Tait. Image classification using hybrid neural networks. In *Proc. 26th annual int'l ACM SIGIR Conf. on Research and development in information retrieval*, pages 431–432, Jul 2003.

[96] S. Prabhakar, Hui Cheng, J.C. Handley, Zhigang Fan, and Ying wei Lin. Picture-graphics color image classification. In *IEEE Int'l Conf. on Image Processing*, volume 2, pages 785–788, Sep 2002.

[97] S.-B. Dong and Y.-M. Yang. Hierarchical web image classification by multi-level features. In *Proc. IEEE Int'l Conf. on Machine Learning and Cybernetics*, volume 2, pages 663–668, Nov 2002.

[98] J. R. Smith B. Tseng M. R. Naphade, C. Y. Lin and S. Basu. Learning to annotate video databases. In *Proc. SPIE Conf. on Storage and Retrieval on Media databases*, 2002.

[99] S. Wan, E. Izquierdo, "Rate-distortion optimized motion-compensated prediction for packet loss resilient video coding", *Image Processing, IEEE Transactions on,* Volume 16, Issue 5, Pages 1327-1338

[100] E. Izquierdo, V. Guerra, "An Ill-Posed Operator for Secure Image Authentication", *IEEE Transactions on Circuits and Systems for Video Technology*, Volume 13, Issue 8, Pages 842-852

[101] N. O'Connor, E. Izquierdo et al., "Region and Object Segmentation Algorithms in the Qimera Segmentation Platform", *3rd International Workshop on Content-Based Multimedia Indexing (CBMI 2003)*, Rennes, France, 22-24 September 2003, Pages 1-8

[102] S. Sprljan, M. Mrak, C. Abhayaratne, E. Izquierdo, "A Scalable Coding Framework for Efficient Video Adaptation", *6th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS 2005).* Switzerland, 2005, 13-15 April 2005, Pages 1-4

[103] E. Izquierdo, S. Kruse, "Image Analysis for 3D Modeling, Rendering, and Virtual View Generation", *Elsevier Journal Computer Vision and Image Understanding,* 1998, Volume 71, Issue 2, Pages 231-253

[104] E. Izquierdo, J-R. Ohm, "Image-based rendering and 3D modeling: a complete framework", *Signal Processing: Image Communication*, Volume 15, Issue 10, 2000, Pages 817-858

[105] N. Ramzan, H. Park, E. Izquierdo, "Video streaming over P2P networks: Challenges and opportunities", *Elsevier Journal Signal Processing: Image Communication*, Volume 27, Issue 5, Pages 401-411

[106] N. Ramzan, S. Wan, E. Izquierdo, "Joint Source-Channel Coding for Wavelet-Based Scalable Video Transmission Using an Adaptive Turbo Code", *EURASIP Journal on Image and Video Processing*, Volume 2007, Pages 1-12

[107] C. Zhang and T. Chen. An active learning framework for content based information retrieval. *IEEE Trans. Multimedia, Special Issue on Multimedia Database*, 4(2):260–268, June 2002.

[108] S. Tong and E. Chang. Support vector machine active learning for image retrieval. *ACM Multimedia*, 2001.

[109] J.Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Analysis Machine Intelligence*, 23(9):947–963, Sep 2001.

[110] C. Zhang and T. Chen. *The Handbook of Video Database Design and Applications*, chapter From Low Level Features to High Level Semantics. CRC Press, 2003.

[111] C. Zhang and T. Chen. Annotating retrieval database with active learning. In *Proc. IEEE Int'l Conf. on Image Processing*, volume 3, pages 595–598, Sep 2003.

[112] J. Laaksonen, M. Koskela, and E. Oja. PicSOM -self-organizing image retrieval with MPEG-7 content descriptors. *IEEE Trans. Neural Networks: Special Issue on Intelligent Multimedia Processing*, 13(4): 841–853, Jul 2002.

[113] J.T. Laaksonen, J.M. Koskela, and E. Oja. Class distributions on SOM surfaces for feature extraction and object retrieval. *Neural Networks*, 17:1121–1133, 2004.

[114] G. Sheilholeslami, W. Chang, and A. Zhang. SemQuery: Semantic clustering and querying on heteroge- neous features for visual data. *IEEE Trans. Knowledge and Data Engineering*, 14(5):988–1002, Sep–Oct 2002.

[115] W. Wang, Y. Wu, and A. Zhang. SemView: A semantic-sensitive distributed image retrieval system. In *Proc. 4th National Conf. on Digital Government Research*, Boston, May 2003.

[116] J. R. Smith, S. Basu, C.-Y. Lin, M. Naphade, and B. Tseng. Integrating features, models, and seman- tics for content-based retrieval. In *Proc. NSF Workshop in Multimedia Content-Based Indexing and Retrieval*, Sep 2001.

[117] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37:1757–1771, 2004.

[118] W. Pedrycz. Algorithms of fuzzy clustering with partial supervision. *Pattern Recognition Letter*, 13: 13–20, 1985.

[119] W. Pedrycz and J. Waletzky. Fuzzy clustering with partial supervision. *IEEE Trans. Systems, Man, and Cybernetics–Part B: Cybernetics*, 27(5):787–795, Oct 1997.

[120] R. Weber, H.-J. Shek, and S. Blott. A quantitative analysis and performance study for similarity search methods in high-dimensional spaces. In *Proc. Conf. on Very Large Databases*, New York, NY, USA, 1998.

[121] U. Guntzer, W.-T. Balke, and W. Kiebling. Optimizing multi-feature queries for image databases. In *Proc. Conf. on Very Large Databases*, Aug 2000.

[122] J.C. Harsanyi and C.-I. Chang. Hyperspectral image classification and dimensionality reduction: An orthogonal subspace projection approach. *IEEE Trans. Geoscience and Remote Sensing*, 32(4):779–785, Jul 1994.

[123] M. Kokare, B.N. Chatterji, and P.K. Biswas. Dimensionality reduction of tree structured wavelet transform texture features for content based image retrieval. In *Proc. 7th IEEE Int'l Conf. on Control, Automation, Robotics, and Vision*, volume 3, pages 1647–1625, Dec 2002.

[124] P. Wu, B.S. Manjunath, and H.D. Shin. Dimensionality reduction for image retrieval. In *Proc. IEEE Int'l Conf. on Image Processing*, volume 3, pages 1647–1652, Sep 2000.

[125] A. Dorado and E. Izquierdo. Fuzzy color signatures. In *Proc. IEEE Int'l Conf. on Image Processing*, volume 1, pages 433–436, Rochester, New York, USA, Sep 2002.

[126] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7 Multimedia Content Description Interface*. J. Wiley, New York, 2002.

[127] E. Izquierdo, I. Damnjanovic, P. Villegas, X. Li-Qun, and S. Herrmann. Bringing user satisfaction to media access: The 1st busman project. In *Proc. 8th IEEE Int'l Conf. on Information Visualisation*, pages 444–449, London, UK, Jul 2004.

[128] J. R. Smith and B. Lugeon. A visual annotation tool for multimedia content description. In *Proc. SPIE Photonics East, Internet Multimedia Management Systems*, Nov 2000.

[129] S.-F. Chang, T. Sikora, and A. Puri. Overview of the MPEG-7 standard. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):688–695, Jun 2001.

[130] B.S. Manjunath, J.-R. Ohm, V.V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans. Circuits and Systems for Video Technology*, 11(6):703–715, Jun 2001.

[131] H. Eidenberger. How good are the visual MPEG-7 features? In *Proc. Visual Communications and Image Processing*, volume 5150 of *Proc. of SPIE*, pages 476–488, 2003.

[132] H. Eidenberger. Statistical analysis of content-based mpeg-7 descriptors for image retrieval. *Multimedia Systems*, 10(2):84–97, Aug 2004.

[133] T. Ojala, M. Aittola, and E. Matinmikko. Empirical evaluation of MPEG-7 XM color descriptors in content-based retrieval of semantic image categories. In *Proc. 16th Int'l Conf. on Pattern Recognition*, volume 1, pages 701–706, 2002.

[134] T. Ojala, T. M¨aenp¨a¨a, J. Viertola, J. Kyllo¨nen, and M. Pietika¨inen. Empirical evaluation of MPEG-7 texture descriptors with a large-scale experiment. In *Proc. 2nd Workshop on Texture Analysis and Synthesis*, pages 99–102, 2002.

[135] P. Stanchev, G. Amato F. Falchi, C. Gennaro, F. Rabitti, and P. Savino. Selection of mpeg-7 image features for improving image similarity search on specific data sets. In *Proc. 7-th IASTED Int'l Conf. on Computer Graphics and Imaging*, pages 395–400, 2004.

[136] I.K. Fodor. A survey of dimension reduction techniques (ucrl-id-148494). Technical report, Center for Applied Scientific Computing Lawrence Livermore National Laboratory, Livermore, CA, Jun 2002.

[137] C. Wild and G. Seber. *Chance Encounters: A First Course in Data Analysis and Inference*. J. Wiley, New York, 2000.

[138] K.Y. Yeung and W.L. Ruzzo. Principal component analysis for clustering gene expression data. *Bioin- formatics*, 17(9):763–774, 2001.

[139] H. Liu and S.T. Huang. Evolutionary semi-supervised fuzzy clustering. *Pattern Recognition Letters*, 24: 3105–3113, 2003.

[140] J. Dunn. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, 3:32–57, 1973.

[141] J. Bezdek. *Pattern recognition with fuzzy objective function algorithms*. Plenum Press, New York, NY, USA, 1981.

[142] A.M. Bensaid, L.O. Hall, J.C. Bezdek, and L. P. Clarke. Partially supervised clustering for image segmentation. *Pattern Recognition*, 29(5):859–871, 1996.

[143] W. Pedrycz. *Knowledge-based clustering: from data to information granules*. J. Wiley, New York, NY, USA, 2005.

[144] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *J. Intell. Inf. Syst.*, 17(2-3):107–145, 2001. ISSN 0925-9902.

[145] M. Roubens. Fuzzy clustering algorithms and their cluster validity. *European Journal of Operational Research*, 10:294–301, 1982.

[146] X.L. Xie and G. Beni. A validity measure for fuzzy clustering. *IEEE Trans. Pattern Analysis and Machine Learning*, 13:841–847, 1991.

[147] Y. Xie, V.V. Raghavan, and X. Zhao. 3m algorithm: finding an optimal fuzzy cluster scheme for proximity data. In *IEEE Int'l Conf. on Fuzzy Systems*, volume 1, pages 627–632, May 2002.

[148] W. Pedrycz. Fuzzy sets in pattern recognition: methodology and methods. *Pattern Recognition*, 23 (1-2):121–146, 1990.

[149] C. Borgelt and R. Kruse. Shape and size regularization in expectation maximization and fuzzy clustering. In *Proc. 8th European Conf. on Principles and Practice of Knowledge Discovery in Databases*, pages 52–62, Germany, 2004. Springer-Verlag.

[150] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3): 264–323, 1999.

[151] J. Bezdek. A convergence theorem for the ISODATA clustering algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2(1):1–8, 1980.

[152] L. I. Kuncheva. How good are fuzzy if-then classifiers? *IEEE Trans. Systems, Man, and Cybernetics– PartB: Cybernetics*, 30(4):501–509, 2000.

[153] A. Smeaton, W. Kraaij, P. Over, and J. Arlandis (Coord.). Trecvid: Trec video retrieval evaluation. http://www-nlpir.nist.gov/projects/trecvid/, 2003.

[154] Y. Jin and B. Sendhoff. Extracting interpretable fuzzy rules from rbf networks. *Neural Processing Letters*, 17(2):149–164, 2003.

[155] G. Klir and T. Folger. *Fuzzy Sets, Uncertainty and Information*. Prentice-Hall, Englewood Cliffs, NJ, 1988.

[156] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.*, 6:707–710, 1966.

[157] S. Guha, R. Rastogi, and K. Shim. Cure: an efficient clustering algorithm for large databases. In *Proc. of the 1998 ACM SIGMOD Int'l Conf. on Management of Data*, pages 73–84, New York, NY, USA, 1998. ACM Press.

[158] W.J. Doll and T. Gholamreza. The measurement of end user computing satisfaction. *MIS Quarterly*, 12(2):259–274, June 1988.

[159] P. Whaite and F.P. Ferrie. From uncertainty to visual exploration. *IEEE Trans. on Pattern Analysis Machine Intelligence*, 13(10):1038–1049, Oct 1991.

[160] D.W. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with nonmetric distances: Image retrieval and class representation. *IEEE Trans. Pattern Analysis Machine Intelligence*, 22(6):583–600, Jun 2000.

[161] W. Lam, C-K Keung, and D. Liu. Discovering useful concept prototypes for classification based on filtering and abstraction. *IEEE Trans. Pattern Analysis Machine Intelligence*, 24(8):1075–1090, Aug 2002.

[162] S-C. Zhu. Statistical modeling and conceptualization of visual patterns. *IEEE Trans. Pattern Analysis Machine Intelligence*, 25(6):691–712, June 2003.

[163] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. *Advances in Knowledge Discovery and Data Mining*, chapter Fast discovery of association rules, pages 307–328. AAAI/MIT Press, 1996.

[164] T. Kliegr, K. Chandramouli, J. Nemrava, V. Svatek, E. Izquierdo, “Combining Image Captions and Visual Analysis for Image Concept Classification”, *9th International Workshop on Multimedia Data Mining*, Las Vegas, NV, 24-27 August 2008, Pages 817

[165] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In *Proc. ACM SIGMOD*, pages 1–12, 2000.

[166] L. Zadeh. The calculus of fuzzy if/then rules. *AI Expert*, 7(3):23–28, Mar 1992.

Appendix A

The Wilcoxon Test

The aim of proposed feature vector structure is to retain the topology of the original descriptor elements so that one could directly relate them to the original descriptor settings.

Considering a two-class problem, say ω and not(ω), where the class ω is in the simplest case a function

$$\omega : \Re^p \mapsto \{0, 1\} \tag{A.1}$$

that partitions the feature space X into a set $Y = \omega^{-1}(1)$, containing the positive examples of ω, and its complement $\overline{Y} = X - Y$, containing the negative examples of ω. Each example is a pair $< \mathbf{x}, \ell$ where ℓ is a label defined as

$$\ell = \omega(\mathbf{x}) = \begin{cases} 1, & \mathbf{x}_i \in Y \\ 0, & \mathbf{x}_i \notin Y \end{cases} . \tag{A.2}$$

Then each value of the image description is analysed individually to find salient features regarding to ω. Individual feature selection is carried out applying statistical hypothesis testing.

A co-ordinate of the feature vector is marked as salient -or retained feature when the *two-sided alternative hypothesis* H_1 is true. Conversely, we are looking for samples in which the *null hypothesis* is false. In such a case, the distribution of the feature in Y is different to that in $\overline{Y}$, which can be written as

$$H_1 : \mathbf{y} \neq \overline{\mathbf{y}} \ , \tag{A.3}$$

where $\mathbf{y} = Y[k]$, $\overline{\mathbf{y}} = \overline{Y}[k]$ and $1 \leq k \leq p$ is the co-ordinate.

In the two-sided alternative there is not strong prior reason for expecting a shift in a particular direction [137]. Otherwise, the possibilities are:

$$H_1 : \mathbf{y} > \overline{\mathbf{y}}, \mathbf{y} \text{ is shifted to the right of } \overline{\mathbf{y}} \ , \tag{A.4}$$

$$H_1 : \mathbf{y} < \overline{\mathbf{y}}, \mathbf{y} \text{ is shifted to the left of } \overline{\mathbf{y}} \ , \tag{A.5}$$

and the null hypothesis is

$$H_0 : \mathbf{y} = \overline{\mathbf{y}}, \mathbf{y} \text{ is the same as that in } \overline{\mathbf{y}} \ . \tag{A.6}$$

The Wilcoxon test is based upon ranking the $N_Y + N_{\overline{Y}} = N_X$ features of the combines sample. Each feature value has a rank established in ascending order: the smallest has rank 1, the 2nd smallest rank 2, etc. The Wilcoxon rank-sum test statistic is the sum of the ranks for features from one of the samples [137].

Tab. A.1 describes the rank-sum calculation for a small sample of 20 descriptions. Vectors $\mathbf{y}$ and $\overline{\mathbf{y}}$ contain instances from the 1st feature of each vector. As can be observed, each feature in a tie is associated to an average rank.

An estimated P-*value* close to zero signals that the null hypothesis is false and consequently the feature is marked as salient.

Tab. A.1: Rank-sum calculation

Given	$\mathbf{y}$	14	14	10	12	25	20	10	14	15	15
	$\overline{\mathbf{y}}$	12	24	21	17	9	24	12	13	16	16
Calculate the Rank											
Feature value:	9	10	10	12	12	12	13	14	14	14	...
Vector:	$\overline{\mathbf{y}}$	$\mathbf{y}$	$\mathbf{y}$	$\mathbf{y}$	$\overline{\mathbf{y}}$	$\overline{\mathbf{y}}$	$\overline{\mathbf{y}}$	$\mathbf{y}$	$\mathbf{y}$	$\mathbf{y}$	...
Rank:	1	2	3	4	5	6	7	8	9	10	...
Average Rank:	1	2.5	2.5	5	5	5	7	9	9	9	...
Compute the Sum $w_{\mathbf{y}}$											
For the $\mathbf{y}$ group, $w_{\mathbf{y}} = 2.5 + 2.5 + 5 + 9 + 9 + 9 + 11.5 + 16 + 19 = 95$											
Estimate the P-*value*											
P-*value* $(W_{\mathbf{y}} \leq 95) = 0.1357$ and P-*value*$(W_{\mathbf{y}} \geq 95) = 0.2713$											

For larger samples [137], the distribution of $W_{\mathbf{y}}$ as if it were `Normal`$(\mu_{\mathbf{y}}, \sigma_{\mathbf{y}})$, where

$$\mu_{\mathbf{y}} = \frac{N_Y(N_Y + N_{\overline{Y}} + 1)}{2} \tag{A.7}$$

and

$$\sigma_{\mathbf{y}} = \sqrt{\frac{N_Y N_{\overline{Y}}(N_Y + N_{\overline{Y}} + 1)}{12}} \tag{A.8}$$

More precisely,

$$\mathtt{pr}(W_{\mathbf{y}} \geq w_{\mathbf{y}}) \approx \mathtt{pr}(Z \geq z) \ , \tag{A.9}$$

where

$$z = \frac{(W_{\mathbf{y}} - \mu_{\mathbf{y}})}{\sigma_{\mathbf{y}}} \tag{A.10}$$

and

$$Z \sim Normal(0,1) \ \ . \tag{A.11}$$

For the two-sided test, if $w_{\mathbf{y}}$ falls in the lower tail

$$P - \mathit{value} = 2\mathtt{pr}(W_{\mathbf{y}} \leq w_{\mathbf{y}}) \ \ , \tag{A.12}$$

whereas if $w_{\mathbf{y}}$ is in the upper tail

$$\mathtt{P}{-}\mathit{value} = 2\mathtt{pr}(W_{\mathbf{y}} \geq w_{\mathbf{y}}) \ \ . \tag{A.13}$$

Appendix B

Rule Generation

B.1 Frequent Pattern Mining

Continually recurring relationships between low-level features and high-level concepts are found applying a knowledge discovery technique. This technique derives from a solution of the frequent pattern mining problem known as *association rule mining* (Agrawal et al. [163]). An *association rule* can be formally described as follows:

Let $I = \{i_1, \ldots, i_n\}$ be a set of literals called *items* and $T = \{t_1, \ldots, t_m\}$ be a set of transactions where each transaction t_k is a set of items such that $t_k \subset I$.

Let $t_k = X$ be a transaction with a set of items corresponding to *itemset* X. Transaction set T is said to *contain* itemset Y if $Y \subset X$.

The support of itemset X, denoted as $sup(X)$, is the number of transactions in T containing X.

Given a user-specified *support threshold minsup*, X is called a *frequent itemset* or *frequent pattern* if $sup(X) \geq minsup$. The problem of mining frequent itemsets is to find the complete set of frequent itemsets in T with respect to a given support threshold *minsup*.

An association rule is an implication of the form $X \Rightarrow Y$ where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$.

The rule $X \Rightarrow Y$ holds in the transaction set T with support s if $s\%$ of transactions in T contain $X \cup Y$. In other words, if $sup(X \cup Y) \geq minsup$.

The rule $X \Rightarrow Y$ holds in T with confidence c if $c\%$ of transactions in T containing X also contain Y. In other words, if $\frac{sup(X \cup Y)}{sup(X)} \geq minconf$

Given a set T of transactions, the problem of mining association rules is to generate all association rules having support and confidence greater than the minimum user-defined support *minsup* and the minimum user-specified confidence *minconf*.

B.2 An Abstract Example

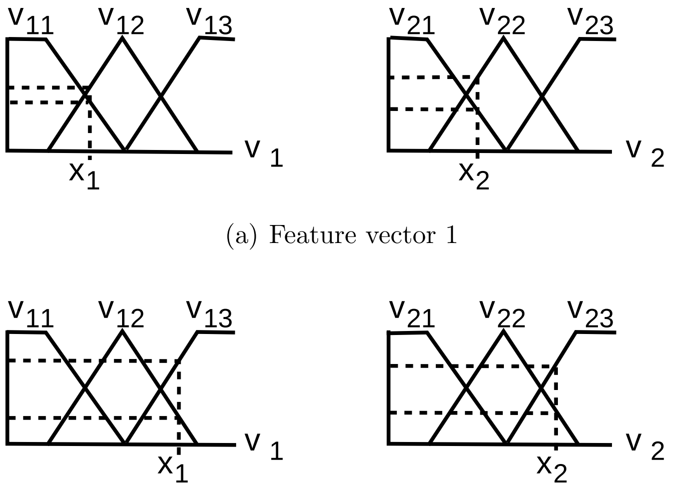
The following abstract example illustrates the usage of association rule mining to generate a built-in knowledge rule base.

Let $\mathbf{x}_i = [x_1, x_2]^T$ be a feature vector extracted from an image **i**. Let ℓ_j a symbol representing a high-level concept associated by visual interpretations of **i**'s content.

Let $v = \{v_1, v_2\}$ be a set of fuzzy variables associated to each feature, $v_1 = \{v_{11}, v_{12}, v_{13}\}$, and $v_2 = \{v_{12}, v_{12}, v_{12}\}$ be fuzzy sets defined for each fuzzy variable.

Assuming two feature vectors in the space, $\mathbf{x}_1, \mathbf{x}_2$, the transaction set T is populated combining information extracted from the fuzzy domain and provided by the interpreter.

The first step is to instantiate the fuzzy variables as depicted in Fig. B.1 and Tab. B.1. The second step is to filter out fuzzy values equal to zero. The third step is to concatenate names of remaining fuzzy sets and the concept attached to the image where the features were extracted.



(a) Feature vector 1

(b) Feature vector 2

Fig. B.1: Fuzzy variables v_1, v_2 are instantiated with membership degrees of each feature value to corresponding fuzzy sets

Tab. B.1: Fuzzy values

Vector	Feature value (i)	v_{i1}	v_{i2}	v_{i3}
1	x_1	0.4200	0.5800	0.0000
1	x_2	0.6700	0.3300	0.0000
2	x_1	0.0000	0.2000	0.8000
2	x_2	0.0000	0.2910	0.7090

The result is a set of transactions as shown in Tab. B.2. Therefore, transactions consists of two parts: left-part groups fuzzy sets' names and right-part is a concept-related symbol (concept).

Tab. B.2: Transaction set

Vector	Transaction
1	$t_1 = v_{11}\ v_{21}\ \ell_j$
1	$t_2 = v_{11}\ v_{22}\ \ell_j$
1	$t_3 = v_{12}\ v_{21}\ \ell_j$
1	$t_4 = v_{12}\ v_{22}\ \ell_j$
2	$t_5 = v_{12}\ v_{22}\ \ell_j$
2	$t_6 = v_{12}\ v_{23}\ \ell_j$
2	$t_7 = v_{13}\ v_{22}\ \ell_j$
2	$t_8 = v_{13}\ v_{23}\ \ell_j$

Given $X = v_{12}\, v_{22}$ and $Y = \ell_j$, the rule $X \Rightarrow Y$ holds in the transaction set $T = \{t_1, \ldots, t_8\}$ has support $s = 0.25$ because 25% of transactions in T contain $X \cup Y$. The rule $X \Rightarrow Y$ holds in T has confidence $c = 1.00$ because 100% of transactions in T containing X also contain Y.

Then, the following inference rule is added to the built-in knowledge rule base:

$$\text{IF } \mathbf{v}_1 \ \text{ is } \ v_{12} \text{ AND } \mathbf{v}_2 \ \text{ is } \ v_{22} \text{ THEN classify as } \ell_j \tag{B.1}$$

The rule mining process has been tested by fixing the minconf to the value 1.0 and varying the minsup with a step=0.01 until a neck is reached. Fig. B.2 shows the behaviour of the process and the location of the neck.

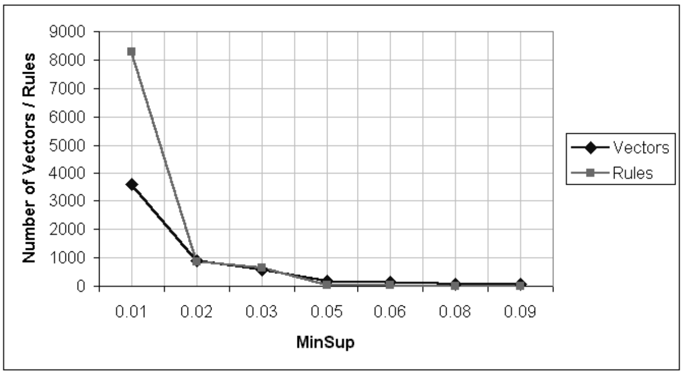


Fig. B.2: Neck or stop condition for the rule mining process

The association rules mining process generates all possible combinations of items in the set of transactions satisfying the minsup and minconf conditions. Therefore, a filter is required to discriminate meaningful rules from meaningless ones.

FP-Tree algorithm proposed by Han et al. [165] is used for mining the association rules. According to recent publications, FP-Tree is currently the most efficient algorithm for mining frequent patterns.

Appendix C

Fuzzy Reasoning Model

C.1 Fuzzy Systems

A system implementing the fuzzy reasoning model involves three modules: fuzzification, fuzzy inference, and defuzzification. The fuzzy inference module uses a built-in base of if-then rules. This kind of systems supports both multiple inputs with a single output (MISO) and multiple inputs with multiple outputs (MIMO). A system overview is depicted in Fig. C.1.

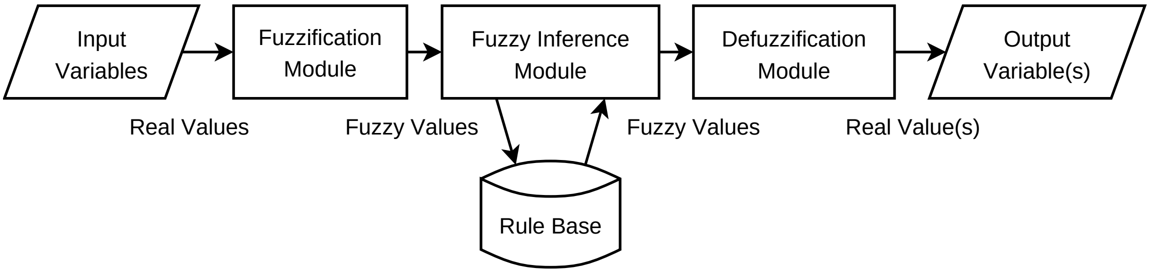


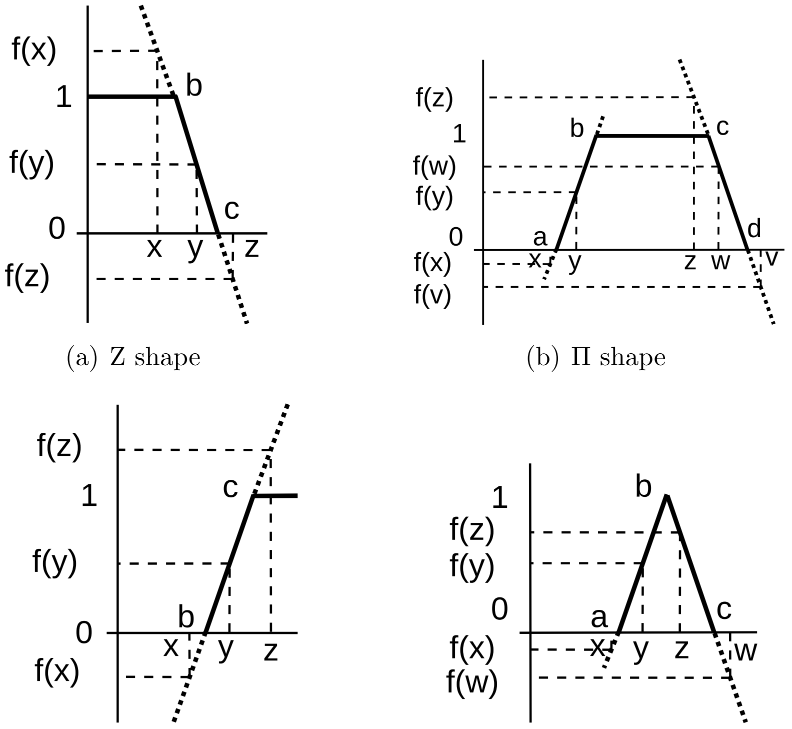
Fig. C.1: Overview of a system implementing a fuzzy reasoning model

Each input variable is represented in the fuzzy domain using a number of fuzzy sets. These fuzzy sets are ascribed to one or more rules in arrays called *antecedents*. Similarly, each output variable is associated to one or more rules in arrays called *consequents*. The *fuzzy inference* works with rules of the form

$$\texttt{IF}\ \textit{antecedents}\ \texttt{THEN}\ \textit{consequents} \tag{C.1}$$

C.2 Fuzzification

The *fuzzification* module maps input variables from the real domain into the fuzzy domain. This mapping is carried out by computing the degree of truth (or membership values) of an input variable to each fuzzy set associated to it. Membership values can be approximated by piecewise linear functions. Typically, four types of linear functions are used: Z, S, Π, and Λ. These functions' names obey to the shapes they look like. Fig. C.2 shows the usage of piecewise linear functions when computing membership degrees. Tab. C.1-Tab. C.4 describes the normalization of membership values into the interval [0,1].



(a) Z shape

(b) Π shape

(c) S shape

(d) Λ shape

Fig. C.2: Piecewise linear functions used to compute membership degrees. v, w, x, y, z are real-valued variables. $f(\cdot)$ are piecewise linear (membership) functions. a, b, c, d are co-ordinates specifying boundaries shaping the functions

Tab. C.1: Z-shape membership function estimation

i	Equation	$\mu(i)$
x	$\mu(x) = max(min(f(x), 1), 0)$	1
y	$\mu(y) = max(min(f(y)), 1), 0)$	$f(y)$
z	$\mu(z) = max(min(f(z), 1), 0)$	0

Tab. C.2: S-shape membership function estimation

i	Equation	$\mu(i)$
x	$\mu(x) = max(min(f(x), 1), 0)$	0
y	$\mu(y) = max(min(f(y)), 1), 0)$	$f(y)$
z	$\mu(z) = max(min(f(z), 1), 0)$	1

Tab. C.3: Π-shape membership function estimation

i	Equation	$\mu(i)$
x	$\mu(x) = max(min(f(x), 1), 0)$	0
y	$\mu(y) = max(min(f(y)), 1), 0)$	$f(y)$
z	$\mu(z) = max(min(f(z)), 1), 0)$	1
w	$\mu(w) = max(min(f(w)), 1), 0)$	$f(w)$
v	$\mu(v) = max(min(f(v), 1), 0)$	0

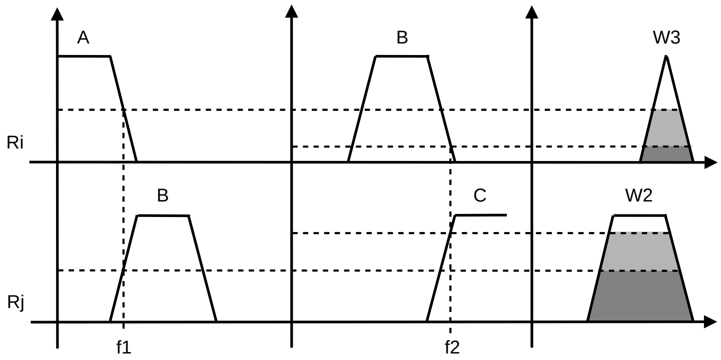
Tab. C.4: Λ shape membership function estimation

i	Equation	$\mu(i)$
x	$\mu(x) = max(min(f(x), 1), 0)$	0
y	$\mu(y) = max(min(f(y)), 1), 0)$	$f(y)$
z	$\mu(z) = max(min(f(z)), 1), 0)$	$f(z)$
w	$\mu(w) = max(min(f(v), 1), 0)$	0

C.3 Fuzzy Inference

Inference rules are not a free form of natural language (Zadeh [166]). They are limited to a set of linguistic terms and a strict syntax. Each antecedent (condition associated to the IF-part) of a rule corresponds to a specific value of a fuzzy input. Antecedents are instantiated by the fuzzification module. Each consequent (conclusion associated to the THEN-part) of a rule corresponds to a fuzzy output.

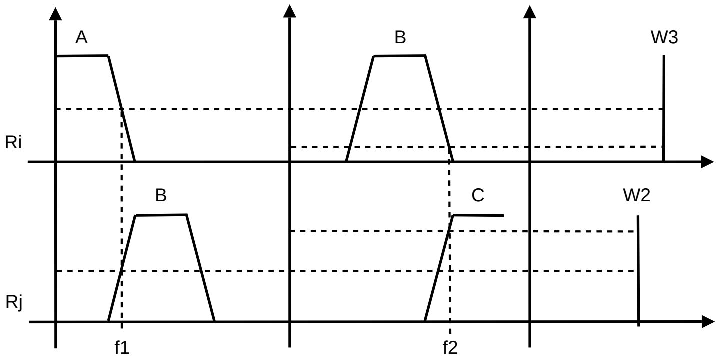
Depending on characteristics of the consequents, the fuzzy inference system (FIS) can be adapted to function as a Mamdani FIS, supporting rules with both antecedents and conse-quents defined in a fuzzy domain (i.e. fuzzy sets) as is depicted in the illustrative example presented by Fig. C.3.



R_i IF (F_1 is $\widetilde{A}$) AND (F_2 is $\widetilde{B}$) THEN Classify as W_3)
R_j IF (F_1 is $\widetilde{B}$) AND (F_2 is $\widetilde{C}$) THEN Classify as W_2)

Fig. C.3: Mamdani fuzzy inference system: antecedents and consequents consist of fuzzy sets

The FIS can also be adapted to function as a Sugeno FIS, which uses rules with consequents defined in a discrete domain (i.e crisp sets). This kind of FIS is suitable for implementing typical classification tasks where each element fo the crisp set represents a class. This mode is illustrated in Fig. C.4.



R_i IF (F_1 is $\widetilde{A}$) AND (F_2 is $\widetilde{B}$) THEN (Classify as W_3)
R_j IF (F_1 is $\widetilde{B}$) AND (F_2 is $\widetilde{C}$) THEN (Classify as W_2)

Fig. C.4: Sugeno fuzzy inference system: consequents consist of crisp sets

The fuzzy inference module combines antecedent values to assign a value to consequents attached to the participating rules. A rule participates in the inference process when all its antecedents have been instantiated.

At this point, a number of antecedents can be true in different degrees. Using an *operator of aggregation*, antecedents are combined to determine the value of the rule (Klir and Folger [155]). This procedure is applied to all participating rules. Tab. C.5 lists several operators of aggregation (fuzzy combinations or T-norms) frequently found in the literature.

Tab. C.5: Operators of aggregation (T-norm)

Operator	Definition
Minimum	$T(x_i, x_j) = min(x_i, x_j)$
Algebraic Product	$T(x_i, x_j) = x_i x_j$
Bounded Product	$T(x_i, x_j) = max(0, x_i + x_j - 1)$
Drastic Product	$T(x_i, x_j) = \begin{cases} x_i & x_j = 1 \\ x_j & x_i = 1 \\ 1 & x_i x_j < 1 \end{cases}$

It is possible to have consequents in different rules associated to the same output variable. To determine the value for this consequent, an *operator of composition* is used. Tab. C.6 lists the type of operator of composition (fuzzy combinations or T-conorms) frequently found in the literature.

C.4 Defuzzification

The *defuzzification* module combines the fuzzy values of an output variable to obtain its correspondence in the real domain. In the former fuzzy systems, the maximum value was assigned. Currently, this method is considered very poor, due to the fact that it ignores the

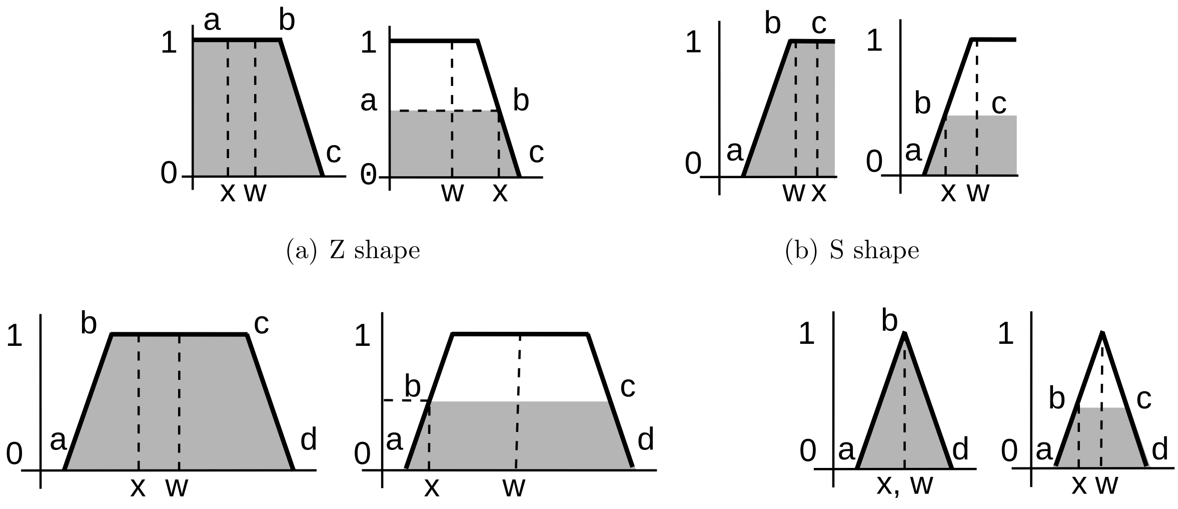
Tab. C.6: Operators of composition (T-conorm or S-norm)

Operator	Definition
`Maximum`	$T(x_i, x_j) = max(x_i, x_j)$
`Algebraic Sum`	$T(x_i, x_j) = x_i + x_j - x_i x_j$
`Bounded Sum`	$T(x_i, x_j) = min(1, x_i + x_j)$
`Drastic Sum`	$T(x_i, x_j) = \begin{cases} x_i & x_j = 0 \\ x_j & x_i = 0 \\ 1 & x_i x_j > 0 \end{cases}$

contribution of all the rules. Therefore, there are other alternatives to calculate the output values. Typically, a method called *Centre of Area*, which combines fuzzy values using a weighted average, is used to obtain the crisp value is used.

$$CoA = \frac{\sum_i A_i W_i}{\sum_i A_i} \tag{C.2}$$

Fig. C.5 indicates the calculation of W (weight) and A (area) according to the shape and instance of the fuzzy set.



(a) Z shape

(b) S shape

(c) Π shape

(d) Λ shape

Fig. C.5: Calculation of center of area. W denotes the weight. x is used to illustrate two possible cases of fuzzy value, i.e. $\mu(x) = 1$ and $0 \leq \mu(x) < 1$. a, b, c, d are co-ordinates specifying boundaries shaping the area

In Eq. C.3 and Eq. C.4 area and weight computation of Z-shape membership functions is presented.

$$A = \frac{\mu(x)((c-a)+(b-a))}{2} \tag{C.3}$$

$$W = a + \frac{c-a}{2} \tag{C.4}$$

In Eq. C.5 and Eq. C.6 area and weight computation of S-shape membership functions is described.

$$A = \frac{\mu(x)((c-a)+(c-b))}{2} \tag{C.5}$$

$$W = a + \frac{c-a}{2} \tag{C.6}$$

In Eq. C.7 and Eq. C.8 area and weight computation of Π-shape membership functions is presented.

$$A = \frac{\mu(x)((d-a)+(c-b))}{2} \tag{C.7}$$

$$W = a + \frac{d-a}{2} \tag{C.8}$$

In Eq. C.9 and Eq. C.10 area and weight computation of Λ-shape membership functions is described.

$$A = \frac{\mu(x)((d-a)+(c-b))}{2} \tag{C.9}$$

$$W = a + \frac{d-a}{2} \tag{C.10}$$

Finally, values obtained applying Eq. C.2 are assigned to the output variables.

List of Notation

$\mathbf{i} = (i_{11}, i_{21}, \ldots, i_{M1}, i_{12}, \ldots, i_{MN})$	Digital image
D	Set of image descriptors
$\mathbf{d}_j\ (1 \leq j \leq N_D)$	Image description (with embedded semantic)
$\lVert\mathbf{d}_{colour}\rVert$	Number of constutient elements used by colour descriptors
$\lVert\mathbf{d}_{texture}\rVert$	Number of constutient elements used by texture descriptors
X	Set of p-dimension feature vectors
$\mathbf{x}_i\ (1 \leq i \leq N_X)$	Feature vector
$x\ (x \in \Re)$	Feature value
Ω	Set of classes
$\omega_k\ (1 \leq k \leq N_\Omega)$	Class (or semantic category)
$\mathrm{f} : X \mapsto \Omega$	Multi-class classifier
$\mathrm{f}_k : \Re^n \mapsto \{0, 1\}$	Binary classifier
L	Lexicon
$\ell_j\ (\ell_j \in L;\ \omega_k \multimap \ell_j)$	Concept-related symbol (label) associated with class ω_k
$\mathcal{L} \subset L$	Set of concept-related symbols
$\mathcal{A}$	Image annotation process
$\mathcal{S}_{(0)}$	Abstract non annotated image space
$\mathcal{S}_{(t)}$	Abstract annotated image space
$\mathcal{S}_{(0)} \circlearrowleft \mathcal{S}_{(1)} \circlearrowleft, \ldots, \circlearrowleft \mathcal{S}_{(t)}$	Annotation session
$J(X, \mathbf{V}, \mathbf{U})$	Criterion function
$\mathbf{V}$	Set of c $(2 \leq c \leq N_X)$ cluster prototypes
$\mathbf{U}$	Matrix belonging to the set of all possible fuzzy partitions $\Im$
$C = \{c_1, c_2, ldots\}$	Set of clusters
u_{ij}	Degree of membership of vector $\mathbf{v}_i$ to cluster j
$\mathbf{v}_j$	p-dimension prototype
$d^2(\cdot)$	Any distance norm expressing similarity between two arrays

$m\ (1 < m < \infty)$	Fuzzy exponent
X/E	Quotient set that forms a partition of the feature space
$\phi : X \mapsto X/E$	Clustering mechanism used as a classification function
X^d	Labeled data set
X^u	Unlabeled data set
$\mathbf{b}$	Indicator vector of (un)labeled data
$\mathbf{F}$	Matrix of known membership degrees
β	Scaling factor to keep balance between unsupervised and supervised data
δ	Criterion used in the Picard Iteration
ϵ	Maximum number of epochs, fuzzy c-means
I	Set of literals called items
s	Rule support (Association rules mining)
$v = \{v_1, v_2, \ldots, v_{N_v}\}$	Set of fuzzy variables
T	Set of transactions
$t_{i,j} =< \mathbf{x}_i, \ell_j >$	Transaction
$\mu_{\widetilde{A}}(\mathbf{x}_i) \in [0, 1]$	Membership function
$u_1, u_2, \ldots, u_{N_D}$	Receptive fields
y	RBFN output
w_j	Weights used in the RBFN model and semantic profiles
Φ	Activation matrix
$Y = \omega^{-1}(1)$	Set of positive examples of ω
$\overline{Y} = X - Y$	Set of negative examples of ω
H_0	Null hypothesis
H_1	Two-sided alternative hypothesis
$\mathbf{y} = Y[k]\ (1 \leq k \leq p)$	Vector consisting of N_Y instances of k-th feature
$\bar{\mathbf{y}} = \overline{Y}[k]\ (1 \leq k \leq p)$	Vector consisting of $N_{\overline{Y}}$ instances of k-th feature
$\mathbf{I}_{color} = \mathbf{I}_{CLD} \mid \mathbf{I}_{CSD} \mid \mathbf{I}_{SCD}$	Indexes of retained colour features
$\mathbf{I}_{texture} = \mathbf{I}_{EHD} \mid \mathbf{I}_{HTD}$	Indexes of retained texture features
$\mathbf{x} = [\mathbf{x}_{colour} \vert \mathbf{x}_{texture}]^T$	Feature vector as an aggregation of colour and texture descriptor elements
α	Confidence level