www.forgetit-project.eu

ForgetIT

Concise Preservation by Combining Managed Forgetting and Contextualized Remembering

Grant Agreement No. 600826

Deliverable D4.1

Work-package	WP4: Information Consolidation and Concentration
Deliverable	D4.1: Information analysis, consolidation and concentration for preservation - State of the Art and Approach
Deliverable Leader	Vasileios Mezaris, Papadopoulou Olga
Quality Assessor	Walter Allasia
Estimation of PM spent	5PM
Dissemination level	PU
Delivery date in Annex I	31-07-2013 (M6)
Actual delivery date	31-07-2013
Revisions	0
Status	Final
Keywords:	textual and multimedia similarity, redundancy, semantic multimedia analysis, information condensation and consolidation

Disclaimer

This document contains material, which is under copyright of individual or several ForgetIT consortium parties, and no copying or distributing, in any form or by any means, is allowed without the prior written agreement of the owner of the property rights.

The commercial use of any information contained in this document may require a license from the proprietor of that information.

Neither the ForgetIT consortium as a whole, nor individual parties of the ForgetIT consortium warrant that the information contained in this document is suitable for use, nor that the use of the information is free from risk, and accepts no liability for loss or damage suffered by any person using this information.

This document reflects only the authors' view. The European Community is not liable for any use that may be made of the information contained herein.

© 2013 Participants in the ForgetIT Project

List of Authors

Olga Papadopoulou / CERTH
Vasileios Mezaris / CERTH
Mark A. Greenwood / USFD
Berker Logoglu / TT

Contents

List of Authors 3

Contents 4

Executive Summary 6

1 Introduction 7

2 Glossary 9

3 Analysis of Textual Similarity and Redundancy 11

3.1 Document Similarity Assessment . . . 11

3.1.1 Semantic Text Analysis . . . 11

3.2 Textual Summarization and Redundancy Removal . . . 12

3.3 Planned ForgetIT Approach . . . 15

4 Analysis of Multimedia Quality, Similarity and Redundancy 16

4.1 Image Quality Assessment . . . 16

4.2 Video Quality Assessment . . . 19

4.3 Multimedia Similarity . . . 21

4.3.1 Visual Information Descriptors . . . 21

4.3.2 Multimedia Metadata . . . 24

4.3.3 Similarity measures . . . 25

4.4 Clustering Multimedia Items . . . 26

4.5 Redundancy of Media Items and Media Fragments . . . 28

4.6 Planned ForgetIT Approach . . . 29

5 Semantic Multimedia Analysis for Condensation 31

5.1 Concept Detection . . . 31

5.1.1 Low-level feature extraction . . . 31

5.1.2 Machine Learning for Concept Detection 33
5.2 Face Detection and Clustering 34
5.3 Event Detection 35
5.4 Planned ForgetIT Approach 36

6 Information Condensation and Consolidation **39**
6.1 Textual Content Condensation and Presentation 39
6.1.1 Semantic Redundancy and Diversity Analysis 40
6.2 Image and Video Condensation 40
6.3 Planned ForgetIT Approach 42

7 Conclusion **44**

References **45**

Executive summary

The present document reports on the current state of the art in textual and multimedia content analysis for condensation, from the perspective of information preservation, and sketches our first thoughts on the corresponding approaches that WP4 should adopt or further extend, in order to support the overall goals of the project and requirements summarised in D9.1. The topics covered by this document include:

- Analysis of textual similarity and redundancy
- Analysis of multimedia similarity and redundancy
- Semantic multimedia analysis for condensation
- Information condensation and consolidation

The structure of the document is as follows: Section 3 addresses textual similarity and redundancy assessment. It includes an outline of the state of the art in document similarity assessment and text semantic analysis techniques, a review of text summarization techniques, and a short sketch of the directions that we will follow in ForgetIT for textual information processing. Section 4 discusses the processing of non-textual multimedia content, i.e. image and video. It covers image and video quality assessment, the evaluation of the similarity between two media items and the clustering and detection of potentially redundant media items, and concludes with an outline of the planned approach of ForgetIT in these areas. Section 5 presents the state of the art on analysis of multimedia contents (detection of concepts, faces and events and their clusterization) and our planned approach with respect to the deeper analysis of multimedia content. Section 6 discusses the possible approaches for using the results of text and multimedia analysis towards information condensation, i.e. the generation of summaries of media collections. Section 7 gives some final remarks and conclusions.

This document represents the input source for deliverables D4.2, D4.3 and D4.4 where the selected technologies and methodologies for text and multimedia analysis and condensation will be implemented and evaluated in the project test set.

1 Introduction

As a result of the explosive development of digital media capture, transmission and consumption devices and networks that we have been experiencing in the last years, and the prevalence of digital media in our everyday lives, the amount of multimedia data that is created and stored every day has reached unprecedented levels, and continues to increase. For instance, surveys show that the amount of pictures taken every year is counted in billions. Supporting the users or owners of these digital media items in identifying the best preservation options for their content, given any inevitable storage space and preservation cost limitations is among the goals of ForgetIT.

Within ForgetIT, the goal of WP4 is to develop methods that enable the condensation of content, both textual and multimedia. This is intended to support the gradual forgetting approach, where content is presented with varying levels of details with the passage of time or ceasing importance. The condensation and consolidation of content will rely on content analysis methods of varying degrees of complexity, including techniques for the identification of redundancy, the detection of similarity through content analysis, and deeper semantic analysis. In the condensation process, we will also take into account aspects such as the quality, diversity and coverage of the multimedia items.

The present document reports on the current state-of-the-art in the research areas that are of interest to WP4, i.e. textual and multimedia content analysis for condensation, from the perspective of information preservation, and based on this review it goes on to sketch our thoughts on the approaches that WP4 should adopt or further extend in these areas, in order to support the overall goals of the project. More specifically, the topics covered by this document include:

- Analysis of textual similarity and redundancy
- Analysis of multimedia similarity and redundancy
- Semantic multimedia analysis for condensation
- Information condensation and consolidation

The structure of the document is as follows: Section 3 addresses the topics of textual similarity and redundancy assessment. It starts with an outline of the state-of-the-art in document similarity assessment techniques, including techniques based on the semantic analysis of textual content. Then proceeds to presenting in some more detail techniques for the summarization of textual information, and concludes with a short sketch of the directions that we will follow in ForgetIT for textual information processing, stressing that the selection of techniques that we will adopt depends a lot on the specifics of the project's use cases. Section 4 moves on to discuss the processing of non-textual multimedia content, most notably image and video. It addresses the issues of image and video quality assessment, the evaluation of the similarity between two media items (involving the extraction of features from the content and their assessment using appropriate similarity

measures), and the clustering and detection of potentially redundant media items. Section 4 concludes with an outline of the planned approach of ForgetIT in these areas. Section 5 presents the state of the art on analysis of multimedia contents and our planned approach with respect to the deeper analysis of multimedia content. Such analysis includes concept detection in image and video content, face detection and clustering, as well as event detection. Section 6 discusses the possible approaches for using the results of analysis techniques such as those described in Sections 3 to 5 towards textual and visual information condensation, so that good summaries of media item collections can be produced and used in the overall ForgetIT approach. Finally, some concluding remarks are given in Section 7.

This document represents the input source for deliverables D4.2, D4.3 and D4.4 where the selected technologies and methodologies for text and multimedia analysis and condensation will be implemented and evaluated in the project test set.

2 Glossary

List of important terms and acronyms presented in the document:

ANSI	American National Standards Institute
AP	Affinity Propagation
BoW	Bag of Words
CEDD	Color and Edge Directivity Descriptor
CHOG	Compressed HOG (Histogram of Oriented Gradients)
CLD	Color Layout Descriptor
CSD	Color Structure Descriptor
CSSD	Curvature Scale Space Descriptor
DCD	Dominant Color Descriptor
DCT	Discrete Cosine Transform
DMOS	Difference Mean Opinion Score
DURF	Dense SURF (Speeded Up Robust Feature Descriptor)
DVQ	Digital Video Quality
EDH	Edge Direction Histogram
EHD	Edge Histogram Descriptor
EOH	Edge Orientation Histogram
EXIF	EXchangeable Image file Format
FCM	Fuzzy C-Mean (for clustering)
FCTH	Fuzzy Color and Texture Histogram
FFT	Fast Fourier Transform
FR IQA	Image Quality Assessment with Full Reference
GBR	Gabor Texture Descriptor
GCM	Global Color Moments
GMM	Gaussian mixture model
GoF	Group of Frames (in compressed video)
GoP	Group of Pictures (in compressed video)
GPS	Global Position System
H2MP	Homogeneous and Heterogeneous Message Propagation
HOG	Histogram of Oriented Gradients
HTD	Homogeneous Texture Descriptor
HVS	Human Visual System
ITS	Institute for Telecommunication Services
IQA	Image Quality Assessment
JPEG	Joint Photographic Experts Group
JP2K	JPEG2000 lossless compression standard
KNN	k-Nearest Neighbour
LAB	Locally Assembled Binary
LBP	Local Binary Patterns
LDP	Local-Difference-Patterns
LFCC	Linear Frequency Cepstral Coefficients
LSH	Locality Sensitive Hashing

LSVM	Linear SVM (Support Vector Machine)
MBH	Motion Boundary Histogram
MOS	Mean Opinion Score
MOVIE	Motion-based Video Integrity Evaluation
MPEG	Motion Picture Expert Group
MPEG-7	Standard ISO/IEC 15938 (for digital item metadata representation)
MSE	Mean Squared Error
MSG	Modulation Spectrogram
MS-SSIM	Multi-Scale Structural SIMilarity
MVG	Multivariate Gaussian Model
NLG	Natural Language Generation
NR IQA	Image Quality Assessment with No Reference (or blind)
NR-B	No-Reference Bitstream
NR-P	No-Reference Pixel
NSS	Natural Scene Statistics
NTIA	National Telecommunications and Information Administration
NVC	Natural Visual Characteristics
PCA	Principal Components Analysis
PHOG	Pyramidal HOG (Histogram of Oriented Gradients)
PSNR	Peak Signal to Noise Ratio
PSSIM	Perceptual Structural Similarity
PVQM	Perceptual Video Quality Metric
RBF	Radial Basis Function
RGB	Red Green Blue
RMS	Root Mean Square
RR IQA	Image Quality Assessment with Reduced Reference
SAHN	Sequential Agglomerative Hierarchical Non-overlapping (clustering)
SCD	Scalable Color Descriptor
SIFT	Scale Invariant Feature Transform
SNR	Signal-to-Noise Ratio
SOM	Self Organised Map
SURF	Speeded Up Robust Feature Descriptor
SSIM	Structural SIMilarity
SQFD	Signature Quadratic distance Function
SVM	Support Vector Machine
SVR	Support Vector Regression
TBD	Texture Browsing Descriptor
TF.IDF	Term Frequency-Inverse Document Frequency
QFD	Quadratic Form Distances
VEML	Video Event Markup Language
VERL	Video Event Representation Language
VSNR	Visual Signal-to-Noise Ratio
VQA	Video Quality Assessment
VQM	Video Quality Metric
XML	extensible Markup Language
ZMD	Zernike Moment Descriptor

3 Analysis of Textual Similarity and Redundancy

3.1 Document Similarity Assessment

Determining the similarity between documents is both an important step in most automatic summarization algorithms (see Section 3.2) as well as a useful result in its own right. Most approaches to textual similarity are statistically based and operate without regard to the semantics or syntax of language. The most widely used similarity measure is the cosine measure, which is usually applied in conjunction with the term frequency-inverse document frequency, TF.IDF, weighting function [1]. The extracted features can be considered in a vector space and the similarity between vectors can be easily evaluated by the projection of one vector to the other. The cosine similarity is relatively easy to compute over a collection for which we have access to an inverted index; this makes it difficult to apply directly to web documents for which we do not have accurate term distribution information.

Whilst the cosine measure considers terms in isolation, it is possible to use other measures to determine textual similarity which are still easy to compute but which take into account more of the structure of the document. The ROUGE [2] measure was originally developed as an evaluation metric for automatic summarization systems but can also be used as a similarity measure in it's own right. ROUGE looks at sequences of terms, n-grams, when computing similarity. This wider term context allows for some elements of language and content to be captured without requiring deep semantic or syntactic processing.

3.1.1 Semantic Text Analysis

Semantic analysis of text allows us to take a step beyond information extraction (IE) by associating textual mentions with ontological data (i.e. the semantics) stored externally to the document. This allows us to get a deeper understanding of documents, which in turn allows us to make more informed decisions when summarizing documents or determining the similarity of documents. For example, just because two documents contain the word *apple* does not mean that they are both referring to the fruit and so using the concept referred to rather than the actual words would alter the result of most similarity algorithms. Such analysis also frees us from standard keyword based queries when searching large corpora, allowing questions to be asked which can never be answered by traditional search systems [3].

Whilst these techniques can, and should, be used to improve the performance of document similarity and summarization algorithms, in general they relate to the context surrounding a document by pulling in extra information not explicitly contained within the documents. More details can therefore be found in deliverable D6.1 of WP6 on contextualization. Results of the development carried out within WP6 will be incorporated within

the textual summarization and redundancy approaches developed within WP4 where appropriated, as described in current deliverable D4.1.

3.2 Textual Summarization and Redundancy Removal

Textual summaries are extremely common in online and printed media. Table 1 lists some examples of types of summaries which are commonly encountered. These classes have been described as either critical, indicative or informative, where critical summaries attempt to appraise and evaluate a work in some way, indicative summaries aim to enable a reader to decide whether or not to read the whole document [4, 5] and informative summaries try to capture the content of the original document.

Summary	Purpose
Movie review	Critical
Journal abstract	Indicative
Novel blurb	Indicative
News report	Informative
Football highlights	Informative

Table 1: Example classes of summary

Given the aims of the project it is likely that extractive techniques for generating informative summaries will be the most useful. Researchers have developed methods for summarizing single and multiple documents, either by combining sentences and phrases extracted from the original texts (called *extractive summarization*, e.g., [6, 7]) or by using Natural Language Generation (NLG) to create *abstractive*, interpretative summaries (also called concept-to-text generation) (e.g., [8, 9]). The former type of summary, the extract, is composed entirely from text which can be found in the original document(s). Overall, extractive summarization is arguably easier to explain and implement, however, the resulting summaries reflect strongly the original document(s), which could be problematic on short texts, such as often found in social media or short diary entries etc.

Extractive summaries are generally produced according to the following two steps:

1. Score textual units (sentences, phrases, paragraphs, etc.) according to some representation of the document or document set.

2. Generate summaries by selecting high scoring textual units until some desired compression ratio has been achieved.

The textual unit to be included in a summary could be a word, phrase, sentence or whole paragraph depending on the application. Different methods for scoring textual units have been developed, including the aforementioned word frequencies (TF.IDF), sentence position in the document [10], and centroid-based methods [11].

On the other hand, abstractive summarization algorithms tend to be much more complex. The advantage of the abstractive approaches is that they enable succinct summaries of the content, independent of its original presentation in the source document collection [12], as well as the generation of different (personalized) summaries from the same formal input [8].

In addition to being described by their form (abstractive or extractive) and their purpose (critical, indicative or informative) summaries can also be classified by whether they are derived from single or multiple documents. The two types are considered somewhat separate, because multi-document summarization must address different challenges to single document summarization such as repeated text between documents, order of publishing and inter-document references.

Summaries may be topic-centric (generic), user-focused (i.e. personalized) or query focussed. The former class of summary is meant simply to summarize the content with no bias as to whom can benefit from it. Query focussed summarization, on the other hand, involves building a summary to meet a specified information need; in this sense it is related to the task of question answering, and is especially useful for generating answers to definition style questions [13]. User focussed or personalized summarization must model in some way the information needs and interests of a specific user, arranging a summary containing details which they alone may find salient.

Another approach is to simplify documents to reduce their length, rather than trying to extract the salient information. Within any passage of text it is quite likely that some of the words or phrases will be redundant; that is, removing them has no effect on the meaning of the text. While full syntactic or semantic parsing can be used to identify redundant phrases, shallower approaches are often both easier to implement and understand, and can be performed in real time making them more applicable to interactive use cases. The following is not intended as an exhaustive list of redundant phrases types but aims to illustrate common examples which have previously been shown safe to remove without affecting the meaning of the text [14, 15][1].

- **Gerund clauses** often comment on, rather than add to the content of a sentence and therefore tend not to include essential information. For example, “More than 800 lives were lost when the ferry*, carrying passengers from the Estonian capital Tallinn to Stockholm,* sank within minutes early yesterday morning in the Baltic Sea 40 km south west of the Finnish island of Uto.”.

- **Leading Adverbs:** In some sentences the lead word does not actually contribute to the meaning of the sentence. Certainly given a sentence in isolation the words “*and*” and “*but*” at the beginning of the sentence can be safely removed. Similarly, adverbs when they appear as the first word in a sentence can be safely removed. More importantly if a summary is constructed from a set of independent sentences the presence of these words at the beginning of sentences can often disrupt the flow of information.

[1] In the examples which follow redundant text is written in *italics*.

- **Sentence Initial Expletives** are phrases of the form *it + be-verb* or *there + be-verb*. Such phrases can be useful in expressing emphasis but usually result in longer sentences than is strictly necessary to convey the information. For example the sentence "*It is* the governor *who* signs or vetoes bills." can easily be re-written, without changing the meaning of the sentence, as "The governor signs or vetoes bills.". Further examples include, "*There are* four rules *that* should be observed." which becomes "Four rules should be observed." and "*There was* a big explosion*, which* shook the windows, and people ran into the street." which becomes "A big explosion shook the windows, and people ran into the street.". From these examples it is clear that no information has been lost and the results are shorter sentences which are still well formed and easy to read.

- **Redundant Category Labels:** Certain words imply their general categories and so a sentence does not usually have to contain both the word and it's category label. For example, users will know that pink is a color and that shiny is an appearance so the sentence "During that time period, many car buyers preferred cars that were pink *in color* and shiny *in appearance*." as "During that period, many car buyers preferred cars that were pink and shiny." without altering the meaning of the sentence.

- **Unnecessary Determiners and Modifiers:** Sentences sometimes contain one or more extra words or phrases which seem to determine narrowly or modify the meaning of a noun without actually adding to the meaning of the sentence as a whole. Although these words or phrases can, in the appropriate context, be meaningful they can often be eliminated. For example "Any *particular type of* dessert is fine with me." can easily be re-written as "Any dessert is fine with me." without any alteration in the meaning of the sentence.

- **Circumlocutions** are indirect or roundabout expressions of several words but which could easily be written more succinctly, often as a single word. It is usually possible to replace both *"the reason for"* and *"due to the"* simply with the word *because*. Unfortunately there are no hard and fast rules which state exactly which expressions can be replaced or with what. For instance, the previous two examples could, in most instances, also be replaced by the word *since*.

- **Unnecessary That and Which Clauses:** When a clause is used to convey meaning which could be presented in a phrase or even a single word the sentence length is increased without any increase in the information conveyed to the reader. Often the unnecessary clauses are of the form *that + be-verb* or *which + be-verb* which can be easily simplified. For example "All applicants *that are* interested in the job must..." can be simplified to "All applicants interested in the job must..." without any change in the meaning of the sentence, and similarly "All components *which are* needed for..." can be simplified to "All components needed for.." without any change in the meaning of the sentence.

- Sentences which uses the **noun form of a verb** often contain extra words (often the verb *be*) to allow the text to flow correctly. Changing these nouns back to their verb forms therefore reduces the length of the phrase. For example "The function of this

department is *the* collect*ion of* accounts." can be reduced to "The function of this department is to collect accounts.".

3.3 Planned ForgetIT Approach

Within the ForgetIT project we intend to make use of many of the techniques described within this section to solve the problems of detecting textual similarity and reducing redundancy within the archived documents. The exact techniques used will be subject to the requirements of the use cases being developed within work packages 9 and 10. Initial work will therefore focus on integrating scalable, domain independent approaches to textual similarity and redundancy into the ForgetIT framework as the basis from which to develop use case specific solutions.

4 Analysis of Multimedia Quality, Similarity and Redundancy

This section introduces the state of the art on the analysis of multimedia quality, similarity and redundancy. Starting from the image quality assessment presented in Section 4.1 with Full-Reference (FR), Reduced-Reference (RR) and No-Reference (NR) Image Quality Assessment (IQA) algorithms, Section 4.2 reports the current methodologies for Video Quality Assessment (VQA). Then Section 4.3 deals with the evaluation of the similarity between multimedia contents, while Section 4.4 describes the techniques for clustering the several features extracted from them. Section 4.5 gives an overview of how to manage the redundancy of media items and Section 4.6 describes the approaches that will be adopted in ForgetIT for managing the VQA, IQA, clustering and redundancy.

4.1 Image Quality Assessment

The rapid advances in digital media technology have led to a dramatic growth in the amount of images and videos residing along communication networks, such as the World Wide Web, enterprise networks, etc. During different types of image processing operations, e.g. image acquisition, compression, transmission, storage and retrieval, digital images undergo a wide variety of distortions and degradations. Therefore, the quantification of image visual quality is of major importance for a wide range of research areas (e.g., image processing, computer vision, image acquisition, storage, transmission and display system technologies), as in this way the distorted images can be controlled and possibly restored. For instance, in a personal photo album collection application, distorted images can be automatically identified as of lower importance than non-distorted images of the same scene, and receive lower priority for inclusion in a summary of the photo album.

Image quality assessment (IQA) has been a topic of intense research over the last years and a diverse variety of relevant methods has appeared in the literature. IQA techniques can be divided into two main categories: a) objective and b) subjective. Objective quality assessment techniques quantify the visibility of differences between the original and the distorted image using image features and appropriate mathematical operations to automatically provide an image quality score [16]. That is, human viewers do not intervene in the IQA procedure. Some well-known objective image quality measures are the MSE (Mean Squared Error), the PSNR (Peak Signal to Noise Ratio), the SSIM (Structural Similarity), and the MS-SSIM (Multi-Scale Structural Similarity) [17].

On the other hand, subjective techniques exploit a set of human viewers to rate the quality of each image and utilize particular measures to combine the subjective quality measures [18]. To this end, the most commonly used measures are the MOS (Mean Opinion Score) and the DMOS (Difference Mean Opinion Score) [19, 17], which provide a mean rate for the quality of an image.

Subjective image quality techniques can be expensive in terms of human resources and time requirements, especially for large-scale datasets. Therefore, subjective techniques are usually exploited to build the ground truth for IQA datasets, which are then used for learning and validating objective measures. For the above reason, in the rest of this chapter we concentrate on objective IQA techniques. These methods can be further subdivided to full-reference (FR) IQA, reduced-reference (RR) IQA and no-reference (NR) IQA, depending on the amount of information provided regarding the undistorted image. In the following, we examine in more detail the above three categories.

Full-Reference IQA Algorithms

Full-reference IQA (FR-IQA) techniques aim to predict the visual quality of a distorted image assuming that the reference image is fully available. The visual distortion is estimated using one of the quality measures mentioned above. Until recently, the most widely used measures were the MSE and the PSNR which are based on point-wise differences of pixels values between the reference and the distorted image. However, these measures are unable to exploit the image content information and the Human Visual System (HVS) characteristics of the image [20]. More advanced techniques, such as HVSMSE [21], PSNRHVS and PSNRHVSM [22], were developed in order to correlate the human perception with the traditional quality measures. The above techniques were further extended aiming to represent error signal in a perceptually meaningful way. A comprehensive review of such methods can be found in [23]. Finally, in addition to the HSV characteristics of image content many techniques exploit the image structure. The most popular methods in this category are the Structural Similarity (SSIM) [24], the improved Multi-Scale Structural Similarity (MS-SSIM) [25] and the Perceptual Structural Similarity (PSSIM) [26].

Reduced-Reference IQA Algorithms

Reduced reference IQA (RR-IQA) methods assume that in addition to the distorted images some partial information about the undistorted image is also available. These methods are more suitable in scenarios where original images are not usually provided. These methods typically consist of a feature extraction module at the sender side of the communication channel and a quality analysis module at the receiver side, which exploits the extracted features to assess the degradation imposed in the transmitted image. The side image information can be transmitted with one of the following ways: i) using the same channel where the original image is transmitted, ii) using an ancillary channel [27], and, iii) embedded in the image [28].

No-Reference IQA Algorithms

No-reference (or blind) IQA (NR-IQA) methods rate image quality without using any information concerning the original image. These methods are more suitable for real world scenarios, e.g., in communication network applications and large image collections, where

the undistorted images are unavailable. The majority of NR -IQA algorithms are distortion specific, i.e., image quality is quantified assuming that the distortion type (e.g., compression, blur, etc.) is known. Recently, the research community aims towards the development of completely blind techniques, where the assessment of a variety of distortions can be achieved. To this end, NR-IQA techniques are divided into two groups: a) distortion-unaware NR-IQA methods and b) distortion-aware NR-IQA methods, described in the following.

- Distortion-unaware NR-IQA methods: This class of techniques have witnessed great progress during the last few years. These methods usually utilize features extracted using the Natural Scene Statistics (NSS) model of the image [29]. For instance, BIQI [30], DIVIINE [31] and BRISQUE [32], use NSS and a supervised learning method to parametrize and quantify image distortion. In the two first cases (BIQI, DIVIINE), an image transformation is used (e.g., wavelet transform, DCT, etc.) and the retrieved coefficients are utilized to represent the image. Subsequently, feature vectors formed using the above coefficients and the corresponding ground-truth labels are exploited to train a set of Support Vector Machine-based (SVM-based) classifiers, one for each distortion type (e.g., JPEG, JP2K, white noise, blur, fast fading etc.). Similarly, using the same set of feature vectors, the same number of Support Vector Regression-based (SVR-based) modules are trained for providing a respective quality score. In summary, these methods first classify the image distortion using NSS models and then assess the image quality using techniques that refer to the particular distortion category of the image. On the contrary, the BRISQUE algorithm directly utilizes locally normalized luminance coefficients to quantify the distortion (i.e., without performing any coordinate transformation) and the quality evaluation is performed in a one-stage framework. In general, due to the learning step, supervised learning methods present higher computational complexity. Nevertheless, they correlate well with human subjective quality assessments.

  There is also a class of NSS-based methods that do not require a training dataset of distorted images annotated with human judgements. For instance, in this category belong the BLIINDS [33] and the NIQE [34] algorithms. In the former method, the image is partitioned into equally sized blocks, which are thereafter subjected to a local 2D DCT transformation. Subsequently, for each block, a generalized Gaussian distribution model is estimated and features are extracted by utilizing a set of functions related with the Gaussian model parameters. In the final stage, a Bayesian probabilistic model is used for image quality predictions. Similarly, in the NIQE method, NSS coefficients are computed for the entire image and for each equally sized patch. Finally, the visual quality of a test image is estimated by computing the multivariate Gaussian model (MVG) [35] of a corpus of training undistorted images and the corresponding MVG of the test image and comparing them.

- Distortion-aware NR-IQA methods :
  - Blur / Sharpness: Blur is usually quantified using edge detection algorithms. For instance, in [36] a NR sharpness measure is developed based on the DCT Kurtosis of image blocks. In [37], the blur effect is quantified using the average

extent of the edges, while in [38] an iterative edge refinement is used. In [39], a blur quantification method is developed based on ratio and mean factors of edge blurriness and noise.

Edge-based measures are sensitive not only to the threshold choice at the edge-classification step, but also to the presence of noise. For this reason, several methods utilizing a non edge-based measure have appeared in the relevant literature. For instance, in [40] a normalized Gaussian algorithm for blur estimation is proposed, and, in [41] the image quality is measured by quantifying the difference between levels of blur on the same image. In [42], a noise-immune wavelet-based sharpness measure is proposed, while in [43], sharpness is quantified using the degree of local phase coherence of complex wavelet coefficients.

- Contrast: The measurement of contrast level relies on the computation of well known contrast measures, such as RMS contrast, Michelson contrast, Weber contrast [44], and other. In addition, histogram-based techniques have been developed for the efficient and accurate measurement of contrast. In [45], the amount of contrast distortion is quantified by fitting the histogram of the distorted image to the histogram of a model function. In [46], a contrast enhancement technique is presented based on the computation of the contrast on local edge detections, whereas, in [47], the global contrast is quantified using salient region detection.
- Compression: A variety of distortion types may be introduced by the application of image compression. Blurring, blocking, ringing and fast fading are considered as the most common distortions generated during this process. Recently, researchers have developed NR-IQA algorithms which aim to quantify the distortion induced by the well-known compression algorithms JPEG and JPEG2000 [48]. JPEG NR-IQA methods measure distortions caused by compression using a variety of techniques, including hermite transformation to model blurred edges [49], importance map weighting of spatial blocking scores [50], and computation of block strengths in the Fourier domain [51]. On the other hand, most JPEG2000 IQA proposed algorithms are based on the measurement of the edge spread [52], while others are based on feature computation in the spatial domain [53] or on NSS [54].

At this point, we should note that there are several publicly available IQA databases, e.g., [55] and [56], providing reference images, distorted images subjected in a variety of distortion, as well as the associated human opinion scores (MOS, DMOS).

4.2 Video Quality Assessment

Video Quality Assessment (VQA) is similar to IQA in many aspects such that the methods are classified into FR, RR and NR as well. The FR and RR video quality measures are further classified into traditional point-based measures, Natural Visual Characteristics

oriented measures, and Perceptual (HVS) oriented measures [57]. The traditional video quality measures are signal-to-noise ratio (SNR), peak-signal-to-noise ratio (PSNR), and mean squared error (MSE). They are computationally simple, clear and easy to implement but disregard the viewing conditions and the characteristics of human visual perception.

The Natural Visual Characteristics (NVC) measures are further classified into Natural Visual Statistics and Natural Visual Features based methods [57]. Structural Similarity (SSIM) and its variants Multi-Scale SSIM (MS-SSIM) and Speed-SSIM are well known statistics-based measures. SSIM is first introduced as an IQA measure [58] and then extended to video [59]. It measures the structural distortion frame by frame in a video using the luminance component. MS-SSIM is an extension of the SSIM that provides more flexibility by incorporating the variations of the image resolution and viewing conditions [24] whereas Speed-SSIM is another extension that incorporates statistical models of visual speed perception [60]. Among the feature-based methods, Video Quality Metric (VQM) software tools is a well known collection of methods [61]. They are developed by the Institute for Telecommunication Services (ITS), the research branch of the US National Telecommunications and Information Administration (NTIA). Due to their performance they are also adopted by the American National Standards Institute (ANSI) as standard. The NTIA VQM provides several quality models based on the video sequence under consideration and with several calibration options prior to feature extraction in order to produce highly efficient quality ratings.

HVS-based measures can be classified into frequency and pixel domain. In the frequency domain, in one of the early works, Watson et al. introduced Digital Video Quality (DVQ) model [62]. The measure is based on DCT and incorporates aspects of early visual processing, including light adaptation, luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation. A more recently introduced popular measure is Motion-based Video Integrity Evaluation (MOVIE) [63]. It captures temporal distortions as well as spatial distortions. In the pixel domain, a well known metric is Perceptual Video Quality Metric (PVQM) that is introduced by Hekstra et al. [64]. It uses a linear combination of three distortion indicators; edginess, temporal decorrelation, and error. Another pixel domain measure is Visual Signal-to-Noise Ratio (VSNR) [65]. It is introduced by Chandler et al. for still images but it has shown good performance in assessing video quality when applied on a frame-by-frame basis. The model incorporates visual masking and visual summation concepts to identify the perceptually detectable distortions.

As introduced above, although many useful measures have been proposed for VQA, most of them are very complex and require the original video for estimating the quality. As in the case of ForgetIT scenarios, since many systems and applications can not access the reference video, NR measures are required too. Although human observers can usually assess the quality of a video without using the reference, designing a no reference measure is a difficult task thus there is much less work on NR compared to FR and RR.

As Yamada et al. states, NR models can be categorized into two types: the No-Reference Pixel (NR-P) type, and the No-Reference Bitstream (NR-B) type [66]. NR-P type algo-

rithms use decoded video frames whereas NR-B algorithms use bit stream information. Generally, NR-P methods try to introduce video quality measures based on measurements of three artifacts: blockiness, blurriness and noiseness. In a typical work, Farias and Mitra propose one measure for each of the above artifacts [67]. They evaluate the performance of each measure individually and then obtain a model for overall annoyance based on a combination of the measures. Keimal et al. uses a very similar set of measures; blockiness, bluriness, activity and predictability [68]. The first three features are extracted from individual frames of a video whereas predictability is dependent on the perceived visual quality at transitions between frames. They propose a new measure by modelling these four features and verify it on HD videos. They show that the proposed NR measure outperforms PSNR and performs equally well as the top FR measures.

As mentioned above, NR-B algorithms use encoded bit streams and typically exploit DCT coeefficients [69, 70]. Such methods are more suitable for applications to video transmitted over IP networks. In one such work, Yamada et al. proposes a method based on a hybrid of NR-B and NR-P approaches [66]. The method estimates video quality degradation caused by packet loss. Macroblocks containing errors are accurately detected using bitstream information, and the effectiveness of error concealment for these macroblocks is evaluated using both bitstream and decoded-frame information.

4.3 Multimedia Similarity

The representation of multimedia items as well as the computation of their similarity (or dissimilarity), i.e., the degree of how much they are related, are important issues for efficient multimedia indexing, search and clustering. The similarity (or dissimilarity) between two items is typically found in two steps. First, a feature extraction technique is utilized so that a feature vector representation is used to describe the image or video. Secondly, exploiting the desired feature vectors, a similarity (or dissimilarity) measure is evaluated providing a numerical value of the similarity between the two multimedia items. In the following, a review of the state-of-the-art regarding feature extraction techniques and similarity measures is provided.

4.3.1 Visual Information Descriptors

Feature extraction techniques can be roughly divided into two categories based on the type of the feature descriptor used for describing the image (or video keyframe): a) methods that use global feature descriptors, which describe image characteristics such as color, texture, etc., computed along the entire image, b) approaches that exploit local feature descriptors to describe image attributes in different image regions. In the following subsections we provide a review of the two main categories of visual information descriptors.

4.3.1.1 Global Image and Video Descriptors

Global descriptors capture the general characteristics of the image. The well-known MPEG-7 standard specifies a set of global descriptors categorized according to the image characteristic they capture, e.g., color, texture, shape, etc. A brief overview of MPEG-7 [71] descriptors is given in [72]. In the following the major MPEG-7 descriptors are reviewed.

Regarding color information, the MPEG-7 defines four such descriptors [73]: a) The Dominant Color Descriptor (DCD) [74] describes an image using a small number of dominant values (color percentages, variances, etc.) along with an estimation of the spatial coherency, which represents the overall spatial homogeneity of the dominant colors in the image. b) The Scalable Color Descriptor (SCD) [75] characterizes an image using the image color Histogram in the HSV model. To allow a scalable representation, the Haar transform [76] on the histogram values is performed. c) The Color Structure Descriptor (CSD) [77] captures both the local spatial structure and the global color information. The advantage of this descriptor over the two above is that it can be used for distinguishing between images with the same color information but different structure. d) The Color Layout Descriptor (CLD) [78] is a compact and resolution-invariant descriptor capturing the spatial layout of the dominant colors in the YCbCr color space on an image region or the entire image. The final representation is retrieved using the discrete cosine transform (DCT).

An extension of SCD, known as Group of Frame (GoF) or Group of Pictures (GoP) descriptor, has been presented in [79]. This technique generates a color histogram representing a sequence or a group of images rather than a single image. This allows also to be able to assess the compression algorithm running over the GoPs.

Three texture descriptors are defined in MPEG-7 [80]: a) The Homogeneous Texture Descriptor (HTD) [81] provides a quantitative characterization of the image's texture. It is based on a filter bank approach employing scale and orientation sensitive filters. b) The Edge Histogram Descriptor (EHD) [82] uses five types of edges and 16 local edge histograms to represent image content. c) The Texture Browsing Descriptor (TBD) [83] is based on multiresolution decomposition computed using Gabor wavelets [84]. It consists of two parts: a perceptual browsing component which provides a quantitative characterization of the texture's structuredness, directionality and coarseness and a similarity retrieval component that characterizes the distribution of texture energy in different subbands.

For encoding shape information, MPEG-7 defines region-based descriptors, (e.g., the Zernike moment descriptor (ZMD)), and contour-based descriptors, (e.g., Curvature Scale Space Descriptor (CSSD)) [85]. Moreover, Fourier Descriptors applied in different shape signatures are discussed in [86].

Besides the MPEG-7 descriptors, several other descriptors have been proposed in the literature for capturing color, texture and shape image attributes. For instance, in [87], the Color and Edge Directivity Descriptor (CEDD) incorporates color and texture information

in a single histogram. Similarly, in [88], the Fuzzy Color and Texture Histogram (FCTH) descriptor combines color and texture in one histogram using 4 fuzzy systems. In [89], the above descriptors (CEDD, FCTH) are combined yielding a Joint Composite Descriptor. For texture descriptors, in [90, 91, 92], the Gabor Texture Descriptor (GBR), Wavelet Texture Descriptor and Local Binary Patterns (LBP) respectively, are used for describing images based on texture information. As far as shape descriptors are concerned, popular approaches include the Edge Orientation Histogram (EOH) [93], the Edge Direction Histogram (EDH) [94], and the Histogram of Oriented Gradients (HOG) [95]. Other extensions include the Pyramidal HOG (PHOG) [96] and the Compressed HOG Descriptor (CHOG) [97]. Finally, the authors in [98] propose a totally different global descriptor, called GIST, which exploits several important statistics about a scene. It encodes the amount or strength of vertical and horizontal lines representing the dominant spatial structure of a scene.

4.3.1.2 Local Image and Video descriptors

Local descriptors extract features from regions or points of interest in order to describe the local image structure. They are usually invariant to certain image transformations, such as viewpoint changes, lighting conditions, etc., and in general, any affine transformation of the space such as geometric roto-traslations. Local descriptors are computed in two steps: a) a keypoint detection method is applied to select keypoints of interest, b) a suitable local descriptor technique is applied to provide a feature vector representation of the selected local image patch. In the following, an overview of keypoint detection techniques and local image patch descriptors is provided.

Edge detection algorithms [99] are often used for detecting salient local keypoints, e.g., Sobel operator [100], Prewitt operator [101], Robert's Cross operator [99], Canny edge detector [102], etc. Another interesting method is the Harris-Laplace point detector [103], which uses the Harris corner detector for determining a set of candidate keypoints and for each corner a scale-invariant point is selected making use of the Laplacian operator. Finally, the dense sampling strategy [104] and the random sampling approach [105] are two local region detection methods that recently are getting increasing attention.

The most well-known local descriptor is probably the Scale Invariant Feature Transform (SIFT) descriptor proposed by Lowe in [106]. SIFT is invariant to translation, rotation and scaling transformations and robust to moderate perspective transformations and illumination changes. It is based on local gradient histograms sampled in a square grid around the detected keypoints. SIFT has been extended in several ways as described in the following. In [107, 108], PCA-SIFT and GLOH descriptors were proposed respectively, in order to provide a more compact (of lower dimensionality) representation of SIFT feature vectors. In [109], SIFT + GC augments SIFT vectors with global context components that add curvilinear shape information. In [110], inspired from SIFT, the rotation invariant RIFT descriptor is presented. In [111, 112], KPB-SIFT and CDIKP are proposed, which apply kernel projection techniques to derive a more compact SIFT-based feature vector representation. Finally, in [113], several color-based invariant descriptors (HSV-SIFT,

HUE-SIFT, OpponentSIFT, C-SIFT, rgSIFT, transformed color SIFT and RGB-SIFT) are evaluated in three publicly available datasets.

Another well-known local descriptor is the Speeded up Robust Features descriptor (SURF) proposed in [114]. In comparison to SIFT, this descriptor is faster. Similar to SIFT, several extensions of the original SURF have been proposed. Dense-SURF (DURF), in [115], uses a dense sampling strategy and achieves a much better detection performance than conventional SIFT with lower computational cost. In [116], a novel Colored Local Invariant Descriptor based on SURF was proposed providing photometric invariance. In this work, authors consider the color-based photometric invariants presented in [117], which are derived from the Gaussian opponent color model to enrich the photometric invariance of the extracted feature vectors.

Besides SURF and SIFT, which are the most widely used descriptors, several other local descriptors have been proposed. In [118, 119], a distribution-based descriptor and an affine-invariant descriptor are proposed respectively. Several filter-based descriptors have been proposed in the literature. For instance, in [120, 121] Steerable Filters and Gabor filters are exploited respectively for building local descriptors. Another interesting method, the so called Textons [122], characterize the image textures by quantizing the responses of a linear filter bank. In [123], semantic textons are presented that act directly on image pixels, thus, avoiding expensive computations of filter bank responses. Several other approaches have been proposed including: the multi-scale phase-based local features [124], the multiple support regions [125] and the covariant support regions descriptor [126].

For compact representation of the visual content, usually a 'Bag-of-Words' (BoW) [127] approach is used to represent each image with a BoW vector. First a clustering algorithm (e.g., k-means, fuzzy-c means, etc.) is applied to the derived local descriptor vectors to create a codebook of visual words. Subsequently, using this visual vocabulary for each image a histogram of visual words is created. This is usually the final low-level feature vector representation of the image (the BoW approach is described in detail in Subsection 5.1.1 in the context of concept detection).

4.3.2 Multimedia Metadata

Information residing in administrative metadata of images and videos can provide useful cues for multimedia representation. These metadata may include time and geolocation tags, user tags, creator name, multimedia format, etc. Different protocols and techniques can be used to store multimedia metadata.

MPEG-7 provides a comprehensive and rich metadata standard [128] for the description of multimedia content. This standard defines a way to describe the content of a multimedia item (e.g., audio, video and images) using both subjective and objective descriptors (metadata) using XML. However, due to the extensive model and complexity of MPEG-7, the standard has not been adopted in industry very well [129]. Another used metadata

format is the Dublin Core [130], which is an interoperable online metadata standard focused on networked resources. It consists of 15 elements, such as title, creator, date, descriptions and others, describing the multimedia content. Other metadata standards are: EBUCore, a set of descriptive and technical metadata based on the Dublin Core adapted to media; PBCore, a Metadata and Cataloging Resource for Public Broadcasters and Associated Communities; EXIF [131]; IPTC-IIM [132]; XMP [133] and others. Within the scope of WP4, we will focus mostly on exploiting EXIF metadata, due to their widespread availability. EXIF metadata are produced automatically when the image or the video is captured. Combining them with visual descriptors (Section 4.3.1) we can retrieve a more powerful representation of multimedia items. For instance in [134], time stamps and geo-location information is combined with visual feature vectors for collective multimedia organization. This approach is appealing because timestamps and geo tags provide reliable information and can be processed with low computational cost. In [135], low-level features and metadata statistics, such as exposure time, flash fired etc., are used for indoor-outdoor classification and sunset detection in images. In [136], low-level features, metadata and meta-tags added by social tagging are employed by an agglomerative clustering algorithm. The clustering results are used for representing image search results on the web.

4.3.3 Similarity measures

The degree of relation between multimedia items is assessed using a suitable similarity (or dissimilarity) measure. This measure is applied on the low-level feature representation of the multimedia items and provides a numerical value reflecting the similarity between them. These values can be exploited in different ways, e.g., for classification, construction of efficient tree structures, etc. The choice of the dissimilarity measure depends on several parameters, such as the nature of the data, the application complexity, etc. Below we review several distance measures that have appeared in the multimedia similarity literature.

The most popular dissimilarity measure is the Euclidean distance (L_2) [137], which belongs to the L_p Minkowski family [138]. Other popular measures of this family are the Manhattan (L_1), the Minkowski (L_p) and the Chebyshev (L_∞) distance [139]. The above distances assume that the vector dimensions are independent. However, in several multimedia applications individual vector dimensions may be correlated. In order to overcome this limitation Quadratic Form Distances (QFD) are used [140]. In [141], the Signature Quadratic distance Function (SQFD) is used to extend QFD for computing distances between feature signatures, which represent sets of feature clusters; see [142] for more details. Other popular distance measures for feature signatures are the Hausdorff distance [143], the Perceptually Modified Hausdorff Distance [144], the Earth Mover's Distance [142] and the Weighted Correlation Distance [145].

Taking into account the data distribution, the use of the Euclidean distance (or sum of the square differences) is justified when the data distribution is Gaussian, whereas the

Manhattan distance (or sum of the absolute differences) is suitable for Exponential data distribution. However, real-world data distributions may differ from the above. To this end, in [146] a boosted distance framework is proposed that finds multiple distance measures capturing the underlying data distribution for each element separately. In [147], a new hamming-based distance measure is proposed that exploits the edit distance to compare different bit sequences. Other interesting works are [148] and [149], where 5 and 38 distance metrics were implemented and tested respectively. In the former evaluation the Manhattan distance provided the best results, while in the latter the pattern difference measure and the Meehl index where the best-performing measures.

It is worth to mention the work done in the field of similarity measure adopting the quantum mechanic formalism, introduced by K. Van Rijsbergen [150] few years ago and exploiting a more sophisticated distance measure between multimedia objects: the adoption of bracket formalism and more generalised geometry theory behind the extracted features allows to evaluate the projection of query sets to the overall sample set, leading to the probabilistic measure of having the query represented in our sample. Hence, a probability distribution is substituting the Euclidean distance, enabling overlap of states instead of simple thresholds. The work in [151] and [152] have proposed a tensor representation of features with a linear superimposition for similarity measure the former, and a functional representation in a complex space the latter, with a probability distribution as similarity information. The approaches proposed are still not demonstrating to work reasonably better in respect to the classic and well established methods reported in the current document. Nevertheless some specific areas such as the relevance and pseudo-relevance feedback can benefit of a more probabilistic approach.

4.4 Clustering Multimedia Items

Clustering (or unsupervised learning) refers to the learning problem where the data should be grouped without exploiting any labelling information [153]. The resulting clusters contain data that are more similar to each other in the same group according to a particular similarity (or dissimilarity) measure. Clustering techniques have numerous applications, such as class suggestion in multimedia applications, data compression, density estimation, image segmentation and other.

Clustering methods can be roughly categorized as following: a) Hierarchical Methods: These are multilevel procedures where at each level a different number of clusters is formed. We distinguish two main subcategories: i) Agglomerative: In the first level each sample is considered as a cluster and successively at the next levels clusters are merged in a bottom-up fashion. ii) Divisive: In contrast to agglomerative clustering, here we start with all samples in one cluster and successively each cluster is divided at the next levels in a top-down fashion. b) Graph-based clustering: These algorithms relate data items using a graph representation of them. However, only a single level is used. c) Partitioning Methods: These algorithms create a flat set of clusters, where clusters are unrelated with each other. Popular algorithms of this category are the K-means [154] and K-medoids

[155]. A brief overview of clustering algorithms with respect to the above categories is provided below.

Hierarchical Clustering techniques, firstly proposed by Jonhson in [156], group data points in a tree-shaped structure. A basic Agglomerative Hierarchical clustering algorithm, called Sequential Agglomerative Hierarchical Non-overlapping clustering (SAHN) was introduced in [157]. This iterative algorithm begins by considering all data points as individual clusters and at each iteration the two most similar clusters are merged. A distance measure, such as the Euclidean distance or the Mahalanobis distance, is used as the clustering criterion. The clustering results of this algorithm are usually visualized with a dendrogram. In [158], a hierarchical clustering-based browsing algorithm is used for navigating within an image dataset. In [159], the information bottleneck principle [160] is utilized to design an information-theoretic framework for hierarchical clustering.

In contrast, divisive hierarchical algorithms begin considering all data points into one cluster and in subsequent steps successively split the data into subclusters depending on a particular distance measure. In [161], the selection of the next candidate cluster to be split is discussed. The common Size-priority cluster split approach is first discussed and then four new selection methods, namely, the Average similarity, the Cluster cohesion, the Temporary objective and the Stopping criteria are presented. Similarly in [162], several splitting criteria were proposed in order to discover clusters effectively.

Partitioning methods are the most commonly used algorithms. Partitioning is done by optimizing a certain objective function such as the minimization of the sum of squared distances from cluster centroids. Partitioning methods can be either fuzzy or crisp. In fuzzy algorithms data items can belong to multiple clusters, whereas, in crisp algorithms each data item belongs to a single cluster. K-means is the most widely used algorithm in this category. Several different formulations of this algorithm have been proposed until now, such as the approaches in [154, 163]. The basic steps of the algorithm are: i) Select a set of k centroids. ii) Assign each data point to the closest centroid using a distance measure, typically the Euclidean distance. iii) Recompute centroids. iv) Repeat the above steps until the centres of the clusters converge to a stable solution.

K-medoids [155], is another partitioning algorithm. It works similarly to the K-means algorithm, with the difference that instead of using the average of all points in a partition for computing the respective centroid, the median point of a cluster is used. A comparison of both algorithms is given in [164]. Another interesting algorithm is the one proposed in [165]. In this work, the Fuzzy C-means (FCM) algorithm is combined with the Affinity Propagation (AP) algorithm in order to provide an estimation of the number of clusters C for the FCM.

Graph-based clustering techniques utilize graph theory to represent the dataset and the relations among datapoints as a graph (G) of vertices and edges. For instance in [166], a graph-based image clustering algorithm is proposed, where the vertices correspond to images and the edges correspond to the similarities between images. The partitioning of the graph is usually done using a spectral clustering algorithm that exploits the notion of cut between subgraphs, such as the Ncut algorithm [167]. In order to cluster images

taking into account different image modalities, e.g., low-level features, textual tags, etc., k-partite graphs may be exploited. For instance, in [166], the bipartite Spectral Graph Partitioning is proposed to exploit information contained in both low level features and image surrounding text. Similarly in [168], visual features and image surrounding text are modelled using bipartite graphs. The above graphs are partitioned simultaneously exploiting a Consistent Isoperimetric High-order Co-clustering algorithm.

4.5 Redundancy of Media Items and Media Fragments

A large portion of multimedia data residing in the web, enterprise storage systems or even personal computers is redundant. These media are characterized as duplicate or near-duplicate. The term duplicate refers to multimedia items that are exact copies of the original one, while near-duplicates are altered versions of the original. The common transformations characterizing near-duplicates can be categorized as follows [169, 170]:

- Minor scene changes: slightly different background, absence or presence of objects, etc.
- Camera parameter changes: change of angle, scaling, panning etc.
- Photometric changes: lighting conditions, etc.
- Digitization changes: change in color, contrast, saturation, cropping, resolution etc.

Multimedia tools that can automatically (or semi-automatically) discard redundant images/videos, while at the same time preserving the most significant ones for the specific application, can save huge amounts of storage space and increase user satisfaction. In the following we briefly review indicative works in near-duplicate multimedia identification. In [170], a part based representation of visual scenes and Attribute Relational Graphs are used for detecting near-duplicate images. The proposed approach is evaluated using a subset of the TRECVID 2003 corpus. In [171], PCA-SIFT local features and locality-sensitive hashing (LSH) are exploited for image representation and indexing respectively. This method was applied for near-duplicate image detection using the MM270K [172] database and a fine arts image collection. Finally in [173], the above work was extended using Local-Difference-Patterns (LDP) for image representation and a more efficient LSH strategy. Using this method very good results were obtained for near-duplicate image and video detection on two publicly available datasets.

Exact duplicates can be detected with even simpler methods than near-duplicate ones, but for sure the methods designed for the latter and outlined above can be used for detecting exact duplicates as well.

4.6 Planned ForgetIT Approach

Visual Quality Assessment

The goal within ForgetIT is to develop visual quality assessment techniques that can be applied to natural image collections. As in this scenario reference images cannot be available, we will concentrate on NR-IQA methods. Specifically, the distortion unaware methods BIQI, BRISQUE, DIVIINE, BLINDS and NIQE will be evaluated in terms of prediction accuracy and computational cost and the best of them will be selected for further investigation. The overall evaluation will be performed in publicly available databases, such as the LIVE DB [55], as well as on natural image datasets created within the ForgetIT project.

Preliminary evaluation results have shown that BIQI and BRISQUE seem to outperform the other methods described above. The main limitation of these methods is that they only examine blur and noise distortions in natural images. To this end, we plan to extend the above methods so that contrast distortions are also accounted during IQA. This may be achieved for instance using RMS and Michelson contrast measures [44] or histogram based methods [174].

Moreover, the use of supervised learning algorithms for enhancing the performance of the above IQA methods is another interesting research direction that we plan to investigate. Ultimately, we also wish to investigate new measures for additionally providing aesthetic assessment of natural images, a topic that has received relatively little attention in the relevant literature (e.g., see [175]).

Image Similarity Assessment and Similarity Measures

Image similarity assessment is necessary for near-duplicate image identification, redundancy assessment, dataset partitioning, etc. Towards the above directions, in order to facilitate subsequent steps in the ForgetIT project, suitable features and appropriate similarity measures will be investigated.

For coarse similarity assessment global features will be extracted, such as MPEG-7 color descriptors. For instance, the CSD descriptor may be suitable for distinguishing images in summer holiday photo collection which usually depict several different objects in front of a mostly blue background (sea, sky, etc.). For finer assessment, local descriptors will be exploited. To this end, the popular SIFT [106] and its color variations [113] will be utilized. Moreover, we also wish to investigate the applicability of the SURF descriptor [114] and work on possible color extensions of it. Finally, use of image metadata information will also be considered. In particular, the above visual features will be combined with spatio-temporal metadata of EXIF standard such as timestamps and geo-locations.

For the actual similarity assessment, we sill start with using the Euclidean distance and subsequently, if required, additional measures which may be more efficient for large-scale

processing may also be considered.

Clustering and Redundancy Algorithms

The feature extraction procedures and similarity measures described in the previous section will be utilized for identifying near-duplicate images in the ForgetIT dataset. The next step will include the application of a clustering algorithm for data partitioning. Subsequently, the most representative images within each cluster will be identified, and this information will be used within the ForgetIT project for guiding subsequent preservation choices.

For data partitioning the popular K-means [154] and the recently proposed K-medoid [155] will be investigated, The drawback of these algorithms is that the number of clusters C should be provided in advance. To this end, hierarchical clustering methods may also be applied to allow for an automatic estimation of C.

5 Semantic Multimedia Analysis for Condensation

5.1 Concept Detection

The detection of high level concepts in image and video signals is an important and challenging multimedia analysis task. Concept detection can significantly aid the consolidation and condensation actions by providing a higher level semantic description of the multimedia items. However, the well known 'semantic gap' causes restrictions to the mapping of low-level features to high-level concepts. To date, many techniques tried to deal with this challenge and most of them are generally based on the following procedure:

- Content representation: spatial and/or temporal sampling is applied in order to reduce the amount of the visual image or video information that will be processed. To this end, images may be represented in lower resolution while videos are typically represented using keyframe sequences.
- Low-level feature extraction: as described in Section 4.3.1.2 a suitable feature extraction technique is applied to the image or video keyframes of the dataset, extracting feature vectors which describe the media item.
- Learning: an annotated dataset is then used for creating the concept detectors. This is done by first deriving low-level feature vectors for this dataset, and then using these feature vectors to train a machine learning technique, e.g., SVMs, nearest neighbour classifiers, etc.
- Concept-based content description: this step consists of the application of the trained concept detectors to images or video keyframes for providing a semantic description of the multimedia content.

We should note that several pre- or post- processing steps can be added to the above description. For instance, often multiple concept detection results are further combined using a late fusion step in order to increase the detection performance. However, the aforementioned low-level feature extraction procedure and the machine learning technique are the most important elements in concept detection. For this reason, we review related techniques in more detail in the next subsections.

5.1.1 Low-level feature extraction

Content representation may be achieved either with global or local descriptors. As described in Section 4.3.1.1, global descriptors extract the general characteristics of the image such as color, texture and shape. For example, in [176], color and texture features are used to construct descriptors and train SVMs for high level concept detection. However, the majority of the state-of-the-art techniques make use of local descriptors for content representation. The first step of these techniques consist of the application of

an interest point detector in an image or video keyframe in order to select salient points for extracting local descriptors. Such techniques include the Harris-Laplace point detector [177], dense sampling [178], random sampling [179], the Maximally Stable Extremal Regions [180], the Maximally Stable Color Regions [181] and combination of the above techniques. In the next step, local descriptors are applied for representing the extracted interest points with low-level feature vectors. The SIFT descriptor [106] and variations of it [113, 111, 112] are among the most popular descriptors in the literature. For instance in the competition of ImageCLEF 2009 [182] and ImageCLEF 2010 [183] SIFT and several of its variations, such as HSV-SIFT, HUE-SIFT, OpponentSIFT, rgSIFT, C-SIFT, and RGB-SIFT [113] were extensively applied for the task of concept detection.

Another popular descriptor is SURF [114], which gives improved computational time and comparable performance to SIFT, as shown for instance in [184]. In [115], experimental results using Pascal VOC 2007 dataset show that the combination of dense sampling strategy with the SURF descriptor (DURF) outperforms a SIFT-based algorithm for the task of concept detection. In [108], the GLOH descriptor is introduced that provide a more compact feature representation. In [185], the DAISY descriptor is presented that is faster from SIFT and GLOH. Moreover, it provides improved accuracy and comparable computational time performance with DURF. Finally, it has been shown that the combination of several global and local descriptors provides improved accuracy. For instance in [186], global and local visual features as well as audio features are combined using ensemble fusion. Similarly in [187], the authors report results of their participation in ImageCLEF 2008 visual concept detection task. In this work concept detectors are built using the k-nearest neighbour (KNN) classifier and combinations of local and global features.

Visual word assignment is the last part of the feature extraction procedure. In this step, the low-level features are usually transformed to Bag-of-Words (BoW) vectors. Firstly a visual vocabulary or codebook is constructed by grouping similar keypoints into a large number of clusters and treating each cluster as a visual word. The most common method for the construction of the codebook is the well known K-means clustering algorithm. Subsequently, the previously calculated local descriptors are assigned to the codebook so that each descriptor is mapped to a visual word. The assignment can be either soft or hard. Hard-assignment is typically implemented by finding the Nearest cluster centroid, while soft-assignment allows assigning each descriptor to multiple visual words as explained in [188]. The latter usually provides improved detection accuracy. Moreover, for better performance in terms of computational time the authors in [189, 190] proposed the use of tree-based assignment algorithms. Another interesting tree-based structure method for visual assignment is the semantic texton forests proposed in [123]. The advantage of this technique is the increased efficiency and accuracy over previous tree-based algorithms. Finally the 'Spatial Pyramid' strategy [191] divides images to several rectangular regions and creates a BoW vector for each region separately.

Several other extensions of the above techniques have been proposed in the literature. In [192], a region-based approach is introduced, where a BoW model is constructed for the most common 'region types'. Finally, the authors in [193], construct multiple resolution images and extract local features from all resolution images with dense regions. The

derived local features are then used for the creation of a BoW model with the K-means algorithm.

5.1.2 Machine Learning for Concept Detection

The other major step in concept detection is the use of a machine learning algorithm for implementing the concept detector. These approaches exploit an annotated dataset and feature vectors extracted using a particular feature extraction procedure to learn the desired concepts. Several different learning approaches have been used until now. Among them, Support Vector Machines (SVMs) [194] and its variations are probably the most popular methods in this domain. A brief review of such approaches is presented in the following.

In [195], three types of low-level features, namely, color moments (GCM), local binary patterns (LBP) and edge orientation histogram (EOH), are used as inputs for learning concepts on various TRECVID datasets. In [196], low-level features extracted from videos of the TRECVID 2007 corpus are used to construct a region thesaurus. This thesaurus is then exploited to provide a model vector representation for each video and train a set of SVM-based concept detectors. In [197], SVMs with Radial Basis Function (RBF) kernel and rgSIFT features are used for learning the concepts of the ImageCLEF 2010 visual concept detection and annotation task. Other popular kernels are the Chi-square and the Histogram Intersection kernel. The former has been evaluated in [198], providing superior performance in multicategory classification tasks. Moreover, in [199] an extended version of the latter for the classification of binary strings, such as color histograms, showed very good performance in the Corel database. Recently, methods using ensembles of linear SVMs (LSVMs) are getting increasing attention because they can offer much faster training and testing times especially for large-scale datasets. For instance in [200], feature vectors derived using low-degree mappings are used to train LSVMs. This approach exhibited competitive performance to KSVM in a variety of problems. Similarly, in [200, 201], LSVMs are used to exploit the subclass structure of the data. In particular, mixLSVMs [202] train one LSVM to separate positive from negative observations in subregions of the feature space and incorporate the local LSVMs using a mixture of experts framework. On the other hand, LSSVMs train LSVMs [201] to separate observations belonging to different subclasses, and fuse LSVM decision scores using an ECOC framework [203] for classifying test videos.

Besides SVM-based techniques, several other machine learning approaches have been used for concept detection. For instance, in [204], a multigraph-based learning algorithm, which effectively integrates multiple graphs into the same regularization framework, is used for learning 39 LSCOM-Lite concepts. In [205], multiple-instance learning is used to associate local image regions with keywords and a Bayesian formulation is utilized for the overall image annotation with a set of concepts. In [206], a fuzzy spatial relation ontology is utilized for knowledge-based recognition of brain structure concepts in medical image applications. Finally, in [207], a correlative multi-labelling approach is reported, achieving

very good results in the TRECVID 2005 concept detection task.

5.2 Face Detection and Clustering

Face detection has been extensively studied during the past years. It usually serves as a stepping stone for most facial analysis algorithms in the areas of face tracking, clustering, recognition and others. The objective of face detection is, given an image, to detect the presence of a face and locate the facial region of interest (ROI). This task is challenging as faces appear in different scales, poses, expression, with varying occlusions, etc. Moreover, with the advances in photo camera technology, facial images are produced in large-scale, residing in personal multimedia collections, social media servers, etc. To this end, it is expected that face detection has still to play a fundamental role in the organization and condensation of these vast multimedia collections.

A recent overview of face detection is provided in [208]. In the following, we briefly review recent approaches in this field. One of the most successfully face detectors is the one introduced in the seminal work of Viola and Jones [209]. This algorithm detects faces in real-time following three main steps: a) computation of an integral image [210], which is used for rapid extraction of Haar-like features, b) training of an AdaBoost classifier [211], which is used to select a small number of potentially useful features, c) exploitation of an attentional cascade structure [211], which dramatically increases the speed of the detector by focusing attention on promising image regions. In [212], a color-based segmentation algorithm that combines HSV and RGB color space models is proposed yielding good detection accuracy. A relevant approach, utilizing skin face color was proposed in [213]. In this work a fuzzy classifier exploits RGB and HSV models to detect potential face-depicting regions, and geometrical facial properties are used to refine the detections. However, skin color-based techniques are sensitive to illumination changes and require color images, limiting the applicability of those algorithms. In order to overcome the above limitations different feature types may be exploited, e.g., extending Viola and Jones algorithm. In [214], rotated Haar-like features are proposed that can enrich conventional Viola and Jones features and in [215], the Viola and Jones classifier is extended using additional sets of weak classifiers in the AdaBoost algorithm. In [216], locally assembled binary (LAB) features are introduced, extending the idea of combining Haar-features and locally binary patterns (LBP) for face detection [217]. Experimental results showed that these features are superior to state-of-the-art methods both in terms of detection accuracy and speed.

As explained in the beginning of the section, face detection is usually utilized as a preprocessing step for further face analysis tasks. In this project we are interested in face clustering for multimedia organization. A brief overview of the state-of-the-art methods in this domain is provided in the following. In [218], a dissimilarity matrix is constructed using SIFT-based facial features. Facial images are then clustered using a hierarchical average linkage clustering algorithm. In [219], facial images are represented in the HSV color space and clustered using a spectral graph partitioning algorithm. In [220], a graph-

based clustering algorithm is proposed for partitioning facial image sequences. In [221], the Gabor wavelet transform and the principal component analysis are exploited for facial image representation. This representation is then used to examine the effect of facial poses in face analysis systems. Finally in [222], to alleviate the effect of different poses, clustering is done in two stages. Initially, eye detection results are exploited for pose clustering. Within each pose cluster, K-means is used to group facial images of different individuals.

5.3 Event Detection

Event detection is currently considered as a major step for narrowing the semantic gap between human and machine understanding of the real world [223]. The utilization of event detection results can significantly aid towards more effective organization of multimedia items in relevant applications.

Events consist of dynamic and static objects with or without interactions among them, occur at specific time and place, and their perception depends on the particular observer [224]. For this reason, event detection is a much more challenging task than tasks dealing with the classification of specific objects or actions [225]. A recent review of the current state-of-the-art in this field is provided in [226]. In the following we review recent works in this topic.

In [227], high level visual and temporal features such as material visual types [228] (sky, grass, etc.), event duration, etc., are exploited by a Bayesian belief network for event classification. In [229], simple contextual cues (timestamps, GPS coordinates, color information) are exploited along with a multi-modal clustering algorithm for event mining in personal photo collections. In [230], static and dynamic visual features (SIFT, motion boundary histogram (MBH)), along with audio MFFCs are used for video representation. For each modality an SVM is trained, and the weighted average at the score level is used for detecting the events of the TRECVID MED 2011 task. For the same task, in [231], a variety of features (SIFT, STIP-HOG, STIP-HOF, MFFC, etc.), are used to construct a Gaussian mixture model (GMM) supervector for each feature and video. These vectors are exploited by SVM classifiers for learning the MED 2011 events. In [232], model vector sequences and disriminant analysis are used for video representation and the nearest neighbor classifier is combined with the median Hausdorff distance for detecting the MED 2010 events. Similarly in [233], model vector sequences are used to represent videos, and subclass SVMs are combined using an error-correcting output framework (ECOC) for detecting the MED 2012 events.

Event modelling is an important research domain closely related with event detection. Event models can significantly aid event classification and at the same time organize and preserve the detection results and related event information in a human comprehensible format. For completeness, in the following we review recent advances in this domain. In [234], the Video Event Representation Language (VERL) and the Video Event Markup Language (VEML) are presented for describing events in videos. In [235], the E-event

model following a number of fundamental event properties, analysed in [236], is proposed. In [237], the E*-event model is proposed, extending E [235], utilizing the pattern oriented framework of ABC and DOLCE ontologies along with a graph-based design. In [223, 238], the Joint content-event model is proposed taking into account the event requirements presented in [236], and at the same time offering a referencing mechanism for linking event elements with video content. Finally, several annotation tools exploiting proprietary models for event-based video descriptions have been proposed such as AVISA [239], ELAN [240], and other.

5.4 Planned ForgetIT Approach

Concept Detection

In our most recent work on concept detection, a set of 25 different feature extraction procedures are used to detect the concepts on an image or video shot as explained in the following [233]. A 1x3 spatial pyramid decomposition scheme is used, i.e., the entire image is the pyramid cell at the first level, and three horizontal image bars of equal size are the pyramid cells at the second level [191, 241]. In the case of videos, a shot segmentation algorithm is first exploited, e.g., [242], or keyframes are extracted at fixed intervals. A shot is represented with one or more extracted keyframes or shot tomographs [4]. An interest point detector is then applied to the extracted keyframe. In particular, we exploit dense sampling or the Harris-Laplace corner detector [243]. The statistical properties of the extracted keypoints are captured using the SIFT descriptor and two of its color variations, i.e., RGB-SIFT and opponentSIFT [241]. For each pyramid level a visual word codebook of 1000 words is created using k-means and the extracted low-level descriptors. The descriptors are then represented with Bag-of-Words (BoW) feature vectors in R^{4000} employing both hard and soft assignment according to [188]. Therefore, in total 24 feature extraction procedures are utilized derived from every combination of representation type (keyframe, tomograph), sampling strategy (dense sampling, Harris-Laplace detector), descriptor type (SIFT, RGB-SIFT, opponentSIFT) and assignment technique (hard, soft). We also utilize a global visual descriptors (HSV histograms) as 25th feature. For learning the different concepts an appropriate annotated dataset is used and a bag of linear SVM (LSVM) strategy is applied for each feature extraction procedure [244]. In particular, for each concept and feature extraction procedure the k-means algorithm is applied to partition the dataset into 5 subclasses, and for each subset a LSVMs is trained. Thus, in total 125 (25 x 5) LSVM classifiers are trained for each concept. A late fusion strategy is applied to combine the scores of the subclass LSVM referring to the same concept.

A major issue in online classification of large-scale datasets with a large bag of classifiers is the high computational cost that may be required for scaling the test observations. To this end, we will investigate a new scaling strategy, where during training we will find a few representative ranges (instead of 125) and during testing we will restrict ourselves to using only these representative ranges, rather than any possible range. Additionally, we will investigate the possibility of creating color variants of SURF-based descriptors

[114]. These descriptors could be used to replace the SIFT-based descriptors described above. The main advantage of SURF descriptors over the SIFT ones is the computational efficiency. To this end, a further significant speed-up of the overall feature extraction procedure may be achieved. Achieving such improvements in the computational cost of concept detection will then give us room for experimenting with adding further descriptors and learners to our current approach, which is at present infeasible due to computational limitations.

Face Detection and Face Clustering techniques

For Face detection techniques we plan to use cascade classifiers and more precisely an implementation of them in the openCV C++ library. Using a cascade classifier, a training step and then a detection step are required. For the training step we will use OpenCV functions such as $opencv\_haartraining$ and $opencv\_traincascade$. The advantage of the second one over the first is that supports both Haar features proposed by Viola and Jones [245] and LBP (Local Binary Pattern) features proposed in [246]. LBP features are several times faster than Haar and their quality depends on the chosen training dataset and the training parameters. By applying the $opencv\_traincascade$ function a trained cascade classifier is saved in an $.xml$ file. After that, for the detection step, the Cascade Classifier Class implementation will be used exploiting the aforementioned $.xml$ file. Moreover, already trained OpenCV cascade classifiers may also been used, such as $haarcascade\_frontalface\_alt.xml$ (trained by Haar features for face detection), $haarcascade\_eye\_tree\_eyeglasses.xml$ (detects human eyes in image or video) and $lbpcascade\_frontalface.xml$ (which exploits LBP features).

Event Detection techniques

For video representation both the visual and audio modalities will be exploited. From the visual modality two different feature representations will be derived as explained in the following.

a) Low-level visual features: A similar strategy to the one applied for training the concept detectors will also be followed here. According to our plan, firstly, each video is decoded and one frame every 6 seconds is selected to represent a video with a sequence of keyframes. A dense sampling strategy is combined with a 1x3 spatial pyramid approach to extract salient image points at each pyramid level [191, 241]. Each selected point is then represented with a 384-dimensional feature vector using the opponentSIFT color descriptor [241]. The derived feature vectors are then exploited to learn 1000 visual words for each pyramid level using the Bag-of-Words (BoW) method. Subsequently, a video frame is described with a feature vector in R^{4000} using the visual codebook and the soft assignment technique. A low-level feature representation of the video is then derived by averaging the feature vectors referring to the same video.

b) Visual model vectors: A set of SVM-based concept detectors referring to the SIN

TRECVID task concepts [233, 247] will be used to derive an intermediate-level feature vector representation of the video. In particular, a model vector [248, 249] is derived for each video keyframe by concatenating the responses of the concept detectors.

c) Audio features: We will describe audio content in video using linear frequency cepstral coefficients (LFCC) and possibly also modulation spectrogram (MSG) features [250]. These features are complementary in the sense that LFCCs capture the short-time audio characteristics while MSGs describe long-term audio attributes.

An appropriate annotated video collection will be used to learn the required events. In particular, for each event and each feature type described above we will learn an event detector. Then, an appropriate fusion technique will be applied to provide an overall event decision. As event detectors we have decided to investigate DA techniques [251, 252] and their combination with SVM classifiers [194]. The optimization criterion of DA seeks for a reduced dimensionality subspace where noise features or features that are irrelevant to the classification problem at hand are effectively discarded. The kernel trick has been used to extend the conventional LDA to kernel DA (KDA) for non-linearly separable data classes, providing a more effective lower-dimensionality representation. DA techniques provide a data representation that aids classification; however, they do not directly provide a classification function. Instead, in the DA subspace an appropriate classifier is needed (usually the nearest neighbor technique is preferred) for the classification of unlabeled observations. The use of SVMs for classification in the DA subspace is rather an unexplored direction. SVMs have shown excellent performance in various real-world problems outperforming other state-of-the-art methods. To this end, we will explore the combination of KDA techniques with linear SVMs (LSVMs). In particular, conventional and subclass-based KDA techniques combined with LSVMs will be investigated. Finally, different fusion techniques will be investigated for combining event classifiers along different features, such as the geometric mean of detection scores or the use of SVMs for late fusion. Finally, event detection using a distributed systems infrastructure is another possible research direction that may be investigated.

6 Information Condensation and Consolidation

6.1 Textual Content Condensation and Presentation

Whilst single document summarizarion and redundancy removal techniques (see Section 3) are useful in reducing the size of a text collection they ignore the fact that often the same information is repeated across documents. Repeated information may be the main focus of multiple documents (i.e. two news articles that both report the same story) or may be ancillary information not related to the main content (e.g., job titles often accompany the first mention of each person within a news article and hence will be repeated across a collection). Multi-document summarization, and related techniques, exploit this cross-document redundancy to further reduce the size of a given text collection.

There are two different types of corpora for which multi-document summarization can be a useful technique [253]:

- a large corpora containing a mixture of dis-similar documents for which the user wishes to obtain a broad overview
- a corpora of topically related documents for which the user would like a focused summary

Whilst these may seem like very different scenarios a common approach would be to generate document clusters from the large corpora resulting in a number of smaller topic based corpora which can then be summarised. The remainder of this section focuses on the second scenario as being more related to the ForgetIT project (it is envisaged that summarization will be used to reduce the size of a single archived arcifact, i.e. a collection of documents, rather than the summarization of the entire archive). The main approaches to summarizing a collection of topically related documents are:

- **Common Sections:** constructs the summary from the common sections (i.e. the intersection) of the documents in the corpora.
- **Common and Unique Sections:** as above but augmented with a number of unique relevant sections.
- **Centroid Document:** constructs the summary by applying single document summarization to the centroid document.
- **Centroid Document plus Outliers:** as above but augmented with a number of relevant sections extracted from outlying documents.
- **Latest Document plus Outliers**: as above but uses the most recent document rather than the centroid.

6.1.1 Semantic Redundancy and Diversity Analysis

Whilst it may be the case that a multi-document summary of a corpora is required, often a more focused entity/event centric summary might be more appropriate. In such instances a definition based summary can be produced.

The aim of a definition based summary is to extract nuggets of information for a given entity or event from a corpus to assemble a summary. For example, the definition of a person should approximate a biography and include, for example, their fullname, where and when they were born, what they are famous for etc. A number of systems for generating definition style summaries have been reported in the literature [254, 255, 256, 257] due to the inclusion of a definition question answering track at TREC [13].

Most definition style summarization systems work in a similar fashion. Firstly relevant sentences, those containing the entity of interest (or a co-referenced mention) are selected from the corpus. Sentences are then clustered, usually based upon a word overlap based measure (often measures such as ROUGE, described in [2] are used), and finally the summary is constructed by taking one sentence from each cluster in turn until either one sentence from each cluster has been added or until a pre-set length limit has been reached. While the main aim of such approaches is to reduce redundancy they also ensure that the generated summaries are diverse and cover information from across the corpus.

Note that such an approach assumes that only the main entity of interest is known. If work in deliverable D6.1 on contextualization can be incorporated then other entities of importance could be determined and included during summarization. Conversely it is unclear how important such summarization may be in light of contextualization; if the defining characteristics of an entity are available in an ontology and linked to the textual mention would the user require a definition style summary? Of course a definitional style summary will be very useful in those situations where contextualization has been unable to identify an entity (either because the entity is not known or because of ambiguity etc.).

6.2 Image and Video Condensation

The volume of personal photos and videos are exponentially increasing due to high quality mobile camera phones and affordable digital cameras. Whether they are stored in local drives or in cloud, to manage these personal collections, to be able to quickly skim through and eventually understand the most representative ones a summarization system is needed. Jain et al. [258] states that mobile phones are shifting how people shoot and preserve photos such that the *plan-shoot-process-share-organize-reflect* behaviour is now being replaced with *shoot-share-forget*. This statement supports the need for such a system in ForgetIT's *preserve-or-forget* framework.

There are various works in literature for such a task. Li et al. [259] uses only image content and time. The framework consists of a two-stage partitioning. Firstly, the photos

are partitioned in time; secondly, this is followed by content key photo selection based on a very basic feature: color histogram. There are also works that exploit the textual data such as tags that are associated with the photos [260, 261]. To achieve summarization using both visual and textual information, Xu et al. [262] uses GIST scene descriptor to represent visual content and proposes a new approach called *Homogeneous and Heterogeneous Message Propagation (H2MP)* that extends affinity propagation, which is an exemplar based clustering algorithm.

Clustering algorithms that are mentioned in Section 4.4 play a key role in most of the photo album/image collection summarization approaches. In one of the early work Krishnamachari [158] uses hierarchical clustering along with histogram based visual features. Hierarchical clustering is also used to group photos temporally using timestamps [263]. Self-organizing map (SOM) is another clustering algorithm that is exploited for image collection summarization. Deng [264] uses SOMs trained on low-level content-based image retrieval features along with Kullback-Leibler dissimilarity measure.

Low-level features are not the only visual cues that are exploited in image collection summarization algorithms. Cascia and Morana [265] uses faces along with time and low-level visual features (RGB histogram and Gabor filter bank) that represent background in images. Mean-shift clustering algorithm is used along with these extracted cues.

Nowadays, photos are no more just intensity values since cameras capture a considerable amount of metadata associated with the taken photo [258]. That being the case, more modern and novel methods try to exploit those available contextual data from the EXIF metadata associated with the photo and/or mobile sensory data such as GPS. Jain et al. [258] calls all the extractable attributes as *Extractable Mobile Photo Tags*. Such tags are photographer name, people in the photo, location, event, environment, objects, scene concepts and time. It should be noted that some of these tags are inferred from the metadata. Another work that uses EXIF metadata along with the visual content is by Jang et al. [266]. The work also focuses on multiple camera usage and utilize this knowledge obtained from EXIF for better clustering. Both temporal and spatial clustering methods are utilized. In his work, Sinha [267, 268] also exploits both visual content and context data. He states that effective subset summary should satisfy three properties:

- Quality
- Diversity
- Coverage

To satisfy these properties, they define metrics for each property and use optimization algorithms.

An additional cue that can be used in addition to the above mentioned cues is the ”user statistics”. In a very recent work, Guldogan et al. [269] proposes an approach for selecting representative images from photo albums based on user’s preferences. Counting the number of clicks and duration of viewing for each image are used for understanding ”interesting” images for each user, and their personal perception and interest on the

album.

6.3 Planned ForgetIT Approach

In this section we present our first thoughts for combining the multimedia analysis techniques presented in the preceding parts of this document, in order to fulfil the goals of ForgetIT in terms of content summarization. Aspects such as quality, diversity and coverage of multimedia data are considered during this process.

Different multimedia information types are used, such as textual (Section 3.3), image/video quality (Section 4.6) and visual information (Section 5.4). In more detail the cues that can be utilized for achieving our goal are:

- low-level visual features
- higher-level visual information (detected concepts, events)
- visual quality assessment results
- faces
- textual similarity assessment results
- any available user statistics (e.g., # of clicks, duration of view)
- EXIF Metadata (e.g., timestamps, location (GPS), camera parameters)

Within the ForgetIT project scope, it is planned to exploit most of the above cues depending on their availability as well as investigate combinations of them. The overall planned process flow is given in Figure 1.

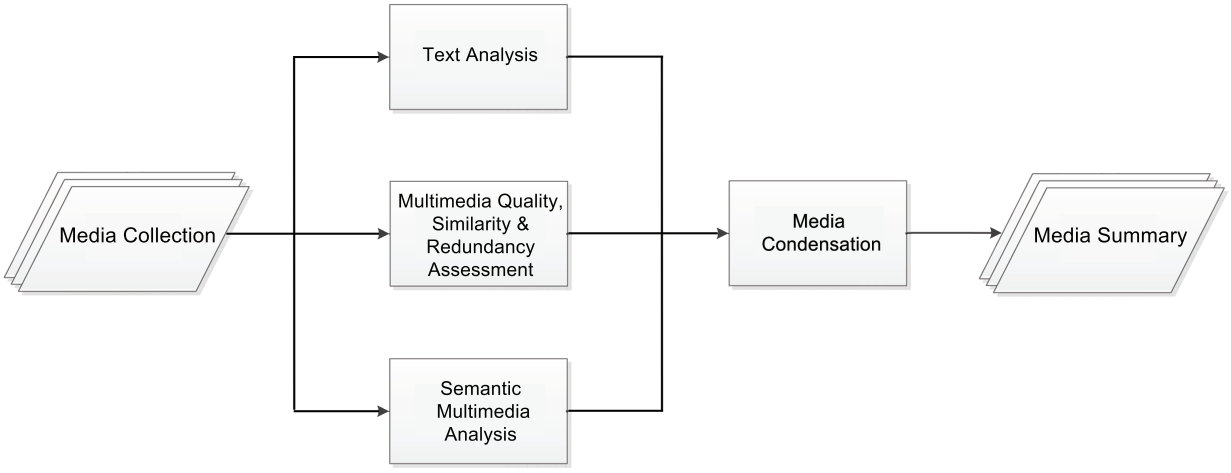

Figure 1: The planned ForgetIT media collection summarization process flow

Using several of the techniques described in Section 3, textual image information and relevant similarity results are used for reducing dataset redundancy. Similarly, visual low-level features are exploited for image quality assessment and near-duplicate image identification, as described in Section 4. The above results are also used in a first stage reducing the size of the collection. In parallel to the above procedures higher level semantics will be extracted for each image using a set of visual concept detectors as well as suitable event and face detection algorithms. The higher level information will be further enriched using EXIF geo-locations and timestamps from image headers. Using appropriate partitioning algorithms and according to our preliminary evaluation results, redundancy reduction will be applied in different stages of the information flow. That is, the clustering algorithm may be applied in a particular feature modality or in an image feature representation resulting in the fusion of several different feature types.

Concerning the type of the clustering algorithm (Section 4.4), the most suitable algorithm will be selected at each stage according to the evaluation results. However, a hierarchical one seems the most suitable for the last stage of the summarization process, where results of the different feature modalities are exploited, as shown in Figure 1.

7 Conclusion

In this document we reported on the state-of-the-art in the research areas related to text and multimedia analysis for condensation, and based on this review we sketched the different techniques that we will start to examine and evaluate in ForgetIT. It should be noted here that the presented state of the art review is not and could not possibly be exhaustive: text and multimedia analysis are multi-faceted topics that are of great interest to a large and vibrant research community, thus have already resulted in a large number of publications. We tried to include in the review presented in this document the most important and recent advances in these areas.

With respect to the planned ForgetIT approach, we should stress that the use of text and multimedia analysis techniques in preservation applications is a rather unexplored domain. Further to this, based on our review of the literature and our experience, it has become clear that to a greater or a lesser extent, depending on the category of techniques, the exact approaches that we will follow in ForgetIT depend on the specifics of the use cases that are being defined in the project, and on the datasets that are being collected. Therefore, the planned ForgetIT approaches outlined in this document should be treated as just our first set of ideas, which in many cases we are already working on implementing and testing, so that we can get a more precise assessment of the further future steps that we need to take; it is neither an exhaustive plan, nor a plan carved in stone. The technical details and results of our first set of ForgetIT analysis approaches will be presented in D4.2, due in month 12 of the project, and these will of course also give rise to further, more concrete plans on the technical developments within WP4 for the next two years.

References

[1] Salton Gerald, Singhal Amit, Mitra Mandar, and Buckley Chris. Automatic text structuring and summarization. *Information Processing & Management*, 33(2):193–207, 1997.

[2] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics, 2003.

[3] Mark A. Greenwood, Valentin Tablan, and Diana Maynard. GATE Mímir: Answering Questions Google Can't. In *Proceedings of the 10th International Semantic Web Conference (ISWC2011)*, October 2011.

[4] Inderjeet Mani. *Automatic Summarization*. John Benjamins Publishing Company, 2001.

[5] Ani Nenkova and Kathleen McKeown. Automatic summarization. *Foundations and Trends in Information Retrieval*, 5(2-3):103–233, 2011.

[6] H. Saggion, K. Bontcheva, and H. Cunningham. Robust Generic and Query-based Summarisation. In *Proceedings of the European Chapter of Computational Linguistics (EACL), Research Notes and Demos*, 2003.

[7] Regina Barzilay and Kathleen R. Mckeown. Sentence fusion for multi-document news summarization. *Computational Linguistics*, 31:297–328, 2005.

[8] K. Bontcheva and Y. Wilks. Tailoring Automatically Generated Hypertext. *User Modeling and User-Adapted Interaction*, 2004. Special issue on Language-Based Interaction.

[9] Jin Yu, Ehud Reiter, Jim Hunter, and Chris Mellish. Choosing the content of textual summaries of large time-series data sets. *Natural Language Engineering*, 13(1):25–49, March 2007.

[10] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, April 1958.

[11] Dragomir R. Radev, Hongyan Jing, Malgorzata Styś, and Daniel Tam. Centroid-based summarization of multiple documents. *Information Processing and Management*, 40(6):919–938, November 2004.

[12] Giuseppe Carenini and Jackie Chi Kit Cheung. Extractive vs. NLG-based abstractive summarization of evaluative text: the effect of corpus controversiality. In *Proceedings of the Fifth International Natural Language Generation Conference*, INLG '08, pages 33–41, 2008.

[13] Ellen M. Voorhees. Overview of the TREC 2003 Question Answering Track. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[14] Daniel M. Dunlavy, John M. Conroy, Judith D. Schlesinger, Sarah A. Goodman, Mary Ellen Okurowski, Dianne P. O'Leary, and Hans van Halteren. Performance of a Three-Stage System for Multi-Document Summarization. In *Proceedings of the Document Understanding Conference (DUC 2003)*, 2003.

[15] Mark A. Greenwood. *Open Domain Question Answering*. PhD thesis, The University of Sheffield, 2006.

[16] Manish Narwaria and Weisi Lin. Objective image quality assessment based on support vector regression. *Neural Networks, IEEE Transactions on*, 21(3):515–519, 2010.

[17] Eric A. Silva, Karen Panetta, and Sos S. Agaian. Quantifying image similarity using measure of enhancement by entropy. In *Defense and Security Symposium*, pages 65790U–65790U. International Society for Optics and Photonics, 2007.

[18] T. M. Kusuma U. Engelke, H.-J. Zepernick. Subjective quality assessment for wireless image communication: The wireless imaging quality database. In *Int. Workshop on Video Processing and Quality Metrics (VPQM)*, 2010.

[19] Zhou Wang and Alan C. Bovik. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 2(1):1–156, 2006.

[20] Xinbo Gao, Wen Lu, Dacheng Tao, and Xuelong Li. Image quality assessment and human visual system. In *Visual Communications and Image Processing 2010*, pages 77440Z–77440Z. International Society for Optics and Photonics, 2010.

[21] David B. Lowe and Athula Ginige. Image quality assessment using an image activity weighting and the hvs response. *Image and Vision Computing NZ*, 93:169–176, 1993.

[22] Yubing Tong, Hubert Konik, Faouzi Cheikh, and Alain Tremeau. Full reference image quality assessment based on saliency map analysis. *Journal of Imaging Science and Technology*, 54(3):30503–1, 2010.

[23] Zhou Wang, Hamid R. Sheikh, and Alan C. Bovik. Objective video quality assessment. *The handbook of video databases: design and applications*, pages 1041–1078, 2003.

[24] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, volume 2, pages 1398–1402. IEEE, 2003.

[25] Zhou Wang, Alan C. Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.

[26] Venkata D. Rao and Pratap L. Reddy. Image quality assessment based on perceptual structural similarity. In *Pattern Recognition and Machine Intelligence*, pages 87–94. Springer, 2007.

[27] Zhou Wang and Eero P. Simoncelli. Reduced-reference image quality assessment using a wavelet-domain natural image statistic model. In *Proc. of SPIE Human Vision and Electronic Imaging*, volume 5666, pages 149–159, 2005.

[28] Zhou Wang, Guixing Wu, Hamid R. Sheikh, Eero P. Simoncelli, En-Hui Yang, and Alan C. Bovik. Quality-aware images. *Image Processing, IEEE Transactions on*, 15(6):1680–1689, 2006.

[29] Anuj Srivastava, Ann B. Lee, Eero P. Simoncelli, and S-C Zhu. On advances in statistical modeling of natural images. *Journal of mathematical imaging and vision*, 18(1):17–33, 2003.

[30] Anush Krishna Moorthy and Alan Conrad Bovik. A two-step framework for constructing blind image quality indices. *Signal Processing Letters, IEEE*, 17(5):513–516, 2010.

[31] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *Image Processing, IEEE Transactions on*, 20(12):3350–3364, 2011.

[32] Anish Mittal, Anush Moorthy, and Alan Bovik. No-reference image quality assessment in the spatial domain. 2012.

[33] Michele A. Saad, Alan C. Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *Image Processing, IEEE Transactions on*, 21(8):3339–3352, 2012.

[34] Anish Mittal, Rajiv Soundararajan, and A Bovik. Making a Completely Blind Image Quality Analyzer. 2012.

[35] Shaoquan YU, Anyi ZHANG, and Hongwei LI. A Review of Estimating the Shape Parameter of Generalized Gaussian Distribution. *Journal of Computational Information Systems*, 8(21):9055–9064, 2012.

[36] Jorge Caviedes and Sabri Gurbuz. No-reference sharpness metric based on local edge kurtosis. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 3, pages III–53. IEEE, 2002.

[37] EePing Ong, Weisi Lin, Zhongkang Lu, Xiaokang Yang, Susu Yao, Feng Pan, Lijun Jiang, and F Moschetti. A no-reference quality metric for measuring image blur. In *Signal Processing and Its Applications, 2003. Proceedings. Seventh International Symposium on*, volume 1, pages 469–472. IEEE, 2003.

[38] Srenivas Varadarajan and Lina J. Karam. An improved perception-based no-reference objective image sharpness metric using iterative edge refinement. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 401–404. IEEE, 2008.

[39] Min Goo Choi, Jung Hoon Jung, and Jae Wook Jeon. No-reference image quality assessment using blur and noise, 2009.

[40] Hao Hu and Gerard de Haan. Low cost robust blur estimator. In *Image Processing, 2006 IEEE International Conference on*, pages 617–620. IEEE, 2006.

[41] Frederique Crete, Thierry Dolmiere, Patricia Ladret, and Marina Nicolas. The blur effect: perception and estimation with a new no-reference perceptual blur metric. *Human Vision and Electronic Imaging XII*, 6492:64920I, 2007.

[42] Niranjan D. Narvekar and Lina J. Karam. A no-reference perceptual image sharpness metric based on a cumulative probability of blur detection. In *Quality of Multimedia Experience, 2009. QoMEx 2009. International Workshop on*, pages 87–91. IEEE, 2009.

[43] Rania Hassen, Zhou Wang, and Magdy Salama. No-reference image sharpness assessment based on local phase coherence measurement. In *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 2434–2437. IEEE, 2010.

[44] Eli Peli. Contrast in complex images. *JOSA A*, 7(10):2032–2040, 1990.

[45] Luis Miguel Sanchez-Brea, Juan Antonio Quiroga, Angel Garcia-Botella, and Eusebio Bernabeu. Histogram-based method for contrast measurement. *Applied optics*, 39(23):4098–4106, 2000.

[46] Azeddine Beghdadi and Alain Le Negrate. Contrast enhancement technique based on local detection of edges. *Computer Vision, Graphics, and Image Processing*, 46(2):162–174, 1989.

[47] Ming-Ming Cheng, Guo-Xin Zhang, Niloy J. Mitra, Xiaolei Huang, and Shi-Min Hu. Global contrast based salient region detection. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 409–416. IEEE, 2011.

[48] Glenn S Maitz and David Gur. Joint photographic experts group (jpeg) compatible data compression of mammograms. *Journal of Digital Imaging*, 7(3):123–132, 1994.

[49] Lydia Meesters and Jean-Bernard Martens. A single-ended blockiness measure for JPEG-coded images. *Signal Processing*, 82(3):369–387, 2002.

[50] Rémi Barland and Abdelhakim Saadane. Reference free quality metric for JPEG-2000 compressed images. In *Signal Processing and Its Applications, 2005. Proceedings of the Eighth International Symposium on*, volume 1, pages 351–354. IEEE, 2005.

[51] Shan Suthaharan. No-reference visually significant blocking artifact metric for natural scene images. *Signal Processing*, 89(8):1647–1652, 2009.

[52] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual blur and ringing metrics: application to JPEG2000. *Signal Processing: Image Communication*, 19(2):163–172, 2004.

[53] ZM Parvez Sazzad, Y Kawayoke, and Y Horita. No reference image quality assessment for JPEG2000 based on spatial features. *Signal Processing: Image Communication*, 23(4):257–268, 2008.

[54] Hamid Rahim Sheikh, Alan Conrad Bovik, and Lawrence Cormack. No-reference quality assessment using natural scene statistics: JPEG2000. *Image Processing, IEEE Transactions on*, 14(11):1918–1927, 2005.

[55] Hamid R. Sheikh, Zhou Wang, L. Cormack, and Alan C. Bovik. Live image quality assessment database. *[online] http://live. ece. utexas. edu/research/quality*, 2003.

[56] Eric C. Larson and Damon Chandler. Categorical image quality (CSIQ) database. *[Online], http://vision. okstate. edu/csiq*, 2010.

[57] Shyamprasad Chikkerur, Vijay Sundaram, Martin Reisslein, and Lina J. Karam. Objective video quality assessment methods: A classification, review, and performance comparison. *Broadcasting, IEEE Transactions on*, 57(2):165–182, 2011.

[58] Zhou Wang and Alan C Bovik. A universal image quality index. *Signal Processing Letters, IEEE*, 9(3):81–84, 2002.

[59] Zhou Wang, Ligang Lu, and Alan C Bovik. Video quality assessment based on structural distortion measurement. *Signal processing: Image communication*, 19(2):121–132, 2004.

[60] Zhou Wang and Qiang Li. Video quality assessment using a statistical model of human visual speed perception. *JOSA A*, 24(12):B61–B69, 2007.

[61] Margaret H. Pinson and Stephen Wolf. A new standardized method for objectively measuring video quality. *Broadcasting, IEEE Transactions on*, 50(3):312–322, 2004.

[62] Andrew B. Watson, James Hu, and John F. MCgowan. Digital video quality metric based on human vision. *Journal of Electronic imaging*, 10(1):20–29, 2001.

[63] Kalpana Seshadrinathan and Alan Conrad Bovik. Motion tuned spatio-temporal quality assessment of natural videos. *Image Processing, IEEE Transactions on*, 19(2):335–350, 2010.

[64] Andries P. Hekstra, J.G. Beerends, D. Ledermann, F.E. De Caluwe, S. Kohler, R.H. Koenen, S. Rihs, M. Ehrsam, and D. Schlauss. PVQM–A perceptual video quality measure. *Signal processing: Image communication*, 17(10):781–798, 2002.

[65] Damon M. Chandler and Sheila S. Hemami. VSNR: A wavelet-based visual signal-to-noise ratio for natural images. *Image Processing, IEEE Transactions on*, 16(9):2284–2298, 2007.

[66] Toru Yamada, Yoshihiro Miyamoto, and Masahiro Serizawa. No-reference video quality estimation based on error-concealment effectiveness. In *Packet Video 2007*, pages 288–293. IEEE, 2007.

[67] Mylene C.Q. Farias and Sanjit K. Mitra. No-reference video quality metric based on artifact measurements. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3, pages III–141. IEEE, 2005.

[68] Christian Keimel, Tobias Oelbaum, and Klaus Diepold. No-reference video quality evaluation for high-definition video. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1145–1148. IEEE, 2009.

[69] Atsuro Ichigaya, Masaaki Kurozumi, Naohiro Hara, Yukihiro Nishida, and Eisuke Nakasu. A method of estimating coding PSNR using quantized DCT coefficients. *Circuits and Systems for Video Technology, IEEE Transactions on*, 16(2):251–259, 2006.

[70] Deepak S. Turaga, Yingwei Chen, and Jorge Caviedes. No reference PSNR estimation for compressed pictures. *Signal Processing: Image Communication*, 19(2):173–184, 2004.

[71] MPEG-7 ISO/IEC 15938-3. *FCD information technology - multimedia content description interface - part 3: Visual*. 2002.

[72] Shih-Fu Chang, Thomas Sikora, and A. Purl. Overview of the MPEG-7 standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):688–695, 2001.

[73] Bangalore S. Manjunath, J.-R. Ohm, Vinod V. Vasudevan, and Akio Yamada. Color and texture descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, 11(6):703–715, 2001.

[74] Nai-Chung Yang, Wei-Han Chang, Chung-Ming Kuo, and Tsia-Hsing Li. A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval. *Journal of Visual Communication and Image Representation*, 19(2):92–105, 2008.

[75] Leszek Cieplinski. MPEG-7 color descriptors and their applications. In *Computer analysis of images and patterns*, pages 11–20. Springer, 2001.

[76] Radomir S Stanković and Bogdan J Falkowski. The Haar wavelet transform: its status and achievements. *Computers & Electrical Engineering*, 29(1):25–44, 2003.

[77] Dean S. Messing, Peter van Beek, and James H. Errico. The MPEG-7 colour structure descriptor: Image description using colour and local spatial information. In *Image Processing, 2001. Proceedings. 2001 International Conference on*, volume 1, pages 670–673. IEEE, 2001.

[78] R. Balasubramani and Dr V. Kannan. Efficient use of MPEG-7 color layout and edge histogram descriptors in cbir systems. *Global Journal of Computer Science and Technology*, 9(4), 2009.

[79] A. Mufit Ferman, S. Krishnamachari, A. Murat Tekalp, A. Abdel-Mottaleb, and R. Mehrota. Group-of-frame/picture color histogram. *Global Journal of Computer Science and Technology*, 2000.

[80] Peng Wu, Yong Man Ro, Chee Sun Won, and Yanglim Choi. Texture descriptors in MPEG-7. In *Computer Analysis of Images and Patterns*, pages 21–28. Springer, 2001.

[81] Yong Man Ro, Munchurl Kim, Ho Kyung Kang, BS Manjunath, and Jinwoong Kim. MPEG-7 homogeneous texture descriptor. *ETRI journal*, 23(2):41–51, 2001.

[82] Chee Sun Won, Dong Kwon Park, and Soo-Jun Park. Efficient use of mpeg-7 edge histogram descriptor. *Etri Journal*, 24(1):23–30, 2002.

[83] P. Wu, B.S. Manjunath, S. Newsam, and H.D. Shin. A texture descriptor for browsing and similarity retrieval. *Signal Processing: Image Communication*, 16(1):33–43, 2000.

[84] Wei-Ying Ma and BS Manjunath. A texture thesaurus for browsing large aerial photographs. *Journal of the American Society for Information Science*, 49(7):633–648, 1998.

[85] Dengsheng Zhang and Guojun Lu. Evaluation of MPEG-7 shape descriptors against other shape descriptors. *Multimedia Systems*, 9(1):15–30, 2003.

[86] Dengsheng Zhang and Guojun Lu. A comparative study of curvature scale space and Fourier descriptors for shape-based image retrieval. *Journal of Visual Communication and Image Representation*, 14(1):39–57, 2003.

[87] Savvas A. Chatzichristofis and Yiannis S. Boutalis. CEDD: color and edge directivity descriptor: a compact descriptor for image indexing and retrieval. In *Computer Vision Systems*, pages 312–322. Springer, 2008.

[88] Savvas A. Chatzichristofis and Yiannis S. Boutalis. Fcth: Fuzzy color and texture histogram-a low level feature for accurate image retrieval. In *Image Analysis for Multimedia Interactive Services, 2008. WIAMIS'08. Ninth International Workshop on*, pages 191–196. 2008.

[89] Savvas A. Chatzichristofis, Yiannis S. Boutalis, and Mathias Lux. Selection of the proper compact composite descriptor for improving content based image retrieval. In *Proceedings of the 6th IASTED International Conference*, volume 134643, page 064, 2009.

[90] Zhi Li, Guizhong Liu, Haixia Jiang, and Xuemin Qian. Image copy detection using a robust gabor texture descriptor. In *Proceedings of the First ACM workshop on Large-scale multimedia retrieval and mining*, pages 65–72. ACM, 2009.

[91] Yong Xu, Xiong Yang, Haibin Ling, and Hui Ji. A new texture descriptor using multifractal analysis in multi-orientation wavelet pyramid. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 161–168. IEEE, 2010.

[92] Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using Local Binary Patterns with an application to facial expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(6):915–928, 2007.

[93] David Gerónimo, Antonio López, Daniel Ponsa, and Angel D Sappa. Haar Wavelets and Edge Orientation Histograms for On–Board Pedestrian Detection. In *Pattern Recognition and Image Analysis*, pages 418–425. Springer, 2007.

[94] Xinbo Gao, Bing Xiao, Dacheng Tao, and Xuelong Li. Image categorization: Graph edit distance+ edge direction histogram. *Pattern Recognition*, 41(10):3179–3191, 2008.

[95] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[96] Anna Bosch, Andrew Zisserman, and Xavier Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 401–408. ACM, 2007.

[97] Vijay Chandrasekhar, Gabriel Takacs, David M Chen, Sam S Tsai, Yuriy Reznik, Radek Grzeszczuk, and Bernd Girod. Compressed histogram of gradients: A low-bitrate descriptor. *International journal of computer vision*, 96(3):384–399, 2012.

[98] Matthijs Douze, Hervé Jégou, Harsimrat Sandhawalia, Laurent Amsaleg, and Cordelia Schmid. Evaluation of GIST descriptors for web-scale image search. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 19. ACM, 2009.

[99] Raman Maini and Himanshu Aggarwal. Study and comparison of various image edge detection techniques. *International Journal of Image Processing (IJIP)*, 3(1):1–11, 2009.

[100] Wenshuo Gao, Xiaoguang Zhang, Lei Yang, and Huizhong Liu. An improved Sobel edge detection. In *Computer Science and Information Technology (ICCSIT), 2010 3rd IEEE International Conference on*, volume 5, pages 67–71. IEEE, 2010.

[101] Lei Yang, Dewei Zhao, Xiaoyu Wu, Hui Li, and Jun Zhai. An improved Prewitt algorithm for edge detection based on noised image. In *Image and Signal Processing (CISP), 2011 4th International Congress on*, volume 3, pages 1197–1200. IEEE, 2011.

[102] Bill Green. Canny edge detection tutorial. *From web resource. www. pages. drexel. edu/weg22/cantut. html*, 2002.

[103] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004.

[104] Frederic Jurie and Bill Triggs. Creating efficient codebooks for visual recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 604–610. IEEE, 2005.

[105] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 524–531. IEEE, 2005.

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

[107] Yan Ke and Rahul Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506. IEEE, 2004.

[108] Krystian Mikolajczyk and Cordelia Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, 2005.

[109] Eric N Mortensen, Hongli Deng, and Linda Shapiro. A SIFT descriptor with global context. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 184–190. IEEE, 2005.

[110] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1265–1278, 2005.

[111] Gangqiang Zhao, Ling Chen, Gencai Chen, and Junsong Yuan. KPB-SIFT: a compact local feature descriptor. In *Proceedings of the international conference on Multimedia*, pages 1175–1178. ACM, 2010.

[112] Yun-Ta Tsai, Quan Wang, and Suya You. CDIKP: a highly-compact local feature descriptor. In *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*, pages 1–4. IEEE, 2008.

[113] Koen van de Sande, Theo Gevers, and Cees Snoek. Evaluation of color descriptors for object and scene recognition. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[114] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded up robust features. In *Computer Vision–ECCV 2006*, pages 404–417. Springer, 2006.

[115] Ran Tao. Visual concept detection and real time object detection. *arXiv preprint arXiv:1104.0582*, 2011.

[116] Ali Jalilvand, Hamidreza Shayegh Boroujeni, and Nasrollah Moghadam Charkari. CWSURF: A novel coloured local invariant descriptor based on SURF. In *Computer and Knowledge Engineering (ICCKE), 2011 1st International eConference on*, pages 214–219. IEEE, 2011.

[117] J.-M. Geusebroek, Rein van den Boomgaard, Arnold W.M. Smeulders, and Hugo Geerts. Color invariance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(12):1338–1350, 2001.

[118] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, 2002.

[119] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using affine-invariant regions. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–319. IEEE, 2003.

[120] X. Shi, A.L. Ribeiro Castro, R. Manduchi, and R. Montgomery. Rotational invariant operators based on steerable filter banks. *Signal Processing Letters, IEEE*, 13(11):684–687, 2006.

[121] Tai Sing Lee. Image representation using 2D Gabor wavelets. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(10):959–971, 1996.

[122] Jun Zhang, Heng Zhao, and Jimin Liang. Continuous rotation invariant local descriptors for texton dictionary-based texture classification. *Computer Vision and Image Understanding*, 2012.

[123] Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[124] Gustavo Carneiro and Allan D. Jepson. Multi-scale phase-based local features. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–736. IEEE, 2003.

[125] Hong Cheng, Zicheng Liu, Nanning Zheng, and Jie Yang. A deformable local image descriptor. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[126] Liu Yawei, Li Jianwei, and Zhang Xiaohong. A new local feature descriptor: Covariant support region. In *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, volume 4, pages 346–351. IEEE, 2009.

[127] Pierre Tirilly, Vincent Claveau, and Patrick Gros. Language modeling for Bag-of-Visual words image categorization. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 249–258. ACM, 2008.

[128] Phillipe Salembier, Thomas Sikora, and BS Manjunath. *Introduction to MPEG-7: multimedia content description interface*. John Wiley & Sons, Inc., 2002.

[129] Mathias Lux. Caliph & emir: Mpeg-7 photo annotation and retrieval. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 925–926. ACM, 2009.

[130] Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. Dublin core metadata for resource discovery. *Internet Engineering Task Force RFC*, 2413:222, 1998.

[131] Hsiang-Cheh Huang and Wai-Chi Fang. Metadata-based image watermarking for copyright protection. *Simulation Modelling Practice and Theory*, 18(4):436–445, 2010.

[132] Core, IPTC. Photo Metadata, 2010.

[133] Henrik Eriksson. The semantic-document approach to combining documents and ontologies. *International Journal of Human-Computer Studies*, 65(7):624–639, 2007.

[134] Antoine Pigeau and Marc Gelgon. Organizing a personal image collection with statistical model-based icl clustering on spatio-temporal camera phone meta-data. *Journal of Visual Communication and Image Representation*, 15(3):425–445, 2004.

[135] Matthew Boutell and Jiebo Luo. Bayesian fusion of camera metadata cues in semantic scene classification. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–623. IEEE, 2004.

[136] Masaharu Hirota, Shohei Yokoyama, Naoki Fukuta, and Hiroshi Ishikawa. Constraint-based clustering of image search results using photo Metadata and low-level image features. In *Computer and Information Science 2010*, pages 165–178. Springer, 2010.

[137] Chaur-Chin Chen and Hsueh-Ting Chu. Similarity measurement between images. In *Computer Software and Applications Conference, 2005. COMPSAC 2005. 29th Annual International*, volume 2, pages 41–42. IEEE, 2005.

[138] Sung-Hyuk Cha. Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1, 2007.

[139] Peter Grabusts et al. The choice of metrics for clustering algorithms. In *Proceedings of the 8th International Scientific and Practical Conference*, volume 2, 2011.

[140] Tomáš Skopal, Tomáš Bartoš, and Jakub Lokoč. On (not) indexing quadratic form distance by metric access methods. In *Proceedings of the 14th International Conference on Extending Database Technology*, pages 249–258. ACM, 2011.

[141] Christian Beecks, Merih Seran Uysal, and Thomas Seidl. Signature quadratic form distances for content-based similarity. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 697–700. ACM, 2009.

[142] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[143] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. Comparing images using the Hausdorff distance. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 15(9):850–863, 1993.

[144] Bo Gun Park, Kyoung Mu Lee, and Sang Uk Lee. Color-based image retrieval using perceptually modified Hausdorff distance. *Journal on Image and Video Processing*, 2008:4, 2008.

[145] Wee Kheng Leow and Rui Li. The analysis and applications of adaptive-binning color histograms. *Computer Vision and Image Understanding*, 94(1):67–91, 2004.

[146] Jie Yu, Jaume Amores, Nicu Sebe, Petia Radeva, and Qi Tian. Distance learning for similarity estimation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(3):451–462, 2008.

[147] Abraham Bookstein, Shmuel Tomi Klein, and Timo Raita. Fuzzy Hamming distance: a new dissimilarity measure. In *Combinatorial Pattern Matching*, pages 86–97. Springer, 2006.

[148] AKMSSA Vadivel, AK Majumdar, and Shamik Sural. Performance comparison of distance metrics in content-based image retrieval applications. In *Proc. of Internat. Conf. on Information Technology, Bhubaneswar, India*, pages 159–164, 2003.

[149] Horst Eidenberger. Distance measures for MPEG-7-based retrieval. In *Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, pages 130–137. ACM, 2003.

[150] Cornelis Joost van Rijsbergen. *The geometry of information retrieval*. Cambridge University Press, 2004.

[151] Elisa Maria Todarello, Walter Allasia, and Mario Stroppiana. Mpeg-7 features in hilbert spaces: querying similar images with linear superpositions. In *Quantum Interaction*, pages 223–228. Springer, 2011.

[152] Federico F Barresi, Giuseppe Battista, Jacopo Pellegrino, and Walter Allasia. Quantistic approach for classification of images. In *MMEDIA 2013, The Fifth International Conferences on Advances in Multimedia*, pages 7–11, 2013.

[153] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern classification*. John Wiley & Sons, 2012.

[154] Parvesh Kumar and Siri Krishan Wasan. Comparative study of K-means, pam and rough K-means algorithms using cancer datasets. In *Proceedings of CSIT: 2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)*, volume 1, pages 136–140, 2009.

[155] Vit Niennattrakul and Chotirat Ann Ratanamahatana. Clustering multimedia data using time series. In *Hybrid Information Technology, 2006. ICHIT'06. International Conference on*, volume 1, pages 372–379. IEEE, 2006.

[156] Stephen C. Johnson. Hierarchical clustering schemes. *Psychometrika*, 32(3):241–254, 1967.

[157] Peter H.A. Sneath, Robert R. Sokal, et al. *Numerical taxonomy. The principles and practice of numerical classification*. 1973.

[158] Santhana Krishnamachari and Mohamed Abdel-Mottaleb. Image browsing using hierarchical clustering. In *Computers and Communications, 1999. Proceedings. IEEE International Symposium on*, pages 301–307. IEEE, 1999.

[159] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan. Unsupervised image-set clustering using an information theoretic framework. *Image Processing, IEEE Transactions on*, 15(2):449–458, 2006.

[160] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[161] Chris Ding and Xiaofeng He. Cluster merging and splitting in hierarchical clustering algorithms. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 139–146. IEEE, 2002.

[162] Sotiris K. Tasoulis, Dimitris K. Tasoulis, and Vassilis P. Plagianakos. Enhancing principal direction divisive clustering. *Pattern Recognition*, 43(10):3391–3411, 2010.

[163] Hong Liu and Xiaohong Yu. Application research of k-means clustering algorithm in image retrieval system. In *Proceedings of the Second Symposium International Computer Science and Computational Technology (ISCSCT 09), Huangshan, PR China*, pages 274–277, 2009.

[164] Shalini S. Singh and N.C. Chauhan. K-means v/s K-medoids: A Comparative Study. In *National Conference on Recent Trends in Engineering & Technology,(13-14 May 2011)*, 2011.

[165] Fei Wang, Yueming Lu, Fangwei Zhang, and Songlin Sun. New Method Based on Fuzzy C-Means Algorithm for Search Results Clustering. In *Trustworthy Computing and Services*, pages 263–270. Springer, 2013.

[166] Bin Gao, Tie-Yan Liu, Tao Qin, Xin Zheng, Qian-Sheng Cheng, and Wei-Ying Ma. Web image clustering by consistent utilization of visual features and surrounding texts. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 112–121. ACM, 2005.

[167] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[168] Manjeet Rege, Ming Dong, and Jing Hua. Graph theoretical framework for simultaneously integrating visual and textual features for efficient web image clustering. In *Proceedings of the 17th international conference on World Wide Web*, pages 317–326. ACM, 2008.

[169] Alejandro Jaimes. *Conceptual structures and computational methods for indexing and organization of visual information*. PhD thesis, Columbia University, 2003.

[170] Dong-Qing Zhang and Shih-Fu Chang. Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 877–884. ACM, 2004.

[171] Yan Ke, Rahul Sukthankar, and Larry Huston. Efficient near-duplicate detection and sub-image retrieval. In *ACM Multimedia*, volume 4, page 5, 2004.

[172] Multimedia Graphics Pack. Media Graphics International 270000. 1998.

[173] Xin Yang, Qiang Zhu, and Kwang-Ting Cheng. Near-duplicate detection for images and videos. In *Proceedings of the First ACM workshop on Large-scale multimedia retrieval and mining*, pages 73–80. ACM, 2009.

[174] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Computer Vision–ECCV 2006*, pages 288–301. Springer, 2006.

[175] Dhiraj Joshi, Ritendra Datta, Elena Fedorovskaya, Quang-Tuan Luong, James Z. Wang, Jia Li, and Jiebo Luo. Aesthetics and emotions in images. *Signal Processing Magazine, IEEE*, 28(5):94–115, 2011.

[176] Phivos Mylonas, Evaggelos Spyrou, Yannis Avrithis, and Stefanos Kollias. Using visual context and region semantics for high-level concept detection. *Multimedia, IEEE Transactions on*, 11(2):229–243, 2009.

[177] Krystian Mikolajczyk and Cordelia Schmid. Indexing based on scale invariant interest points. In *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, volume 1, pages 525–531. IEEE, 2001.

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

[179] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for Bag-of-Features image classification. In *Computer Vision–ECCV 2006*, pages 490–503. Springer, 2006.

[180] Pantelis Elinas. Matching Maximally Stable Extremal Regions using edge information and the Chamfer distance function. In *Computer and Robot Vision (CRV), 2010 Canadian Conference on*, pages 17–24. IEEE, 2010.

[181] P-E Forssén. Maximally stable colour regions for recognition and matching. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[182] Koen E.A. Van De Sande, Theo Gevers, and Arnold W.M. Smeulders. The university of amsterdams concept detection system at imageCLEF 2009. In *Multilingual Information Access Evaluation II. Multimedia Experiments*, pages 261–268. Springer, 2010.

[183] Muhammad Atif Tahir, Fei Yan, Mark Barnard, Muhammad Awais, Krystian Mikolajczyk, and Josef Kittler. The university of surrey visual concept detection system at imageCLEF@ ICPR: working notes. In *Recognizing Patterns in Signals, Speech, Images and Videos*, pages 162–170. Springer, 2010.

[184] A. Moumtzidou, P. Sidiropoulos, Vrochidis S., Gkalelis N., S. Nikolopoulos, V. Mezaris, I. Kompatsiaris, and P. Patras. ITI-CERTH participation to TRECVID 2011. In *In TRECVID 2011 Workshop, Gaithersburg, MD, USA*, volume 12/2012, 2011.

[185] Engin Tola, Vincent Lepetit, and Pascal Fua. Daisy: An efficient dense descriptor applied to wide-baseline stereo. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(5):815–830, 2010.

[186] Shih-Fu Chang, Dan Ellis, Wei Jiang, Keansub Lee, Akira Yanagawa, Alexander C. Loui, and Jiebo Luo. Large-scale multimodal semantic concept detection for consumer video. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 255–264. ACM, 2007.

[187] Jingtian Jiang, Xiaoguang Rui, and Nenghai Yu. Feature Annotation for Visual Concept Detection in ImageCLEF 2008. *Working Notes of CLEF*, 2008.

[188] Jan C. van Gemert, Cor J. Veenman, Arnold W.M. Smeulders, and J.-M. Geusebroek. Visual word ambiguity. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(7):1271–1283, 2010.

[189] Krystian Mikolajczyk and Jiri Matas. Improving descriptors for fast tree matching by optimal linear projection. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[190] Frank Moosmann, Bill Triggs, Frederic Jurie, et al. Fast discriminative visual codebooks using randomized clustering forests. *Advances in Neural Information Processing Systems 19*, pages 985–992, 2007.

[191] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond Bags-of-Features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.

[192] Yannis Kalantidis, Evaggelos Spyrou, Phivos Mylonas, and Stefanos Kollias. Using a region and visual word approach towards semantic image retrieval. In *Semantic Media Adaptation and Personalization (SMAP), 2010 5th International Workshop on*, pages 85–89. IEEE, 2010.

[193] Li Zhou, Zongtan Zhou, and Dewen Hu. Scene classification using a multi-resolution Bag-of-Features model. *Pattern Recognition*, 2012.

[194] V. N. Vapnik. Statistical learning theory. New York: Wiley, 1998.

[195] Duy-Dinh Le and Shin'ichi Satoh. Efficient concept detection by fusing simple visual features. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1839–1840. ACM, 2009.

[196] Evaggelos Spyrou, Giorgos Tolias, and Yannis Avrithis. Large scale concept detection in video using a region thesaurus. In *Advances in Multimedia Modeling*, pages 197–207. Springer, 2009.

[197] Rami Al Batal and Philippe Mulhem. MRIM-LIG at ImageCLEF 2010 Visual Concept Detection and Annotation task. 2010.

[198] Xin Jin, Anbang Xu, Rongfang Bie, and Ping Guo. Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles. In *Data Mining for Biomedical Applications*, pages 106–115. Springer, 2006.

[199] Sabri Boughorbel, J.-P. Tarel, and Nozha Boujemaa. Generalized histogram intersection kernel for image recognition. In *Image Processing, 2005. ICIP 2005. IEEE International Conference on*, volume 3, pages III–161. IEEE, 2005.

[200] Yin-Wen Chang, Cho-Jui Hsieh, Kai-Wei Chang, Michael Ringgaard, and Chih-Jen Lin. Training and testing low-degree polynomial data mappings via linear SVM. *The Journal of Machine Learning Research*, 99:1471–1490, 2010.

[201] Nikolaos Gkalelis, Vasileios Mezaris, Ioannis Kompatsiaris, and Tania Stathaki. Linear subclass support vector machines. *Signal Processing Letters, IEEE*, 19(9):575–578, 2012.

[202] Zhouyu Fu, Antonio Robles-Kelly, and Jun Zhou. Mixing linear SVMs for nonlinear classification. *Neural Networks, IEEE Transactions on*, 21(12):1963–1975, 2010.

[203] Elizabeth Tapia, José C González, Alexander Hütermann, and Javier García. Beyond boosting: Recursive ECOC learning machines. In *Multiple Classifier Systems*, pages 62–71. Springer, 2004.

[204] Meng Wang, Xian-Sheng Hua, Richang Hong, Jinhui Tang, Guo-Jun Qi, and Yan Song. Unified video annotation via multigraph learning. *Circuits and Systems for Video Technology, IEEE Transactions on*, 19(5):733–746, 2009.

[205] Changbo Yang, Ming Dong, and Farshad Fotouhi. Region based image annotation through multiple-instance learning. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 435–438. ACM, 2005.

[206] Céline Hudelot, Jamal Atif, and Isabelle Bloch. Fuzzy spatial relation ontology for image interpretation. *Fuzzy Sets and Systems*, 159(15):1929–1951, 2008.

[207] Purvi Prajapati, Amit Thakkar, and Amit Ganatra. A Survey and Current Research Challenges in Multi-Label Classification Methods. *International Journal of Soft Computing*, 2.

[208] Cha Zhang and Zhengyou Zhang. A survey of recent advances in face detection. Technical report, Tech. rep., Microsoft Research, 2010.

[209] Paul Viola and Michael J. Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[210] Konstantinos G. Derpanis. Integral image-based representations. *Department of Computer Science and Engineering York University Paper*, 1(2):1–6, 2007.

[211] Mohammad J. Saberian and Nuno Vasconcelos. Boosting classifier cascades. In *NIPS*, volume 23, pages 2047–2055, 2010.

[212] Mrs Sunita Roy and Samir K. Bandyopadhyay. Face detection using a hybrid approach that combines HSV and RGB. 2013.

[213] Devendra Singh Raghuvanski and Dheeraj Agrawal. Human face detection by using skin color segmentation, face features and regions properties. *International Journal of Computer Application*, 38(9), 2013.

[214] Rainer Lienhart and Jochen Maydt. An extended set of Haar-like features for rapid object detection. In *Image Processing. 2002. Proceedings. 2002 International Conference on*, volume 1, pages I–900. IEEE, 2002.

[215] Marcin Wojnarski. Absolute contrasts in face detection with adaBoost cascade. In *Rough Sets and Knowledge Technology*, pages 174–180. Springer, 2007.

[216] Shengye Yan, Shiguang Shan, Xilin Chen, and Wen Gao. Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.

[217] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[218] Panagiotis Antonopoulos, Nikos Nikolaidis, and Ioannis Pitas. Hierarchical face clustering using SIFT image features. In *Computational Intelligence in Image and Signal Processing, 2007. CIISP 2007. IEEE Symposium on*, pages 325–329. IEEE, 2007.

[219] N. Vretos, V. Solachildis, and I. Pitas. A mutual information based face clustering algorithm for movies. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1013–1016. IEEE, 2006.

[220] Bisser Raytchev and Hiroshi Murase. Unsupervised face recognition by associative chaining. *Pattern Recognition*, 36(1):245–257, 2003.

[221] Shaogang Gong, Stephen McKenna, and John J. Collins. An investigation into face pose distributions. In *Automatic Face and Gesture Recognition, 1996., Proceedings of the Second International Conference on*, pages 265–270. IEEE, 1996.

[222] Panpan Huang, Yunhong Wang, and Ming Shao. A new method for multi-view face clustering in video sequence. In *Data Mining Workshops, 2008. ICDMW'08. IEEE International Conference on*, pages 869–873. IEEE, 2008.

[223] Nikolaos Gkalelis, Vasileios Mezaris, and Ioannis Kompatsiaris. A joint content-event model for event-centric multimedia indexing. In *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, pages 79–84. IEEE, 2010.

[224] Norman R. Brown. On the prevalence of event clusters in autobiographical memory. *Social Cognition*, 23(1):35–69, 2005.

[225] Pavan Turaga, Rama Chellappa, Venkatramana S. Subrahmanian, and Octavian Udrea. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(11):1473–1488, 2008.

[226] Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval*, pages 1–29, 2012.

[227] Madirakshi Das and Alexander C. Loui. Event classification in personal image collections. In *Multimedia and Expo, 2009. ICME 2009. IEEE International Conference on*, pages 1660–1663. IEEE, 2009.

[228] Amit Singhal, Jiebo Luo, and Weiyu Zhu. Probabilistic spatial context models for scene content understanding. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–235. IEEE, 2003.

[229] Ivan Tankoyeu, Javier Paniagua, Julian Stöttinger, and Fausto Giunchiglia. Event detection and scene attraction by very simple contextual cues. In *Proceedings of the 2011 joint ACM workshop on Modeling and Representing Events*, pages 1–6. ACM, 2011.

[230] Mohamed Ayari, Jonathan Delhumeau, Matthijs Douze, Hervé Jégou, Danila Potapov, Jérôme Revaud, Cordelia Schmid, Jiangbo Yuan, et al. INRIA@ TRECVID 2011: Copy Detection & Multimedia Event Detection. In *Proceedings of NIST TRECVID Workshop*, 2011.

[231] Nakamasa Inoue, Yusuke Kamishima, Toshiya Wada, Koichi Shinoda, and Shunsuke Sato. Tokyotech+ canon at TRECVID 2011. In *Proceedings of NIST TRECVID Workshop*, 2011.

[232] Anastasia Moumtzidou, Anastasios Dimou, Nikolaos Gkalelis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. ITI-CERTH participation to TRECVID 2010. In *Proc. TRECVID 2010 Workshop. 8th TRECVID Workshop, Gaithersburg, MD, USA*, 2010.

[233] A. Moumtzidou, N. Gkalelis, P. Sidiropoulos, M. Dimopoulos, S. Nikolopoulos, S. Vrochidis, and I. Kompatsiaris. ITI-CERTH participation to TRECVID 2012. In *Proc. TRECVID 2010 Workshop. 8th TRECVID Workshop, Gaithersburg, MD, USA*, 2012.

[234] Alexandre R.J. Francois, Ram Nevatia, Jerry Hobbs, Robert C. Bolles, and John R. Smith. VERL: an ontology framework for representing and annotating video events. *MultiMedia, IEEE*, 12(4):76–86, 2005.

[235] Utz Westermann and Ramesh Jain. {rm E}-A Generic Event Model for Event-Centric Multimedia Data Management in eChronicle Applications. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 106–106. IEEE, 2006.

[236] Utz Westermann and Ramesh Jain. Toward a common event model for multimedia applications. *Multimedia, IEEE*, 14(1):19–29, 2007.

[237] Amarnath Gupta and Ramesh Jain. Managing event information: Modeling, retrieval, and applications. *Synthesis Lectures on Data Management*, 3(4):1–141, 2011.

[238] Nikolaos Gkalelis, Vasileios Mezaris, and Ioannis Kompatsiaris. Automatic event-based indexing of multimedia content using a joint content-event model. In *Proceedings of the 2nd ACM international workshop on Events in multimedia*, pages 15–20. ACM, 2010.

[239] M. Rincón and J. Martínez-Cantos. An annotation tool for video understanding. In *Computer Aided Systems Theory–EUROCAST 2007*, pages 701–708. Springer, 2007.

[240] Han Sloetjes and Peter Wittenburg. Annotation by Category: ELAN and ISO DCR. In *LREC*, 2008.

[241] Koen E.A. Van De Sande, Theo Gevers, and Cees G.M. Snoek. Evaluating color descriptors for object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1582–1596, 2010.

[242] Efthymia Tsamoura, Vasileios Mezaris, and Ioannis Kompatsiaris. Gradual transition detection using color coherence and other criteria in a video shot meta-segmentation framework. In *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on*, pages 45–48. IEEE, 2008.

[243] Chris Harris and Mike Stephens. A combined corner and edge detector. In *Alvey vision conference*, volume 15, page 50. Manchester, UK, 1988.

[244] Bahjat Safadi and Georges Quénot. Evaluations of multi-learner approaches for concept indexing in video documents. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 88–91. LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2010.

[245] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511. IEEE, 2001.

[246] Shengcai Liao, Xiangxin Zhu, Zhen Lei, Lun Zhang, and Stan Z. Li. Learning multi-scale block local binary patterns for face recognition. In *Advances in Biometrics*, pages 828–837. Springer, 2007.

[247] Paul Over, George M Awad, Jon Fiscus, Brian Antonishek, Martial Michel, Alan F Smeaton, Wessel Kraaij, and Georges Quénot. TRECVID 2010–An overview of the goals, tasks, data, evaluation mechanisms, and metrics. 2011.

[248] John R Smith, Milind Naphade, and Apostol Natsev. Multimedia semantic indexing using model vectors. In *Multimedia and Expo, 2003. ICME'03. Proceedings. 2003 International Conference on*, volume 2, pages II–445. IEEE, 2003.

[249] M. Dimopoulos I. Kompatsiaris T. Stathaki N. Gkalelis, V. Mezaris. Video event detection using a subclass recoding error-correcting output codes framework. In *Multimedia and Expo, 2013. ICME'13. Proceedings. 2013 International Conference on*. IEEE, San Jose, CA, USA,, July 2013.

[250] Robert Mertens, Howard Lei, Luke Gottlieb, Gerald Friedland, and Ajay Divakaran. Acoustic super models for large scale video event detection. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 19–24. ACM, 2011.

[251] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Access Online via Elsevier, 1990.

[252] Nikolaos Gkalelis, Vasileios Mezaris, Ioannis Kompatsiaris, and Tania Stathaki. Mixture subclass discriminant analysis link to restricted Gaussian model and other generalizations. In *IEEE Transactions on Neural Networks and Learning Systems*, volume 2, pages 8–11. IEEE, 2013.

[253] Jade Goldstein, Vibhu Mittal, Jaime Carbonell, and Mark Kantrowitz. Multi-document summarization by sentence extraction. In *Proceedings of the 2000 NAACL-ANLPWorkshop on Automatic summarization - Volume 4*, NAACL-ANLP-AutoSum '00, pages 40–48, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[254] Jinxi Xu, Ana Licuanan, and Ralph Weischedel. TREC2003 QA at BBN: Answering Definitional Questions. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[255] Automatic Text, Hang Cui, Mstisalv Maslennikov, Long Qiu, Min-Yen Kan, and Tat-Seng Chua. QUALIFIER in trec-12 qa main task. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[256] Sasha Blair-Goldensohn, Kathleen R. McKeown, and Andrew Hazen Schlaikjer. A hybrid approach for answering definitional questions. In *Proceedings of the 12th Text REtrieval Conference*, 2003.

[257] Mark A. Greenwood and Horacio Saggion. A pattern based approach to answering factoid, list and definition questions. In *Proceedings of the 7th RIAO Conference (RIAO 2004)*, pages 232–243, Avignon, France, April 27 2004.

[258] Ramesh Jain, Mingyan Gao, Setareh Rafatirad, and Pinaki Sinha. Extractable mobile photo tags. In *Proceedings of the 1st international workshop on Mobile location-based service*, pages 3–10. ACM, 2011.

[259] Jun Li, Joo Hwee Lim, and Qi Tian. Automatic summarization for personal digital photos. In *Information, Communications and Signal Processing, 2003 and Fourth Pacific Rim Conference on Multimedia. Proceedings of the 2003 Joint Conference of the Fourth International Conference on*, volume 3, pages 1536–1540. IEEE, 2003.

[260] Hao Xu, Jingdong Wang, Xian-Sheng Hua, and Shipeng Li. Hybrid image summarization. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1217–1220. ACM, 2011.

[261] Minxian Li, Chunxia Zhao, and Jinhui Tang. Hybrid image summarization by hypergraph partition. *Neurocomputing*, 2013.

[262] Xiaowei Xu, Martin Ester, H.-P. Kriegel, and Jörg Sander. A distribution-based clustering algorithm for mining in large spatial databases. In *Data Engineering, 1998. Proceedings., 14th International Conference on*, pages 324–331. IEEE, 1998.

[263] Dong-Sung Ryu, Sun-Young Park, KwangHwi Kim, and Hwan-Gue Cho. A priority queue-based hierarchical photo clustering method using photo timestamps. In *Computer Science and Automation Engineering (CSAE), 2011 IEEE International Conference on*, volume 3, pages 152–156. IEEE, 2011.

[264] Da Deng. Content-based image collection summarization and comparison using self-organizing maps. *Pattern recognition*, 40(2):718–727, 2007.

[265] Marco La Cascia, Marco Morana, and Filippo Vella. Automatic image representation and clustering on mobile devices. *Journal of Mobile Multimedia*, 6(2):158–169, 2010.

[266] Chuljin Jang, Taijin Yoon, and Hwan-Gue Cho. A smart clustering algorithm for photo set obtained from multiple digital cameras. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1784–1791. ACM, 2009.

[267] Pinaki Sinha, Sharad Mehrotra, and Ramesh Jain. Summarization of personal photologs using multidimensional content and context. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 4. ACM, 2011.

[268] Pinaki Sinha. Summarization of archived and shared personal photo collections. In *Proceedings of the 20th international conference companion on World wide web*, pages 421–426. ACM, 2011.

[269] Esin Guldogan, Jari Kangas, and Moncef Gabbouj. Personalized representative image selection for shared photo albums. In *Computer Applications Technology (ICCAT), 2013 International Conference on*, pages 1–4. IEEE, 2013.

[270] Zhou Wang, Alan C. Bovik, and BL Evan. Blind measurement of blocking artifacts in images. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 3, pages 981–984. IEEE, 2000.

[271] Subhasis Saha and Rao Vemuri. An analysis on the effect of image activity on lossy coding performance. In *Circuits and Systems, 2000. Proceedings. ISCAS 2000 Geneva. The 2000 IEEE International Symposium on*, volume 3, pages 295–298. IEEE, 2000.

[272] Peng Fan, Aidong Men, Mengyang Chen, and Bo Yang. Color-SURF: A SURF descriptor with local kernel color histograms. In *Network Infrastructure and Digital Content, 2009. IC-NIDC 2009. IEEE International Conference on*, pages 726–730. IEEE, 2009.

[273] Engin Tola, Vincent Lepetit, and Pascal Fua. A fast local descriptor for dense matching. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[274] EBU. *EBU CORE METADATA SET*, volume version 1.4. 2013.

[275] Jie Yu, Jaume Amores, Nicu Sebe, and Qi Tian. A new study on distance metrics as similarity measurement. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 533–536. IEEE, 2006.

[276] Liviu Octavian and Mafteiu Scai. A new dissimilarity measure between feature-vectors. *International Journal of Computer Application*, 64(17), 2013.

[277] Takeshi Mita, Toshimitsu Kaneko, and Osamu Hori. Joint Haar-like features for face detection. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 2, pages 1619–1626. IEEE, 2005.

[278] R. Lienhart, A. Kuranov, and V. Pisarevsky. Empirical analysis of detection cascades of boosted classifiers for rapid object detection. Technical report, Tech. rep., MRL, Intel Labs, 2002.

[279] S. Charles Brubaker, Jianxin Wu, Matthew D. Mullin, and James M. Rehg. On the design of cascades of boosted ensembles for face detection. Technical report, Tech. rep., Georgia Institute of Technology, Intel LabsGIT-GVU-05-28, 2005.

[280] Changxin Gao, Nong Sang, and Qiling Tang. On selection and combination of weak learners in AdaBoost. *Pattern Recognition Letters*, 31(9):991–1001, 2010.

[281] Dai Dao-Qing and Yan Hong. Wavelets and face recognition. In *Face Recognition*, 2007.

[282] Deng Cai, Xiaofei He, and Jiawei Han. Efficient kernel discriminant analysis via spectral regression. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 427–432. IEEE, 2007.

[283] Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. Correlative multi-label video annotation. In *Proceedings of the 15th international conference on Multimedia*, pages 17–26. ACM, 2007.

[284] Ansgar Scherp, Thomas Franz, Carsten Saathoff, and Steffen Staab. F–a model of events based on the foundational ontology DOLCE + DNS ultralight. In *Proceedings of the fifth international conference on Knowledge capture*, pages 137–144. ACM, 2009.

[285] Nikolaos Gkalelis, Vasileios Mezaris, and Ioannis Kompatsiaris. High-level event detection in video exploiting discriminant concepts. In *Content-Based Multimedia Indexing (CBMI), 2011 9th International Workshop on*, pages 85–90. IEEE, 2011.

[286] Pinaki Sinha and Ramesh Jain. Extractive summarization of personal photos from life events. In *Multimedia and Expo (ICME), 2011 IEEE International Conference on*, pages 1–6. IEEE, 2011.

[287] Cheng-Hung Li, Chih-Yi Chiu, Chun-Rong Huang, Chu-Song Chen, and Lee-Feng Chien. Image content clustering and summarization for photo collections. In *Multimedia and Expo, 2006 IEEE International Conference on*, pages 1033–1036. IEEE, 2006.

[288] Pinaki Sinha, Hamed Pirsiavash, and Ramesh Jain. Personal photo album summarization. In *Proceedings of the 17th ACM international conference on Multimedia*, pages 1131–1132. ACM, 2009.

[289] Stefanie Nowak, Ronny Paduschek, and Uwe Kühhirt. Photo summary: automated selection of representative photos from a digital collection. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 75. ACM, 2011.