THE UNIVERSITY OF HULL

Context-based Multimedia Semantics Modelling and Representation

being a Thesis submitted for the Degree of

PhD in Computer Science

in the University of Hull

by

Emmanuel Uchechukwu Eze, BSc Hons (UNN)

May 2013

Abstract

The evolution of the World Wide Web, increase in processing power, and more network bandwidth have contributed to the proliferation of digital multimedia data. Since multimedia data has become a critical resource in many organisations, there is an increasing need to gain efficient access to data, in order to share, extract knowledge, and ultimately use the knowledge to inform business decisions. Existing methods for multimedia semantic understanding are limited to the computable low-level features; which raises the question of how to identify and represent the high-level semantic knowledge in multimedia resources.

In order to bridge the semantic gap between multimedia low-level features and high-level human perception, this thesis seeks to identify the possible contextual dimensions in multimedia resources to help in semantic understanding and organisation. This thesis investigates the use of contextual knowledge to organise and represent the semantics of multimedia data aimed at efficient and effective multimedia content-based semantic retrieval.

A mixed methods research approach incorporating both Design Science Research and Formal Methods for investigation and evaluation was adopted. A critical review of current approaches for multimedia semantic retrieval was undertaken and various shortcomings identified. The objectives for a solution were defined which led to the design, development, and formalisation of a context-based model for multimedia semantic understanding and organisation. The model relies on the identification of different contextual dimensions in multimedia resources to aggregate meaning and

facilitate semantic representation, knowledge sharing and reuse. A prototype system for multimedia annotation, CONMAN was built to demonstrate aspects of the model and validate the research hypothesis, H_1.

Towards providing richer and clearer semantic representation of multimedia content, the original contributions of this thesis to Information Science include: (a) a novel framework and formalised model for organising and representing the semantics of heterogeneous visual data; and (b) a novel S-Space model that is aimed at visual information semantic organisation and discovery, and forms the foundations for automatic video semantic understanding.

Acknowledgement

My earnest gratitude goes to the Almighty for His inconceivable Grace and immeasurable Love during the course of this research work despite all the perceived difficulties.

I would like to thank the Department of Computer Science and the members of Internet Computing research group for believing in me and offering me the opportunity to embark on this long but highly invigorating and rewarding journey. I owe some gratitude to my initial supervisor, Dr Weihong Haung who provided the initial guidance before he left the University.

I would like to especially thank my supervisor Dr Tanko Ishaya for his excellent guidance and feedback which has not only helped to conclude and present the research, but has helped to shape me as an individual. I remain grateful to Paul Warren for helping to proof-read this thesis. Thanks to my examiners, Dr Darren Mundy and Dr Paul Clough for their insightful perspective to my work which has greatly improved the thesis presentation.

My appreciation goes to my wife, Lilian and children Ernest, Cassandra, and Alexander for showing great understanding and support throughout the course of this research, especially when I had to work day and night and deny them some attention.

Table of Contents

List of Figures

List of Tables

List of Symbols

Symbols	**Descriptions**
$=$	Equality (numbers, sets, etc.)
$\neq$	Inequality
$<$	Less than
$\leq$	Less than or equal
$>$	Greater than
$\geq$	Greater than or equal
$\{x, y, \ldots\}$	Set construction by enumeration
$\{x \mid\!— x —\}$	Set construction by abstraction
$\in$	Set membership
$x : S$	By definition $x \in S$
$\notin$	Negation of set membership
$\subseteq$	Set inclusion
$\subset$	Strict set inclusion
$\varnothing$	Empty set

Num, Card, | |	Cardinality of a set (number of members)
{x, y}	Pair (unordered) of x and y
$\cup$	Union of sets
$\cap$	Intersection of sets
+	Sum (Union of two disjoint sets)
$\sum$	Sum
–	Set difference
Pow(S)	Powerset (set of subsets of S)
$(a_1, a_2, \ldots, a_{n-1}, a_n)$	n-tuple
$\prod$	Cartesian product (n-ary)
$R : A \leftrightarrow B$	Introducing relation R with domain A and codomain B
$F : A \rightarrow B$	Introducing total function F from A to B
$F(x)$	Application of function F to x
$F(a_1, a_2, \ldots, a_n)$	Application of F to $(a_1, a_2, \ldots, a_n)$
OPP	Opposite of a relation
$\wedge$	Conjunction (and)
$\vee$	Disjunction (or)

$\Rightarrow$	Conditional (if . . . then)
$\Leftrightarrow$	Equivalence or biconditional (iff)
$\forall$	Universal quantifier (for all)
$\exists$	Existential quantifier (there exists)
$'$	Prime
$\therefore$	Introduction of a fixed feature
$\subseteq_d$, $=_d$, etc.	*By definition* subset of, equal to, etc.

Chapter 1

Introduction

Most of us would have experienced searching for one thing or another in our homes that we knew existed and yet we were unable to find them when we needed them. It could be a piece of clothing, a piece of jewellery, or a paper-based document. Sometimes finding such items can be a daunting experience that stretches many a person into searching the nook and cranny of their homes and often leaves them with a memorable impression. In this digital age where computers and mobile devices are prevalent, the landscape for information storage and subsequently searching and retrieval has changed. The massive growth of the Internet, the availability of digital multimedia capture devices, and the emerging ubiquitous and pervasive computing have resulted in an increase in multimedia[1] content (Lew *et al*., 2006; Enser, 2008a; Duygulu and Bastan, 2011).

According to The World Bank (2012), the number of mobile phone subscriptions has grown from less than 1 billion in year 2000 to 6 billion in 2012. This figure represents 75 percent of the world's population. With mobile devices becoming more powerful and cheaper, the mobile phones have been transformed from simple voice communications devices to powerful multimedia communications devices (Razikin *et al*., 2011; Hada, 2012; The World Bank, 2012). They are capable of audio-visual capture and playback, multimedia messaging, audio and video streaming, web browsing, video e-Learning, video telephony, and more. The proliferation of these digital multimedia content (such as audio, video, and images), require tools for extracting useful knowledge from the content to enable intelligent and efficient

1 The term *multimedia* generally refers to a complex digital information object, with various components such as text, audio, image, and video.

multimedia organisation, filtering, and retrieval (Moreno *et al.*, 2002; Liu *et al.*, 2007; Enser, 2008a; Pino and Di Salvo, 2011). Multimedia retrieval technologies have been used by experts within specific domains like Bioinformatics, Broadcasting and Video Surveillance. However, with the success of the World Wide Web (WWW) (from now on referred to as the web) and the increased multimedia content generation due to the proliferation of computers, mobile devices, and social media sites; multimedia retrieval technologies are now being required by users of various backgrounds and social context. Recent research efforts have focused on providing domain independent frameworks and tools that facilitates effective and efficient multimedia retrieval on the web (Hunter, 2005; Eze and Ishaya, 2007; Dasiopoulou *et al.*, 2011; Gennaro *et al.*, 2011; Sadallah, Aubert, and Prié, 2011; Fauzi and Belkhatir, 2013). However, the starting point for a good information retrieval system is to represent the content of the document such that efficient search algorithms can be applied on it towards efficient and accurate information retrieval (Goker and Davies, 2009). This process is referred to as *indexing*.

While automated indexing techniques have been successfully applied to textual documents on the web, multimedia semantic indexing involves different, still less mature and less scalable technologies (Ruger, 2011). Enser (2008a), Inoue (2009), and Dasiopoulou *et al.* (2011) suggest that more research effort is required in the indexing and retrieval of multimedia. Internet search engines do not return optimal results for multimedia related searches. The popular saying that "a picture is worth a thousand words" portrays why search results for image and video files are far from optimal (Quelhas *et al*., 2007). The words used in describing the same image for example, usually varies from person to person due their individual background and perception. This is evident by the social tagging of images or videos and often varying comments from other individuals on the same image or video on social media sites like Facebook[2] and Youtube[3].

2 http://www.facebook.com
3 http://www.youtube.com

A true image or video search engine therefore, will need to accommodate all possible interpretations for the image and video files. Tjondronegoro and Spink (2008) in their investigation of multimedia search in 102 search engines found that there are a few search engines offering multimedia search and those few provide limited keyword-based multimedia search functionality based on comments or tags describing the multimedia content. Although these search portals are evolving towards semantic search, inaccurate search result is a major drawback associated with multimedia search on these portals (Tjondronegoro and Spink, 2008). The nature of multimedia content as against traditional text poses a number of challenges including data and knowledge representation, indexing and retrieval, integration, intelligent searching techniques, information browsing, and query processing. Current technology is focusing on extracting and analysing multimedia features, semantics and knowledge (Flinker *et al*., 1995; Eberman *et al.*, 1999; Zhai *et al.*, 2005; Duygulu and Bastan, 2011; Wu *et al.*, 2012). Being able to achieve multimedia content analysis at the semantic level allows for easy indexing, filtering, searching, summarisation, and ultimately the extraction of knowledge (Mylonas *et al.*, 2008; Tjondronegoro and Spink, 2008; Wu *et al.*, 2012).

There is a gradual move away from the era of mere indexing and retrieval to the new era of knowledge extraction. In order to make this transition to multimedia knowledge, research by García and Celma, (2005) and Moutselakis and Karakos, (2009) have identified the need to analyse and transform multimedia content into metadata[4]. Metadata is a representation of the multimedia semantic content and is easier to manipulate using standard information retrieval methods. Multimedia knowledge management systems are necessary for a wide variety of commercial, governmental, and personal applications. These include telemedicine, distance learning, video surveillance, video-on-demand, digital libraries, and many more.

4 A set of data that describes and gives information about other data.

Video media is multimodal, and authors have the capacity to express semantic ideas using at least two or more information channels, where the channels can either be visual, auditory or textual (Snoek and Worring, 2005). This thesis focuses on video media as an instance of multimedia object in exploring and addressing the general problem of semantic video information retrieval.

1.1 Background

According to Denning (1998),

> *"the idea that knowledge should be shared is obviously not new. The pursuit of any significant human activity, typically leads to the acquisition by those involved of know-how and expertise as to how the activity may be successfully conducted. Insofar as what is learned in the process can be captured, and communicated and shared with others, it can enable subsequent practitioners or even generations to build on earlier experience and obviate the need of costly rework or of learning by making the same repetitive mistakes".*

Bock (2005) agrees with Denning (1998) and further encourages the notion of knowledge sharing. Information or knowledge has been handed down from one generation to another by words of mouth, ancient writings, and other forms of communication. Things began to take a new turn with the advent of information technology. Database management systems (DBMS) enabled the storage, modification and retrieval of information from a database. The information in question were basically from textual sources, though recent trends in DBMS have provided support for heterogeneous data management such as Character Large Object (CLOB), Binary Large Object (BLOB), and so on. Query languages like Structured Query Language (SQL), are used for data retrieval from these DBMS. Most information retrieval (IR) systems are based on database management systems. The major difference is that while

a DBMS relies on a strictly defined and standardised query language, which acts on well structured data, information retrieval systems are based on natural language text which is not well structured and can be semantically ambiguous. An information retrieval system is a piece software that stores and manages information on documents (textual and multimedia) such that users of the system are able to find information that they need. Information retrieval search results are therefore ranked (Aly *et al.*, 2012) according to the degree of relevance to the user query. In order to achieve this ranking, semantic information should be extracted from the document sources which is in turn matched against the user query. The problem is not only about knowing how to extract this information but also about knowing how to use it to decide relevance (Baeza-Yates and Ribeiro-Neto, 1999; Ren and Bracewell, 2009; Wu *et al.*, 2012). An information retrieval system often supports three basic processes: *indexing*, the representation of the content of the document; *query formulation*, the representation of the user's information need; *matching*, the comparison of the two representations.

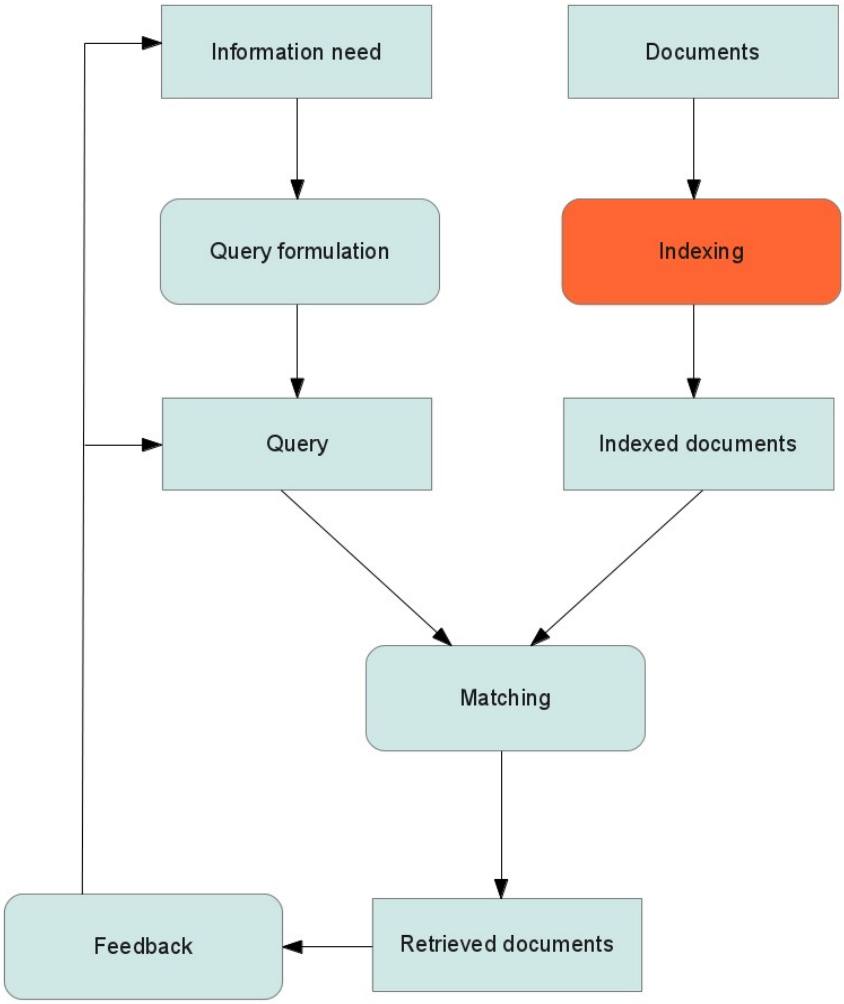

Figure 1.1: Information retrieval process (Hiemstra, 2009)

Figure 1.1 presents a typical information retrieval process with the squared boxes representing data while the rounded boxes represent processes. The main focus and contribution of this thesis is at the indexing stage of the information retrieval process.

The advent and continuous expansion of the web has revolutionised the way in which information is gathered, stored, processed, presented, shared, and used (Enser, 2008a). It is difficult to imagine the web without search engines or information retrieval technologies. Huge amounts of information are added to the web on a daily basis and web search portals like Google, are making efforts to cope with the challenge of

discovering and indexing these contents to enable users to retrieve them. Users of these portals usually consider their activity as retrieving information despite irrelevant search results, disappearing web pages, and broken links (Hider, 2006). The term "Semantic Web", was coined and defined by Berners-Lee *et al.* (2001) as "*a web of data that can be processed directly and indirectly by machines*". Tim Berners-Lee is Director of the World Wide Web Consortium (W3C), which oversees the development of proposed Semantic Web standards and they are actively working on evolving a new web which according to the report by World Wide Web Consortium (2011) will "provide a common framework that allows data to be shared and reused across application, enterprise, and community boundaries". This “new web” is expected to facilitate semantic query and replace the current web of “unstructured” documents (Gennaro *et al.*, 2011).

Visual information (images and videos) is rich in content. A picture for example, may arouse different reactions from different users, at different times and circumstances. Traditional information retrieval systems only deal with textual, unstructured data but the evolution of multimedia databases and the proliferation of multimedia content on the web have added a new challenge to information retrieval. Classical IR systems are therefore not suitable to support the heterogeneous data typical of a multimedia IR system. Text information retrieval is already well established; most data retrieval systems, such as web search engines, are text retrieval systems. However, multimedia information retrieval is less established. The overall effort in any IR system is to ensure accurate and efficient information retrieval but multimedia IR systems are seriously handicapped in the context of efficient and accurate retrieval due to their complex nature (Duygulu and Bastan, 2011). Firstly, most multimedia IR systems lack efficient and automated annotation. Secondly, accurate retrieval results are not guaranteed due to the inability of these systems to appropriately represent multimedia content in a human perceptible fashion. There are a number of open issues involved in such retrieval. Wang and Hua (2011), identified the inefficiencies associated with manual multimedia

annotation and proposed an active learning approach towards improving multimedia annotation and retrieval. Dasiopoulou *et al*. (2011) and Gennaro *et al.* (2011), agrees and identified the issue of poor metadata organisation, poor annotation at the semantic level, and lack of interoperability with other systems.

There are two approaches generally employed in the searching and retrieval of visual information: content-based visual retrieval and Text-based (concept-based) visual retrieval (Natsev *et al.*, 2007; Datta *et al.*, 2008; Petrelli and Auld, 2008; Snoek and Worring, 2009). Content-based visual information retrieval depends on the automatic extraction of computable low-level feature vectors. These feature vectors may be colour histogram, texture statistics, shape, or any other information that can be derived from the visual data itself. Retrieval is based on presenting images or videos similar to what is being searched. This approach has the shortcoming of users not being able to pose appropriate queries that represents their search goals as they often have a sketchy idea of their information need. Also, there are abstract concepts or emotions (e.g. "London City", "happy") that cannot be matched to any set of visual information (Enser *et al.*, 2007; Natsev *et al.*, 2007; Mylonas *et al*., 2008; Snoek and Worring, 2009; Duygulu and Bastan, 2011). Text-based visual retrieval relies on the use of keywords or text to annotate multimedia content. The visual objects and their relationship are captured as high-level semantic description and assigned symbolic labels. The annotation may be manual or automated.

A review of the different techniques for visual information retrieval is presented in chapter 2. The main focus of this thesis is on semantic-based information retrieval which uses text to index and retrieve multimedia objects. Other forms of multimedia information retrieval will be described to provide a perspective of the range of systems.

1.2 Problem Statement

Multimedia low-level features (e.g. colour histogram, texture statistics, shape) are easily measured and computed, but users typically start their information retrieval process by formulating high level semantic query (Hare *et al.*, 2006a; Hider, 2006; Enser, 2008a; Chung and Yoon, 2010; Fauzi and Belkhatir, 2013). Humans think in terms of objects in the media and how they interrelate, and therefore naturally express their information need as such (Hider, 2006; Lew *et. al.*, 2006). This gap between the low-level features and the high-level concepts is generally referred to in the literature as the "semantic gap". Smeulders *et al.* (2000) defines the semantic gap as the "lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation". A major challenge is being able to represent and index the videos at the right level of description and ensure such level matches the user's interest level (Enser, 2008a).

Video annotation can be manual or automated. Manual video annotation is usually done by domain experts who provide textual description of the video content. This can be very time-consuming and subject to errors due to human bias or mistake (Kender and Naphade, 2005; Duygulu and Bastan, 2011; Fauzi and Belkhatir, 2013). The goal of automatic annotation is to predict the descriptive keywords or text for visual information using prior knowledge (Hare *et al*., 2006b; Eze and Ishaya, 2007; Dasiopoulou *et al*., 2011; Hu *et al.*, 2011; Wang and Hua, 2011). While automated video annotation is desirable, it is important that the generated metadata truly represents the video content in its entirety. Intelligent systems are needed that could take low-level feature representation of the visual media in order to provide a high-level knowledge-based semantic representation (Hare *et al.*, 2006b; Hider, 2006; Zhang, 2007; Smeaton, Over, and Kraaij, 2009; Dalakleidi *et al.*, 2011) of the media content. There is need for efficient, accurate, automated video annotation with minimal human intervention (Bloehdorn *et al.,* 2005).

Context has played a key role in ubiquitous computing and recently multimedia semantic retrieval (Ingwersen and Jarvelin, 2005; Borlund *et al*., 2008; Goker, Myrhaug, and Bierig, 2009; Wiliem, Madasu, Boles, and Yarlagadda, 2012; Yi, Peng, and Xiao, 2012). Although the notion of context holds different interpretations with regards to theoretical principles and design practice, they still share some things in common to both application domains. Nunes, Santoro, and Borges (2009) posits that an implementation of context information management in a computational environment, may act as a filter that defines, at any given moment, which piece of knowledge will be taken into consideration in helping perform tasks. In this thesis, c*ontext* is regarded as any additional information that enhances the semantic understanding of a video. The background to context is provided in section 2.4 and a formal definition in relation to this thesis is given in section 3.4.1.

From the survey of visual information retrieval techniques and the taxonomy presented in Chapter 2, it was observed that organising and retrieving multimedia content at the knowledge level is a pertinent research issue. Humans have diverse information interest in the same video content based on the different context and levels of their perceptual ability. Users are more interested in semantic entities rather than visual appearance. Thus, they naturally express their information need at a highly semantic knowledge level. This work focuses on video content representation at the knowledge level, in order to allow for semantic video indexing and subsequent retrieval based on users' information need. The research is therefore driven to investigate the following questions:

1. *In what form is contextual information expressed in web video data?*

2. *How can context be modelled to represent video semantics from web video data?*

3. *How can context be applied in the automatic semantic description of visual features in web video data?*

4. *Can context information improve the representation of semantic knowledge in web video data?*

1.3 Research Hypothesis, Aims, and Objectives

The description and searching of visual data are closely related processes. Without accurate description, there cannot be accurate retrieval. In recognition that visual data can mean different things to the same person at different times or circumstances (Spink and Cole, 2005), this work seeks to investigate the application of *context* to bridging the semantic gap which presents itself as a computational problem in visual information retrieval. An important goal of this research is to create models and framework that facilitate the development of tools for context-based semantic visual information understanding and representation.

The hypotheses underlying this research are as follows:

H_1 - "***Semantic description of video from the web will improve with an increase in contextual knowledge than with less contextual knowledge.***"

H_2 - "***It is possible to develop a generic formalised model for video semantic management and representation.***"

H_3 - "***It is possible to develop a model for automated video semantic understanding.***"

Towards addressing the research challenge, this work aims at bridging the semantic gap between user perception and multimedia content description using contextual knowledge. This helps to facilitate effective multimedia semantic understanding, representation, and knowledge sharing.

In order to validate the research hypothesis above, the following research objectives are defined:

- to provide a review of the state of the art in Multimedia Information Retrieval;
- to provide a context-based framework for web-based video semantic representation;
- to provide a model for video semantic understanding on the web.

The methodology for achieving the outlined objectives is presented in section 1.4.

1.4 Research Methodology

"*... the scientist builds in order to study; the engineer studies in order to build.*"

- Brooks (1996).

Research in Computer Science encompasses a wide variety of different research methodologies to address research questions posed. Some of the prominent research approaches in the literature (Brooks, 1996; Benbasat and Zmud, 1999; Dodig-Crnkovic, 2002; Baskerville and Myers, 2004; Ramesh, Glass, and Vessey, 2004; Jones and Gregor, 2007; Iivari, 2010; Gregor and Hevner, 2011) that are widely applied in Computer Science Research include: Action Research, Constructive Research, Design Science Research, and Formal Methods.

Action research aims at solving practical problems and producing concrete results. The central focus of action research which is to diagnose and solve immediate problems does not allow it to produce results that can be applied outside the context of a particular research project (Kock, McQueen, and Scott, 1997). However, it aims to balance problem-solving actions that are produced, with research efforts to understand fundamental causes which often helps in future predictions and scientific knowledge expansion (Baskerville and Myers 2004; Dodig-Crnkovic, 2010). Kock, McQueen, and Scott (1997) argues that action research is lacking in scientific rigour and generalisation of produced artefacts or principles. Constructive research aims at producing novel constructs that solve both practically and theoretically relevant problems with main focus on creating new solutions rather than researching what already exists (Caplinskas and Vasilecas, 2004). Constructive research is often regarded as the most predominantly used research methodology in Computer Science (Ramesh, Glass, and Vessey, 2004) since it mostly focus on innovative solutions or improvements to research problems. Constructs or artefacts produced in a constructive research process are usually validated by comparing them to existing solutions to demonstrate their relevance and novelty (Dodig-Crnkovic, 2010).

Design Science research (DSR) is somewhat similar to constructive research as its main aim is to develop relevant innovative artefacts in a rigorous fashion that contributes to the development of new knowledge on a scientific level to the application domain (Ulrich, 2006; Gregor and Hevner, 2011; Kuechler and Vaishnavi, 2011). DSR efforts usually target to produce an artefact which is a partial solution to an identified research problem and additionally produce a “design theory” that prescribes the requirements for a class of artefacts to address similar problems (Jones and Gregor, 2007). Design Science Research pays attention to the design and development process through which the research artefacts evolves in order to generalise. Formal research methods are

mathematically-based techniques that can be applied throughout the development and specification of a system to precisely and rigorously describe a system. The specification of the system involves validation and verification at each stage to ensure the completeness, correctness, and consistency of specification (Scheurer, 1994; Kaur, Gulati, and Singh, 2012). Formal methods are usually applied in the development of critical systems or novel models (Kuhn, Chandramouli, and Butler, 2002; Sommerville, 2010) to ensure rigorous specification, correctness, and validity. In Information Science research, formal specification languages (e.g. Z, B, VDM, and Feature Notations) which are based on set theory and logic, are widely applied.

While action research focus mainly on practical problems and concrete results, constructive research aim more at producing research constructs that address both theoretical and practical problems. However, while the research constructs could be a complete or partial practical solution to a problem, it tends to focus on the research problem at hand and oftentimes found wanting in generalisable theory or principles (Ulrich, 2006). Design Science research, unlike constructive and action research, does not always aim at providing concrete solutions to identified research problems. It rather aims at providing a rigorous methodology for producing novel research artefacts which can be building blocks towards solving both practical and theoretical Computer Science problems. DSR ensures that artefacts are abstracted and generalised (Dodig-Crnkovic, 2002) such that they constitute a new scientific knowledge contribution.

In accordance with the statement by Brooks (1996) quoted at the beginning of the section, this research adopts a mixed methods approach incorporating both Design Science Research and formal methods for investigation and evaluation. According to Ulrich (2006), design artefacts or models should emphasise a certain level of abstraction through a rigorous specification which can possibly prove the adequacy of

the artefact and also qualifies it as a scientific knowledge contribution. The motivation for using formal methods stems primarily from their sound mathematical basis and the means of proving that the specification is realisable, precise, consistent, and complete. An exploratory research was initially undertaken by reviewing the literature on multimedia information retrieval (refer to Chapter 2) to gain insight into the open issues on multimedia information retrieval. It was identified from the literature review that there exists a semantic gap which presents itself as a computational problem in visual information retrieval (Smeulders *et al*., 2000; Hare *et al.,* 2006a; Enser, 2008a). The representations that can be computed from raw image data cannot be readily transformed to high-level representations of the semantics that the visual data convey and in which users typically prefer to articulate their queries (Hider, 2006; Enser, 2008a; Fauzi and Belkhatir, 2013).

A review of current approaches for multimedia semantic retrieval was carried out and various shortcomings identified (refer to sections 2.5 and 3.3). The need for a solution was defined which led to proposed *context* approach. The context approach for multimedia semantic understanding and organisation (see Chapter 3) was further designed and developed into a model. The model relies on the identification of different contextual dimensions in multimedia resources to aggregate meaning and facilitate clear semantic representation, knowledge sharing and reuse. Kuhn, Chandramouli, and Butler, (2002) have argued that formal methods are usually a practical means of demonstrating an essential property of critical systems. The context model was abstracted and generalised using formal methods which Sommerville (2010), identified as helping to improve the researchers' understanding of the specification and exposes errors and inconsistencies. The exposed errors and inconsistencies forced the researcher to go through many iterations of specifying and validating the artefact until it became correct, robust, unambiguous, and complete. The S-Space model in chapter 4, which is

an extension of the framework (see Chapter 3) was developed as a result of the need to have support for automatic semantic extraction.

A prototype system for multimedia annotation, CONMAN was built (refer to Chapter 5) to demonstrate aspects of the model and experimentation carried out (see Chapter 6) to validate the research hypothesis, H_1. The originality of this work lies with using context information to represent and manage the semantics of multimedia resources. It has been identified from the literature that context plays a crucial role in human knowledge representation, reasoning, and perception. Therefore, multimedia information retrieval systems need the ability to represent, utilise and reason about context to help improve semantic representation and management of multimedia resources. The model identifies and uses contextual information about the multimedia resources to enhance automatic semantic understanding of such multimedia content. The multimedia content description is subsequently represented as metadata in the form of Extensible Markup Language (XML) or Resource Description Framework (RDF). Capturing and representing all contextual dimensions in different application domains has been identified as an overwhelming task. Thus, the developed prototype uses football video clips as a test application domain. As an important video domain, football clip was chosen due to its intrinsic multimodal nature since its creator uses visual, auditory, and textual channels to convey meaning. Most explanations or examples throughout this thesis are mainly taken from the football domain. A video clip outside the football domain was chosen for the evaluation (refer to Chapter 6) to further buttress the universality of the developed context model.

The models and framework developed in this thesis are further published in peer-reviewed conferences and journals (refer to Appendix C) to further demonstrate their validity, contribution, and relevance to the research community.

1.5 Thesis Contribution

The aim of this research project has been to investigate the application of *context* to bridging the semantic gap which presents itself as a computational problem in visual information retrieval, through creating a framework and models that facilitate the development of tools for context-based semantic multimedia retrieval. The main contributions made include: (a) a novel framework and formalised model for organising and representing the semantics of heterogeneous multimedia data; and (b) a novel Semantic-Space (S-Space) model that is aimed at visual information semantic organisation and discovery, and forms the foundations for automatic multimedia semantic understanding.

The key difference between this work and other existing research efforts in the research area of Multimedia Information Retrieval lies with the context approach of organising and representing high-level multimedia semantics.

1.6 Thesis Organisation

The rationale for the organisation of this thesis is presented in Figure 1.1. The first part which comprises chapters 1 and 2 deals with the formulation of the research problem. Chapter 1 presents the context of the entire thesis, while Chapter 2 focuses on identifying distinct research problems and examining possible approaches towards solving the identified problems. This was achieved through a critical review of the state of the art in semantic multimedia indexing and retrieval. An analysis of the review is presented and precise research questions formulated.

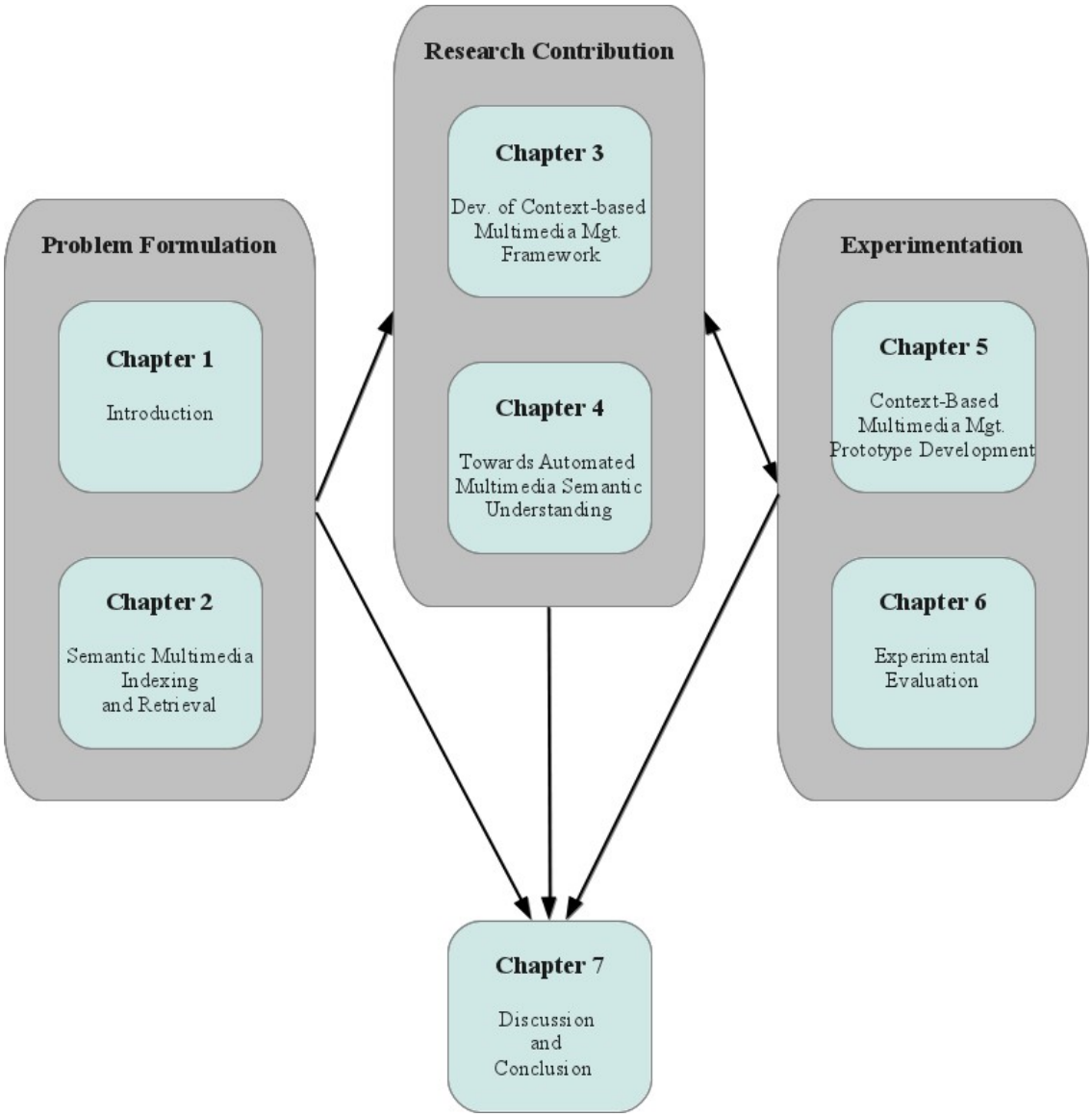

Figure 1.2: Thesis organisation

The first part of the thesis forms the foundation for the second part which is the core research contribution and comprises chapters 3 and 4. Chapter 3 presents the development of the conceptual framework for context-based multimedia semantics organisation and representation; and formalises it into a cross-domain generic model. Chapter 4 presents the development and formalisation of the S-Space model, which implements an important aspect of the context-model (presented in chapter 3). The S-Space model aims to facilitate automatic multimedia semantic understanding.

Part three centres on providing the basis for experimental evaluation of aspects of the core contributions in part two and comprises chapters 5 and 6. Chapter 5 presents the design and development of a prototype that implements aspects of the context model developed in chapter 3, while chapter 6 presents an experimental evaluation using the developed prototype to validate the research hypothesis, H_1. Part four of the thesis is the concluding part and constitute of only chapter 7, which provides a summary of the thesis by drawing from all the parts of the thesis and provides direction for future work.

Chapter 2

Semantic Multimedia Indexing and Retrieval

Information retrieval deals with the representation, storage, organisation of, and access to information items (Baeza-Yates and Ribeiro-Neto, 1999). Information retrieval (IR) usually starts off on one hand by a user with an information need. This need, which may initially be unprecise, somehow needs to be articulated and formulated into a request which embodies the original intent. The request is then converted into a search statement or query. Conversely, information is stored in collections or repositories. Stored information has the potential of being a valuable resource if found, but in order to be found, information need to be represented somehow and then subsequently indexed. An index is an optimised data structure that is built on top of the information objects to speed up searches. The challenge, according to Goker and Davies (2009) is to provide a good match between the query and the stored information in order to ensure that the information presented during the search is of relevance to the user with the original query.

Eidenberger (2011) defines Multimedia Information Retrieval (MIR) as a research discipline of Computer Science that aims at extracting semantic information from multimedia data sources. A broader definition is given by Pino and Di Salvo (2011), where they refer to MIR as a set of theories, algorithms, and systems that aim at extracting multimedia content related to pertinent descriptors or metadata, thus supporting advanced search functions. The later definition by Pino and Di Salvo (2011) is adopted in this thesis. Multimedia Information Retrieval is a multidisciplinary research area that cuts across different fields of Computer Science including digital

signal processing, multimedia applications, information retrieval, database management, artificial intelligence, and so on. Digital signal processing researchers focus on modelling and representation of multimedia content; and object recognition and matching in multimedia retrieval systems (Visser, Sebe, and Bakker, 2002; Snoek *et al.*, 2006; Lavee, Rivlin, and Rudzsky, 2009). Database researchers investigate efficient ways of storing and managing multimedia content in databases (Cosmin, 2010). Researchers in the artificial intelligence discipline approach the MIR issue by finding intelligent methods of representing and modelling human information needs (Lew *et al.*, 2006).

MIR has been the subject of active research in both industry and academia across the world. This reflects the importance of such research and technology, and the fact that there are still many open research issues associated with MIR. This chapter provides a survey of research in visual information retrieval and presents the research challenge which this research work seeks to address. The objective is to present a perspective rather than an exhaustive summary of approaches relevant to the emerging multimedia semantic retrieval. This chapter is structured around the two key approaches generally employed in the searching and retrieval of multimedia information: content-based and text(concept)-based multimedia retrieval (Lew *et al.*, 2006; Natsev *et al.*, 2007; Datta *et al*., 2008; Petrelli and Auld, 2008). Content-based multimedia information retrieval depends on the extraction and matching of computable low-level feature vectors. Text(Concept)-based multimedia retrieval depends on text to provide a high-level semantic description of multimedia content.

The chapter is organised into six main sections. Section 2.1 presents a review of techniques and existing systems in content-based multimedia information retrieval; section 2.2 presents a review of concept-based multimedia retrieval approaches in relation to multimedia semantic understanding and representation; section 2.3 presents a classification of multimedia information retrieval systems; section 2.4 investigates the notion of context as it impacts on information retrieval with emphasis on multimedia

information retrieval; section 2.5 provides an analysis of the research issues to justify the research problem; while section 2.6 concludes the chapter by presenting a summary.

2.1 Content-based Multimedia Information Retrieval

Content-based visual information retrieval (CBVIR) is an instance of multimedia information retrieval that is concerned with the application of computer vision techniques to the visual retrieval problem by relying on the multimedia features (e.g. colour, shape, texture) for indexing and searching multimedia databases (Smeulders *et al.*, 2000; Lew *et al.*, 2006; Natsev *et al.*, 2007; Datta *et al.*, 2008; Patel and Meshram, 2012). Content-based multimedia information retrieval, depends on searching for multimedia content by analysing the actual features of the visual data like colour histogram, texture statistics, shape, or any other information that can be derived from the multimedia data itself. CBVIR systems are predominantly used in application areas like surveillance systems or medical images where sample image queries are available and exact or near matches are desired based on the image/video features. A typical architecture (Juan and Cuiying, 2010) of a visual information retrieval system is presented in Figure 2.1.

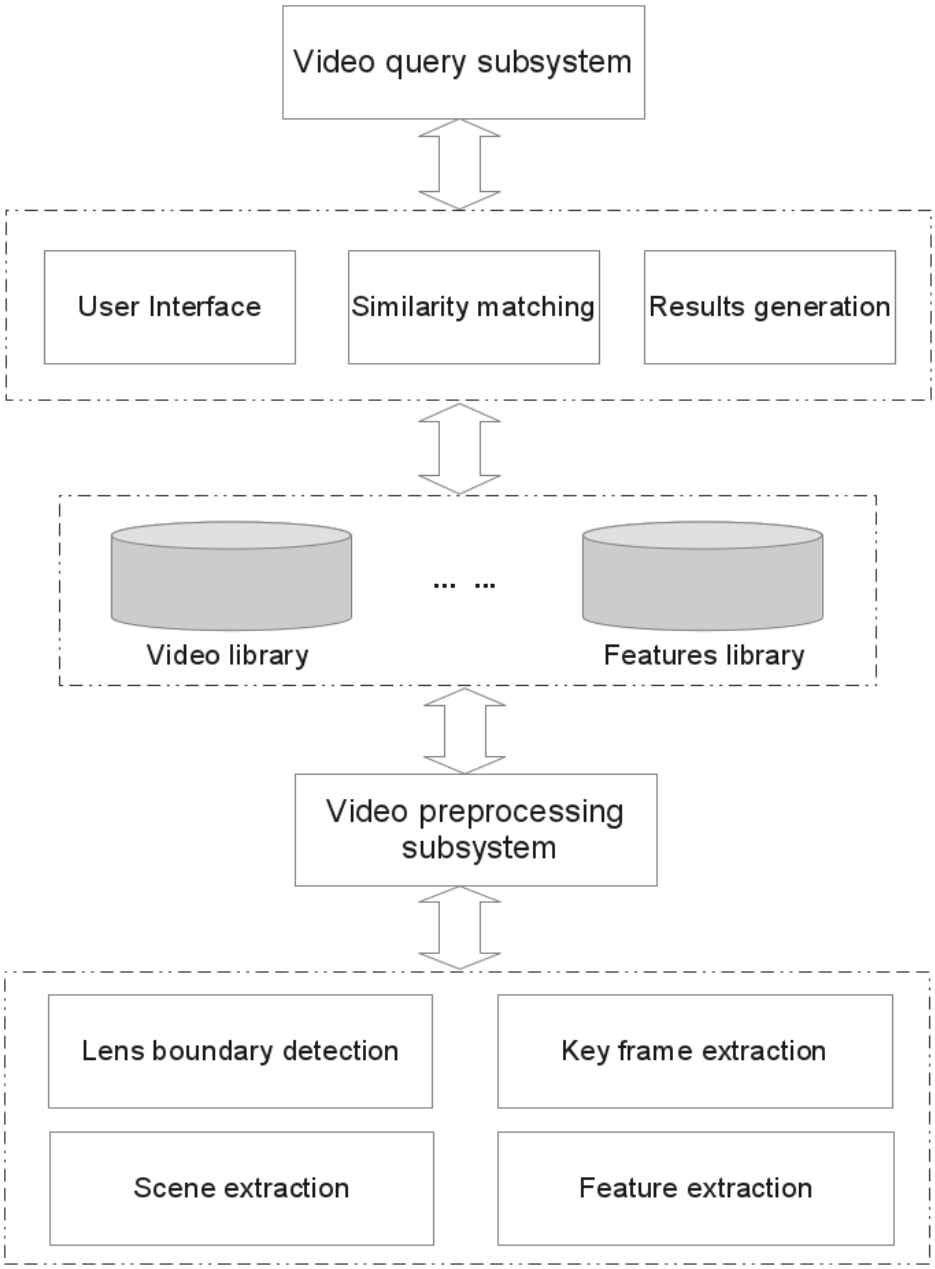

Figure 2.1: Typical video retrieval system (Juan and Cuiying, 2010)

Content-based visual information retrieval is a two-phase process comprising the content representation phase and the retrieval phase (Ren and Bracewell, 2009; Patel and Meshram, 2012). The content representation phase provides some low-level

processing to extract the features of the visual data and subsequently index it in readiness for searching. The retrieval phase interfaces between the user and the index, while displaying the visual content that meets various users' search requirement. These two phases of content-based visual information retrieval are discussed in sections 2.1.1 and 2.1.2 respectively.

2.1.1 Multimedia content representation

Multimedia content representation involves the extraction of multimedia content including low-level visual features from images, shots and scenes from videos, and subsequent indexing. Moving Pictures Expert Group (MPEG) developed MPEG-1 and MPEG-2, which handles multimedia content as signals, and allows for efficient storage, compression, and communication. MPEG-4 further improves coding efficiency by extracting and analysing features of the multimedia signals. MPEG-7 provide tools for the semantic description of multimedia content (Koenen R. (*ed.*), 1999; Martinez, 2003).

Content representation is usually the first of the two major phases in video retrieval systems. It generally involves the partitioning of the video stream into some basic units called shots. Each of these shots is summarised by the selection of key frames, which represents the salient characteristics of that shot. Low-level spatial and temporal features are extracted from these key frames as metadata, which is then indexed in databases. Content representation will be discussed under the following headings, based on the three key steps involved: video segmentation, key frames selection, and indexing.

2.1.1.1 Video segmentation

The primary approach in structuring video information consists of finding its basic units through temporal segmentation in order to represent it as metadata. Video

segmentation is the process of determining the boundaries between consecutive camera shots. A shot refers to a contiguous recording of video frames depicting a continuous action in time and space. Generally, there are two major trends in video segmentation techniques - uncompressed and compressed domains (Sarah De Bruyne et al., 2007).

Choi *et al*. (1997), Mech and Wollborn (1997), Moscheni *et al*. (1998), Wang (1998), Meier and Ngan (1999), Chi-Chun and Shuenn-Jyi (2001), Lo and Wang (2003), Apostoloff and Fitzgibbon (2006), and Yuan *et al.* (2007) proposed several methods of video segmentation in the uncompressed video domain. The histogram-based technique was proposed by Lo and Wang (2003) and Lefevre, Holler, and Vincent (2003). This technique depends on the fact that two frames that exhibit minor changes in the background and object content will also show insignificant variations in their intensity or colour distributions. The histogram-based technique is not very sensitive to camera operations and object motion though it is one of the most widely used shot detection techniques. Further research (Nagasaka and Tanaka, 1992; Zhang *et al.,* 1993; Chi-Chun and Shuenn-Jyi, 2001; Ferman, Tekalp, and Mehrotra*,* 2002; Küçüktunç, Güdükbay, and Ulusoy, 2010) proposed variants of the histogram-based technique to improve performance. In the pair-wise pixel comparison technique proposed by Zhang *et al.* (1993), a pixel is judged different if the difference between the intensity values in two successive frames exceeds a given threshold. The problem with this metric is that it is sensitive to camera motion, illumination changes, object movement and noise since both the feature representation and the similarity comparison are closely related. A different approach is the block-based technique (Chung *et al.*, 2005), which uses local characteristics to increase the robustness to object, and camera movements. Each frame is divided into a number of blocks that are compared against their counterparts in the successive frame. Typically, the similarity or dissimilarity between two frames can be measured by using a likelihood ratio, as proposed by Zhang *et al.* (1993) and Idris and Panchanathan (1997). A shot transition is identified if the number of changed blocks is above the given threshold. This approach has the advantage of providing a better

tolerance to slow and small motions between frames. However, cuts may not be properly detected between two frames that have similar pixel values, but different density functions. Other video segmentation techniques in the uncompressed domain include clustering-based and model-driven techniques.

Wang *et al.* (2003), Sarah De Bruyne *et al*. (2007), and Zhang, Khan, Robertson (2008) proposed methods for segmenting compressed video. The major reason for these methods is to improve efficiency. Some of these techniques as shown in recent research (Yazbek, Mokbel, and Chollet, 2007; Zhang, Khan, and Robertson, 2008) include motion vectors and Discrete Cosine Transform (DCT) coefficient techniques. The DCT coefficients in MPEG video carry information that could be directly used to detect cuts. The MPEG coding standard relies on three different kinds of frames: Intrapictures (I), predicted pictures (P), and Interpolated pictures (B – for bidirectional prediction). A frame of type P is predicted from the previous I frame, while a B frame is predicted and interpolated from its preceding and succeeding I/P frames. The residual error after motion compensation is then DCT encoded. If the residual error exceeds a given threshold for certain blocks, motion compensation prediction is abandoned and straight DCT coding is used. High residual error values are likely to occur in nearly all blocks across a camera shot boundary. In order to detect a camera break, it is sufficient to count the fraction of blocks for which no motion vectors have been computed. If this number exceeds a predefined threshold, a cut is declared. One key problem with the techniques in the compressed domain is that they lack reliability though efficient.

Gargi, Kasturi, and Strayer (2000), Koprinska and Carrato (2001), Lefevre, Holler, and Vincent (2003) , Zhang (2006), and Choroś and Gonet (2008) presented a detailed survey on video segmentation techniques. However, segmenting multimedia data to its basic unit is not the best way to represent multimedia content as it does not represent its semantic properties.

2.1.1.2 Key frames selection

Key frames help to summarise video shots. They basically represent the salient features in a video shot. The simplest way of key frame extraction is to use the *nth* frame of each shot as the shot's key frame. Nagasaka and Tanaka (1991) demonstrated the *nth* frame approach. The drawback of the approach is that it uses a pre-determined single frame that may be unstable and not represent the salient features of the shot properly. Ueda, Miyatake, and Yoshizawa (1991) proposed an approach where two key frames (the first and the last of each shot) are used to represent shots. Another basic approach is to choose several key frames separated by a fixed distance in a shot (Shahrary and Gibbon, 1995). These basic techniques are inadequate for dynamic shots since the visual content of a shot is not considered at all. Other techniques exist which are based on the visual content of the shot.

The cluster-based technique (Doulamis *et al.*, 1998; Zhuang, Rui, Huang, and Mehrotra, 1998; Chang, Sull, and Lee, 1999) depends on grouping the frames of a shot together based on the similarity of their visual content. Key frames are chosen from the representative frame of each cluster. Although the key frames extracted by the cluster-based methods are useful to understand the overall visual content of video, these key frames are not suitable for computing the similarity between two video segments since the temporal information (an important property of a video) of the video frames is not taken into account. The sequential key frame selection methods use both visual and temporal information of frames based on the criterion of reducing the temporal visual-content redundancy by representing several consecutive frames with one key frame. Lee and Kim (2003) proposed a sequential key frame selection method when the number of key frames is given as a constraint. It first selects the pre-determined number of initial key frames and initial time-intervals based on the temporal variation of visual content. Then, it adjusts the positions of key frames and time-intervals by iteration, which reduces the distortion gradually.

Some of the techniques discussed above still suffer the limitation of not properly representing the salient features of a shot. This stage of content representation is very critical as features would be extracted from the selected key frames and hence indexed. The selection of the wrong key frames here will render the entire indexing and retrieval inaccurate. There is the need for a more accurate and reliable key frame selection technique.

2.1.1.3 Indexing

Image or video indexing can be carried out at different levels of abstractions starting from indices (such as name and subject) to much low level aspects (such as motion properties) of video as proposed by Brunelli *et al.* (1996). Video indexing entails the extraction of semantic content from key frames (still images) and representing the frame in the form of metadata, which is then indexed for searching and retrieval. Metadata describe the content, quality, condition, and other characteristics of multimedia data. It is a compact representation that can be indexed for information retrieval.

There are different principles and theories of representing image semantics and achieving image indexing. A prominent model in the literature (Enser and Sandom, 2002; Rafferty and Hidderley, 2005; Hare *et al.*, 2006; Enser *et al.*, 2007) which aids in the image indexing process is that which is derived from the work of the Art Historian, Panofsky. Panofsky (1962) laid out his method as three levels of iconographic/iconological analysis. The first level is the primary or natural subject matter, which consists of perception of the work's basic form. The second level is the secondary or conventional subject matter (iconography), which analyses an image by relating the image to cultural and iconographic knowledge. Such iconological analysis leads to a decision on the meaning of the image. The third level is the tertiary or intrinsic meaning or content (iconology). This level looks at an image from a historical knowledge perspective towards an understanding of the image and demands high-level

semantic reasoning. In the 1980's, Information Scientist Shatford (1986) applied Panofsky's theory of three levels of meaning, to image indexing and extends it to general subject analysis of pictorial work. Only the first and second levels of Panofsky's theory are to be indexed, since the third seems exemplified by the content of a subjective critical review (Shatford, 1986). Shatford (1986), generalised Panofsky's three levels of meaning into Generic (pre-iconographic), Specific (iconographic) and Abstract (iconological) and extended the model further by breaking each of these three levels into four facets: Who, What, Where and When. Enser and Sandom (2002), applied this model in the retrieval of still and moving images. Jaimes and Chang (2000) also extended the work of Panofsky and others, and developed a ten-level pyramid model presented in Figure 2.2. The model allows for indexing different aspects of visual information based on syntax (e.g., colour, texture, etc.) and semantics (e.g., objects, events, etc.), and includes distinctions between visual and non-visual information.

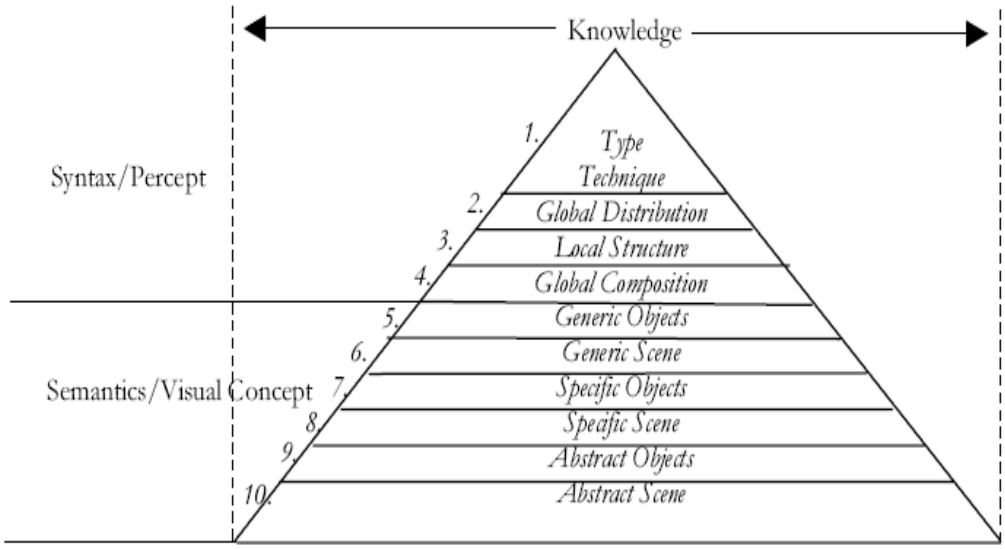

Figure 2.2: Pyramid of Indexing Structure (Jaimes and Chang, 2000)

A further review of the principles and theories of representing image semantics and achieving image indexing was presented by Enser (2008b). Apart from the theories and principles that guide the actual indexing, there are generally two major trends within the research community to extract indices for proper video indexing and annotation. The first trend proposes methods for automatically extracting these indices while the second approach proposes semi-automatic (with human intervention) methods to index the video. Most work in the automatic extraction realm are focused on deriving indices from visual elements (colour, shape etc.), camera motion (panning zooming, etc.), region/object motion in a video frame. Automatic property extraction techniques are slow, clumsy (Burrill, 1994), and performs poorly in the identification of high-level concepts. The semi-automatic indexing approach is advocated, as human intervention is very crucial for a proper semantic video indexing. A detailed review of different approaches for video indexing have been presented by Brunelli *et al*. (1996), Wang *et al.* (2003), Enser (2008b), Mylonas *et al.* (2008), and Hu *et al*. (2011).

A good indexing system would generally be up-to-date. The current web practice is to have crawlers visit sites at defined intervals in order to index the content. This is an issue, since web content providers will have to wait for a crawler to come around and pick up their new content before it can be "found" by people. This issue of "freshness" gets worse with the rapid increase in the content distributed on the web.

2.1.2 Retrieval phase

A major challenge with the image/video retrieval phase is the problem of good query formulation and the ability to match the queries to the media or indexed metadata stored in the database. One technique is Query by Visual Example (QBVE) (Flinker *et al*., 1995; Bimbo, 1999). QBVE is a method of query creation that allows the user to search for images by sending an entire image as a visual query. This technique uses image similarity matching to retrieve relevant result. The problem with this method is

that it is not common for users to have a sample of the image they are trying to retrieve. A slightly different approach is Query by Sketch (Bimbo, 1999). This approach is similar to the QBVE except that in this case users are allowed to draw a sketch of what they wish to retrieve using an interface and a set of drawing objects. A major drawback of the approach is that retrieval is based on how well the sketch was drawn, which basically hinders majority of Internet users from using MIR systems based on this approach. The query by sketch approach generally does not allow for retrieval of image content based on detailed semantics, as there is a limit to how much meaning users can provide with sketches. MIR systems still use the traditional approach of textual query which is matched against a pre-indexed metadata for retrieval. But the main MIR question still remains unanswered - how can the low-level features of the visual media be mapped to the high-level semantic object representation of the content?

Researchers adopt a combination of text and visual query as proposed in WebSeek[5] and VideoCue[6]. Another technique geared towards improving MIR search precision is the relevance feedback (Rui, Huang, Ortega, and Mehrotra, 1998; Su *et al.*, 2011). Users are allowed to assign a score to each of the returned hits and these scores are used to direct the following search phase and improve its result. The next section presents various model of information retrieval.

2.1.2.1 Information retrieval models

This section describe the different type of information retrieval models under three categories: Exact Match, Vector Space and Probabilistic models (Baeza-Yates and Ribeiro-Neto, 1999; Hiemstra, 2009). Consider that each document is described by a set of representative keywords called index terms. An index term is basically a (document) word whose semantics helps in remembering the document's main themes.

5 htpp://www.ctr.columbia.edu/webseek
6 htpp://www.ctr.columbia.edu/VideoQ

Exact Match Models

The exact match retrieval model provides exact matching for documents and ensure they are either retrieved if they match or not. There are two prominent models in the literature (Lashkari, Mahdavi, and Ghomi, 2009; Hiemstra, 2009) that provide exact matching – Boolean and Region models. The Boolean model is based on set theory and Boolean algebra. Queries are specified as Boolean expressions, which have precise semantics. The Boolean model predicts that each document is either relevant or non-relevant. There is no notion of a partial match. The main advantage is the simplicity and clear formalism. The drawback of this model is that it is not easy to translate an information need into a Boolean expression as most users find it very hard to express the queries as such. The model does not provide a ranking of retrieved documents (Hiemstra, 2009). Also, exact matching based on keyword indexing may lead to retrieval of too few or too many documents.

Region models (Jaakkola and Kilpelainen, 1999; Hiemstra, 2009) are similar to Boolean model but the indexed data on which queries are performed, unlike in Boolean model, are separated into segments or regions. The region model attempts to extend the Boolean model beyond the indexed documents and also allow search operations on the actual content. In addition to the Boolean operators, more operators like CONTAINING, CONTAINED_BY, FOLLOWED_BY, etc. are introduced (Hiemstra, 2009) to facilitate improved retrieval within regions. However, this model being an exact match model still suffers from a lack of provision of ranking for retrieved data.

Vector Space models

The vector space model represents each document in a collection as a point in a space generally referred to as a vector within a vector space. The semantic similarity of these

points are a function of the distance between them. Closer points are semantically more similar than distant points (Salton, Wong, and Yang, 1975). The vector space model supports the notion of partial matching by assigning non-binary weights to index terms in queries and documents (Salton, Wong, and Yang, 1975; Raghavan and Wong, 1986; Berry, Drmac, and Jessup, 1999). These weights are used to compute the degree of similarity between the index document and users' query. The main advantages of this model are: (a) its term-weighing scheme improves retrieval performance; (b) its partial matching strategy allows retrieval of documents that approximate the query conditions and sorts it according to the degree of similarity. A disadvantage of the vector space model is a lack of definition of the values of the vector components which is also known as *term weighing*. Some other successful models that are extensions of the vector space model and attempts to address the term weighing problem include Rocchio Algorithm, Latent Semantic Indexing, and Term Discrimination. The Rocchio algorithm (Rocchio, 1971; quoted by Hiemstra, 2009) is a method of relevance feedback which is based on the assumption that users are capable of determining relevant or non-relevant documents amongst query results. Latent Semantic Indexing (Berry, Drmac, and Jessup, 1999; Bradford, 2008) is a technique that projects documents and queries into a space with "latent" semantic dimensions and applies a mathematical technique referred to as Singular Value Decomposition (SVD) to understand terms and concepts within a text collection. Term Discrimination method is similar to term frequency–inverse document frequency (tf-idf) but attempts to assign *weights* to terms such that they are ranked according to their suitability (Salton, Wong, and Yang, 1975). Turney and Pantel (2010) presents a survey of the various use of vector space model and organised them according to the type of matrix involved: term-document, word-context, and pair-pattern.

Probabilistic models

Whilst some extensions to the vector space model introduce *term weighting*, a range of probabilistic retrieval models formally recognise the notion as a means of tracking retrieval relevance. The probabilistic model (Dong, Hussain, and Chang, 2008; Hiemstra, 2009) attempts to capture the IR problem within a probabilistic framework. Given a query q, the probabilistic model assigns to each document d_j, as a measure of its similarity to the query, the ratio P(d_j relevant-to q) / P(d_j non-relevant-to q) which computes the odds of the document d_j being relevant to the query q. The main advantage of this model, in theory is that documents are ranked in decreasing order of their probability of being relevant. This model later evolved into what is referred to as the binary independent retrieval model. Other models which have a strong base in probabilistic theory include the Bayesian network model (Ben-Gal, 2007), which is a probabilistic graphical model where each node represents a random variable, while the edges between the nodes represent probabilistic dependencies among the corresponding random variables. Hidden Markov Models, Neural Networks, and Kalman filters (Griffiths and Yuille, 2006) are further extensions of the Bayesian network model. The Language and Google PageRank models (Brin and Page, 1998; Hiemstra, 2009) are other retrieval models that are grounded in probabilistic theory. Reviews of various probabilistic models for information retrieval have been presented by Griffiths and Yuille (2006); Dong, Hussain, and Chang (2008); and Hiemstra (2009).

2.1.3 Existing multimedia retrieval systems

Most of the early multimedia retrieval systems were largely dominated by the Computer Vision and Image Processing research community. They generally approached the multimedia retrieval problem from bottom-up, by analysing the multimedia low-level features. One of such systems is the Content based retrieval engine (CORE) (Jian-Kang *et al.*, 1995) which adopts multiple feature extraction and

indexing techniques using self-organising neural networks and fuzzy logic. The multimedia analysis and retrieval system (MARS) (Huang, Mehrotra, and Ramchandran, 1996) whose main objective was to provide an integrated multimedia information retrieval and database management infrastructure introduced a relevance feedback architecture. MARS also organised various features (like colour, texture, and shape) into a dynamic user-based retrieval architecture and layout - shape based queries for example, are matched using modified fourier descriptors (MFD). The Virage Media Management System (Bach *et al.*, 1996) is a promising commercial multimedia cataloguing and search tool, which provides a compelling solution to the cataloguing problem. Carrer *et al.* (1997), proposed Video Annotation System (named VANE), which constitutes a semi-automatic tool that facilitates the creation of large video databases. VANE attempted a domain-independent approach and uses the Standard Generalized Markup Language (SGML) as the model of metadata collection. WebSeek (Smith and Chang, 1997) and CueVideo (Ponceleon *et al.*, 1998) are two prototype image and video search engines. They provide tools for searching and browsing image and video repositories using various content-based retrieval techniques that incorporate the use of colour, texture and other properties. The queries can be a combination of text-based searches along with image searches.

The difficulty in translating computable low-level multimedia features (like colour histogram, shape, texture etc.) into high-level semantic concepts that humans can relate to (Hare *et al.*, 2006a; Duygulu and Bastan, 2011; Fauzi and Belkhatir, 2013) led to the evolution of systems that rely on multi-modal approach. Doulamis *et al.* (2000), proposed InforMedia[7] Digital Video Library which integrates automatic speech recognition, natural language processing, image analysis, and information retrieval. Advene (Annotate Digital Video, Exchange on the Net), an on-going project at the University Claude Bernard Lyon 1, attempts to facilitate better visual feature understanding by providing a model and format for sharing annotations about digital

7 http://www.informedia.cs.cmu.edu/

video documents (Sadallah, Aubert and Prié, 2011). A comprehensive survey of techniques, algorithms and tools addressing content-based analysis and visual content representation has been published (Dimitrova, 1999; Antani, Kasturi, and Jain, 2002; Wang *et al.*, 2003; Lew *et al.*, 2006; Snoek and Worring, 2009; Juan and Cuiying, 2010; Dasiopoulou *et al*., 2011; Hu *et al*., 2011).

Various research outcomes from multimedia information retrieval researchers has made evaluation an important issue. Evaluation has become necessary due to the subjective (Kender and Naphade, 2005; Hare *et al.*, 2006a; Duygulu and Bastan, 2011) nature of multimedia annotation by humans and difficulty in generating large test dataset. There is need for some standard benchmark for the evaluation of research outcomes. Towards realising the goal, the National Institute of Standards and Technology (NIST) has sponsored the annual Text REtrieval Conference (TREC) as a means to encourage research within the information retrieval community by providing the infrastructure and benchmark necessary for large-scale evaluation of retrieval methodologies. The TREC Video Track (now referred to as TRECVID) was established by NIST to improve research efforts in content-based video retrieval through large-scale video evaluation benchmarks that the platform provides (Over *et. al.*, 2005; Smeaton, Over, and Kraaij, 2006). TRECVID was started in 2001 and has attracted many participants across the industry and academia and is still very active.

Whilst most reviewed systems focus on extracting meaning from visual data based on the computable low-level features, users of MIR systems are known to process visual information with their cognitive and perceptive abilities (Hare *et al.*, 2006a; Enser, 2008a; Chung and Yoon, 2010). There is limited attempt to fully exploit the extraction of high-level semantic information content expressed in the various modalities of visual data. Content-based MIR techniques are useful in specific application areas like surveillance systems where sample image queries are available and exact or near matches are desired based on the image/video features (Yang, Lovell, and Dadgostar, 2009; Patel and Meshram, 2012). However, it could be very challenging for users to

provide a sketch of certain image concepts based on the features. Existing techniques have been effective for smaller domain-specific collections, but still inadequate for interactive searching on a large scale (Ruger, 2011; Dasiopoulou *et al*., 2011). Furthermore, content-based techniques may not be applicable to the web where general visual data are shared and most users lack specific MIR domain expertise.

2.2 Text(Concept)-based Multimedia Retrieval

Concept-based multimedia retrieval techniques, approach the semantic gap problem from a human perspective (top-down), unlike content-based techniques (presented in section 2.1) which approach the semantic gap problem from the multimedia low-level features (bottom-up). While early research in multimedia retrieval were dominantly content-based, recent research efforts (Hauptmann *et al.*, 2007; Natsev *et al.*, 2007; Zhang, 2007; Snoek and Worring, 2009; Aly *et al.*, 2012; Fauzi and Belkhatir, 2013) have emerged that focus more on concept-based multimedia retrieval approach. Concept-based visual information retrieval systems use high-level semantic textual descriptions to label or annotate semantic concepts in visual data. The objects, settings, scenes, events in the video content are assigned high-level semantic labels by either humans (manual annotation) or machines (automated annotation) which are indexed for subsequent *search, query matching,* and *retrieval*. In manual annotation, humans apply their cognitive abilities to achieve high-level semantic annotations. Such annotation can pose a huge challenge due to the need to have experts in such video content domain perform the annotation and also the long period of time required to have the annotation done. In addition, manual annotation could be subjective and error-prone (Kender and Naphade, 2005; Duygulu and Bastan, 2011) due to varying levels of human perception of the same video content (Spink and Cole, 2005). The machine-driven annotation approach, automatically assigns semantic labels to video segments based on prior

knowledge. Machine-driven automatic multimedia semantic annotation approaches have been proposed in the literature such as the domain-based semantic classifiers proposed by Szummer and Picard (1998), Vailaya *et al.* (2001), Hauptmann (2004), and Wiliem *et al.* (2012). Other general machine-learning approaches have emerged that attempt to map the multimedia low-level features to the high-level semantic concepts (Snoek *et al.*, 2006; Mylonas *et al.*, 2008; Dalakleidi *et al.*, 2011).

Video content cannot be used in its native form in a concept-based visual retrieval system, hence the need to transform and represent the semantics in a machine-understandable format that can be used in such systems (García and Celma, 2005; Dasiopoulou *et al.*, 2011). The ideal way to describe video content is in terms of its objects (Visser, Sebe, and Bakker, 2002; Lavee, Rivlin, and Rudzsky, 2009). This is because humans conceptualise things as objects (e.g. table, chair, bed, phone etc.) and also have the innate ability to give meaning to a collection of objects based on their layout and other factors. Hence, an enclosed wall, with a bed, chair, and dressing table, may be identified as a bedroom and not otherwise. However, it is very difficult to achieve semantic object recognition in video databases. Researchers tend to extract low-level features (shape, colour, texture, etc.) in order to describe video content. Hauptmann *et al.* (2007) through their experiment have shown that high-level concepts could facilitate more accurate video retrieval systems.

This rest of this section presents research efforts leading to machine-driven automatic annotation of visual content and the semantic web which Berners-Lee *et al.* (2001) defined as "a web of data that can be processed directly and indirectly by machines". Metadata representation, ontologies, and knowledge management as they relate to annotation and retrieval of video content are presented in the next sub-sections. Existing concept-based visual retrieval systems are also reviewed.

2.2.1 Metadata Representation

Metadata provides additional descriptive information about multimedia objects due to the absence of self-descriptive and directly accessible formal semantics in multimedia objects. Various relevant metadata representations in the context of visual semantic description and the semantic web are discussed.

The eXtensible Markup Language (XML[8]) defines a set of rules that allows for document encoding in both human-readable and machine-readable formats. The use of XML have been very successful in data encoding on the Internet. However, users are able to add arbitrary structure to their encoded documents but these structures lack semantics (Erdmann and Studer, 2000; Celma et al, 2007). This lack of clear semantics hinders metadata interoperability and makes XML on its own, unsuitable for use in representing multimedia semantics. MPEG-7 MDS (Multimedia Description Schemes) with its XML-based syntax was developed to provide constructs for the definition of multimedia metadata for multimedia content and services description (Koenen (*ed.*), 1999; Zhang, 2007). These constructs are essentially based on the XML Schema Language, extended with basic data types necessary for the definition of Descriptions Schemas for the MPEG7-MDS (Polydoros, Tsinaraki, and Chirstodoulakis, 2006). MPEG-7 facilitates seamless interchange across applications on the Internet and provides standardized tools for describing different aspects of multimedia at different levels of abstraction (Celma *et al.*, 2007). Multimedia Description Schemes (MDS) describe the multimedia content at a number of levels including signal structure, features, models, and semantics. The MDS semantics provide a way to describe what is depicted in the multimedia content from the real world, such as objects, people, and events (Eze and Ishaya, 2007). MDS descriptions are based on the notion of objects and

8 http://www.w3.org/XML

events, abstract notions and their relationship. However, it has been identified that MPEG-7 lacks precise semantics and would require some machine-understandable semantic language in order to make it re-usable and interoperable with other domains (Hunter, 2005; Tous and Delgado, 2010). In order to bridge the lack of proper semantic support in MPEG-7, Hunter (2005) proposed an MPEG-7 ontology so as to make MPEG-7 accessible, re-usable and interoperable with other domains in a machine-understandable language.

The Resource Description Framework (RDF)[9] provides means for adding semantics to a document. It is an infrastructure that enables encoding, exchange and reuse of information structured metadata and has been useful in information modelling. RDF describes resources in the form of subject-predicate-object expressions. RDF lacks cardinality and has limited data types. This led to the development of the Web Ontology Language (OWL) which is a stronger language with larger vocabulary and greater machine understanding than RDF. OWL is a vocabulary extension of RDF and is somewhat the dominating standard in ontology definition. OWL has been developed according to the description logic paradigm and uses RDF(S) syntax[10]. OWL is designed for use by applications that need to process the content of information instead of just presenting information to humans. Eze and Ishaya (2007) posits that "OWL unifies the epistemologically rich modelling primitives of frames, the formal semantics and efficient reasoning support of the description logics and mapping standard Web metadata language proposals". OWL has formal semantics with strong logical foundations to facilitate the representation of multimedia semantics in a well-defined ontology.

9 http://www.w3.org/RDFs
10 http://www.w3.org/TR/owl-features/

2.2.2 Ontology

The English language is inherently ambiguous. The same expression is used to represent different concepts based on the context of use. This is not much of a problem in human communication since humans use gestures, contextual information and innate intelligence to fathom what is being referred to. But modelling these ambiguous concepts in a machine understandable format becomes a serious issue. This is part of why most web search engines return irrelevant results. An example is where a user wants to retrieve video clips that depict “bank” (as a financial institution term) from a large video repository. Existing systems will likely retrieve clips related to data bank, riverbank, and financial bank. This is quite obvious since the concept of ‘bank’ is not strictly defined.

Ontology is seen as a key technology for enabling semantics-driven knowledge processing. It offers an alternative way to deal with the representation of heterogeneous web resources. Ontologies are being employed to provide a framework for sharing a precise meaning of concepts. In the context of computer and information sciences, Gruber (2009) defined ontology as:

> "*... a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application. ....*"

Creating an ontology involves explicitly defining every concept to be represented. This entails a thorough definition of all possible attributes of the concept with all valid constraints and how they interrelate. Ontology can play a crucial role in enabling web-based knowledge processing, sharing, and reuse between people and application

systems. It has been described as the language for the semantic web where intelligent agents interact in a machine-readable web (World Wide Web Consortium (W3C), 2009). An ontology typically contains a hierarchy of concepts within a domain and describe each concept's crucial properties through an attribute-value mechanism. Further relations between concepts might be described through additional logical sentences. Constructing an ontology basically involves the determination of concepts; establishment of the properties for the concepts and how they interrelate; and maintenance of the ontology.

Ontologies have been commonly used in IR to disambiguate meaning and improve precision. One common approach is to extend users' queries by adding semantically related terms to the original query. This ensures that documents that do not necessarily contain terms in the original query may not be retrieved. Paz-Trillo *et al*. (2004) proposed the use of ontologies for the retrieval of art exhibition video information based on an art ontology under development. OntoLog is a tool for annotating video and audio using ontologies which uses an annotation scheme based on hierarchical ontologies, and an RDF-based data model that may be adapted and extended through the use of RDF Schema (Meland *et al.*, 2003). Khan and McLeod (2000), Schreiber *et al.* (2001), Petridis *et al.* (2006), Polydoros, Tsinaraki, and Chirstodoulakis (2006), and Gómez-Romero *et al.* (2011) provide various approaches of using ontologies in multimedia information retrieval. Figure 2.3 shows a stack of various languages for developing ontologies.

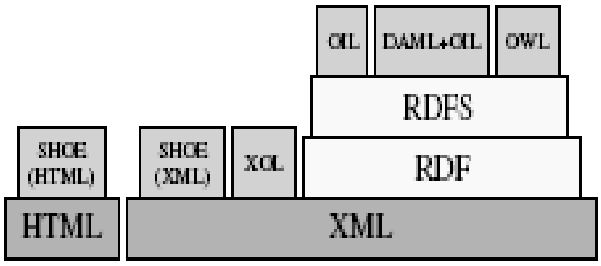

Figure 2.3: The stack of ontology markup languages (Corcho e*t al*., 2003)

The Web Ontology Language (OWL) has been proposed by World Wide Web Consortium (W3C) as the official ontology language for the semantic web. Corcho e*t al*. (2003) has a review of methodologies, tools, and language for building ontologies.

Ontologies allow data organisation at the semantic level and therefore, will be useful in this research towards extracting and representing multimedia semantic knowledge.

2.2.3 Knowledge Management (KM)

Knowledge is a precondition for acting purposefully in a given environment or domain area. Knowledge is synonymous with the notion of truth in philosophy (Conklin, 1974). Knowledge in computer science, is regarded as organised information (refer to Figure 2.4) and it can be contextual, time-sensitive, and requires trusted relationship to foster. In visual information retrieval, the knowledge level is aspired to as the most expressive and perceptive description level and likened to high-level context-sensitive semantic understanding. Knowledge does not come of its own accord – it is acquired. This acquisition of knowledge takes place in the form of experiencing (intuitive knowledge) or learning (erudition).

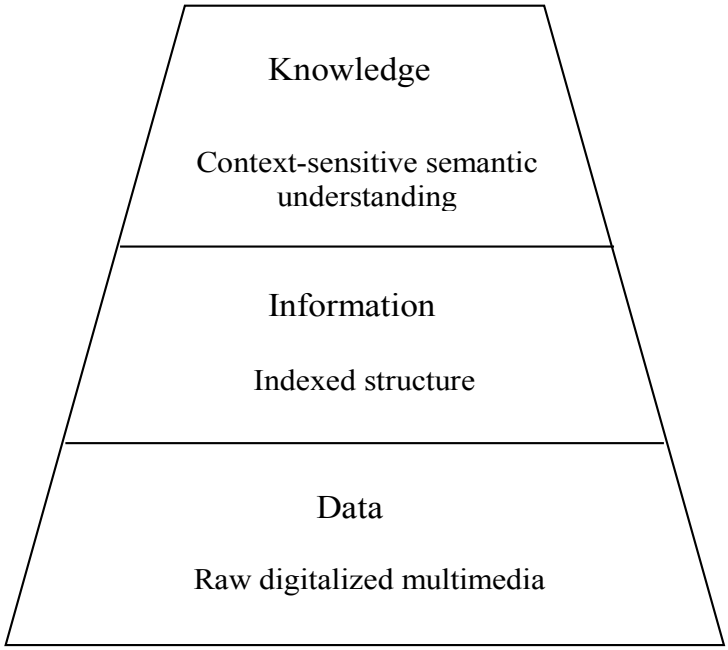

Figure 2.4: Pyramid of Knowledge

In ancient human history, communities flourished due to the knowledge they possessed and shared. This possession of knowledge which is reflected in their attitude towards one another is usually referred to as tradition or culture. This tradition or culture can be likened to what is now known in modern computing as knowledge management. It therefore, suffices to say that a prerequisite for knowledge sharing is that all involved actors must have a common knowledge interest in a given domain. The key point in this analogy is that knowledge management is always domain based and therefore deals with the capturing, organisation, and provision of knowledge to enable well-informed decision-making and knowledge transfer in organisations.

Most KM techniques expect data in a structured format. The hitherto existing KM approaches organise knowledge in portals and rely on text as the medium to transfer knowledge. Due to high availability and the rich semantic content in multimedia data, they are more suitable in KM systems. However, a key challenge is efficiently

representing the unstructured multimedia content into a structured format for effective and intuitive knowledge retrieval. Intelligent KM systems generally help users with their information need by minimising complexity, thereby reducing the users' cognitive load. These systems have a learning component and gain experience over time. They respond to changes and new situations with minimal human intervention. They are context-sensitive and capable of making sense out of ambiguous information. Admittedly, most of the systems that were reviewed in sections 2.1.3 are still far from KM systems. Most of these systems require human intervention at the indexing stage. This creates an unreliable content description as users might miss some important details or introduce their bias. They are also inefficient, as they require a lot of time at the indexing stage. These systems require several changes and improvement before they can evolve to KM systems. Liao (2002) provides a review of KM technologies and applications.

Huang and Tao (2004) identified that current knowledge management applications lack context-awareness in their information and knowledge manipulation, and proposed the addition of context awareness to knowledge management systems based on a set of meta-information elements. Intelligent software agents can make decisions on a set of options based on past experiences. They are fundamentally knowledge-based and would be employed in the proposed system framework in the next chapter, to achieve a fully-fledged multimedia KM system. One possible application area of a multimedia KM system could be in the organisation of the recorded video of the proceedings of a lecture. This can be indexed and archived and students could search and retrieve certain parts of the lecture that might be of interest to them.

Multimedia KM will allow users to reuse knowledge from various semantic dimensions and also help them share knowledge across different domains. A major challenge with multimedia KM is the issue of knowledge extraction and user contribution to the knowledge base.

2.2.4 Existing Concept-based Multimedia Retrieval Systems

Although most visual information indexing and retrieval systems rely on low-level features, other approaches still exist, which use high-level concepts. Hunter, Schroeter, Koopman, and Henderson (2004) proposed Vannotea which was developed at the University of Brisbane and supports distributed knowledge management by allowing collaborative annotation from various users. Petridis *et al.* (2006) developed M-OntoMat Annotizer which is part of the CREAM (Handschuh, Staab, and Studer, 2003) framework and provides an ontology-based manual image and video annotation. The OntoMat project team developed an extension (S-CREAM) of the OntoMat annotation tool to support semi-automatic annotation by learning user annotations and subsequently suggesting annotations based on what was learnt (Handschuh, Staab, and Studer, 2003). OntoLog (Heggland, 2005) is a tool for annotating (describing and indexing) video and audio using ontologies. Fan *et al.* (2004b) proposed a framework called *ClassView* – a hierarchical semantics-sensitive video classifier. The hierarchical structure is derived from the domain-dependent concept hierarchy of video contents in the database using the Expectation-Maximisation (EM) algorithm. This approach does not properly represent semantic concepts in video as the classification is done at the shot level. A single video shot can contain multiple semantic concepts. VisionGo (Luan, Zheng, Wang, and Chua, 2011) is an interactive, semi-automated video annotation and retrieval system that features multiple feedback techniques and motion-icons to enhance dynamic visual semantic understanding. A discussion on issues relevant to visual information retrieval and capturing semantics present in images and video is presented by Colombo, Bimbo, and Pala (1999) and Uren *et al.* (2005).

There are many recent and on-going projects within the multimedia semantics research community which further support the importance of the research domain and the fact

that there are many open issues. The RUSHES (Retrieval of mUltimedia Semantic units for enHanced rEuSability) project (Zhang and Izquierdo, 2011) aims to provide a system for indexing and retrieving raw, unedited audio-visual footage known in the broadcasting industry as "rushes". It adopts a Bayesian network model using the K2 algorithm for semantic context learning and inference. The representations for multimedia content are organised in three semantic levels, namely: Low-level, Mid-Level, and High-Level. The project focused on the mid-level as a means to bridging the multimedia semantic gap between low-level and high-level features as a two-step approach. K-Space (Knowledge Space of Semantic Inference for Automatic Annotation and Retrieval of Multimedia Content) is a European Network of Excellence (NoE) comprising research teams from academia and industry conducting research in semantic inference for semi-automatic annotation and retrieval of multimedia content. K-Space integrative research focuses on three main areas: Content-based multimedia analysis, Knowledge extraction, and Semantic multimedia. They approach the multimedia semantic gap problem from a multimedia knowledge extraction perspective by integrating research efforts on semantic approaches, content- and knowledge-based media engineering (Izquierdo *et al.*, 2007). One of the research outputs on the K-SPACE project is the K-Space Annotation Tool (KAT) which provides a framework for semi-automated multimedia semantic annotation. K-Space evolved into the Semantic Multimedia Research and Technology (SMaRT) Scientific Society which builds on the K-Space goals and involves other semantic multimedia research efforts. The VIDI-Video[11] project is focused on developing a semantic search engine based on a thesaurus for automatic semantic concepts identification (Papadopoulos *et al.*, 2009) towards improving video retrieval. The MESH[12] project advocates a semi-automated multimedia semantic extraction and personalised multimedia summaries with special

11 http://www.vidivideo.info
12 http://www.mesh-ip.eu/?Page=Project

focus in the area of News videos. The Bootstrapping Ontology Evolution with Multimedia Information Extraction (BOEMIE[13]) project (Paliouras, Spyropoulos, and Tsatsaronis, 2011) aims at providing easy access to multimedia semantic content by developing a methodology for knowledge extraction and evolution, using a rich multimedia semantic model based on domain specific multimedia ontologies. A key framework within the BOEMIE project is the media interpretation framework which aims to compute high-level content descriptions of media documents from lower level information extraction results using conceptual and contextual knowledge. The conceptual knowledge are based on formal ontology while the contextual knowledge refers to specific prior knowledge relevant for the high-level interpretation.

There are other professional systems that provide mechanisms for automated visual information semantic annotation. The MediaMill[14] is an online solution from the Intelligent Systems Lab of the University of Amsterdam, which facilitates semi-automated, content-driven semantic visual information search and sharing. The IQ Engines[15] supports semantic object recognition (scenes, landmarks, etc.) in images and provides automated image tagging and stacking for users of the portal. Tineye[16] is a reverse image search engine that identifies and associates images based on image content analysis. CuZero[17], is an interactive visual information search system that bridges the human-computer interface problem by maximizing the use of auto-recommendation systems and a navigation map focused on concept-based visual information exploration (Zavesky and Chang, 2008). Other systems include: Blinkx[18],

13 http://www.boemie.org
14 http://mediamill.cla.umn.edu/mediamill
15 https://www.iqengines.com/
16 http://www.tineye.com
17 http://www.ee.columbia.edu/ln/dvmm/cuzero/
18 http://www.blinkx.com/

Kooaba[19], and VideoSurf[20] which is now part of Bing[21].

Whilst most of the reviewed concept-based multimedia retrieval systems are more expressive (Hare *et al.*, 2006a) and provide various levels of semantic concept detection, they still lack the ability to identify and label semantic concepts at the exact level of human semantic perception based on visual features (Enser *et al.*, 2007). Annotation of abstract concepts like "happiness" are difficult to detect and represent (Hanjalic and Xu, 2005). MIR systems should be able to allow users pose queries that reflect human perceptual expectations based on their search needs and be assured of good retrieval results. Most of the research outcomes like the *ClassView* framework proposed by Fan *et al.* (2004b), focus on concept detection at the shot level. However, MIR users are also interested in finding concepts that are possibly spread across multiple shots. It is therefore necessary to combine the annotation output across multiple shots towards achieving semantic retrieval. MIR systems such as S-CREAM (Handschuh, Staab, and Studer, 2003), MESH[22], and VisionGo (Luan, Zheng, Wang, and Chua, 2011) provide semi-automatic annotation. However, having humans provide annotation can be very time-consuming and subject to bias (Bloehdorn *et al.*, 2005; Kender and Naphade, 2005; Duygulu and Bastan, 2011). Other systems like MediaMill[23] and CuZero[24] have proposed automated annotation to overcome such shortcomings. Nevertheless, such systems are still short of achieving automated semantic understanding at the human perceptual level (Snoek *et al.*, 2006; Duygulu and Bastan, 2011; Hu *et al.*, 2011). The next section presents a taxonomy of multimedia information retrieval.

---

19 http://www.kooaba.com/
20 http://www.videosurf.com
21 http://www.bing.com/videos/browse?FORM=VQFRVS
22 http://www.mesh-ip.eu/?Page=Project
23 http://mediamill.cla.umn.edu/mediamill
24 http://www.ee.columbia.edu/ln/dvmm/cuzero/

2.3 Taxonomy of Multimedia Information Retrieval (MIR)

Considering the multidisciplinary nature of approaches for MIR, different researchers approach the MIR issue from different levels of abstraction. Chang, Smith, Beigi, and Benitez (1997), provided a broad taxonomy of CBVIR systems based on functionality and mode of operation. A summary of this taxonomy is provided in Table 2.1.

Automation	Indexing features of images and video will determine the search functionalities users may use. Preparation of these indexes may be fully automatic or manual. Automatic methods usually are useful for low-level feature extraction only. High-level semantic indexes usually require human input in annotation or system training. Some systems provide interactive tools to assist users in selecting image objects and features and are thus qualified as semi-automatic systems.
Multimedia Feature	Indexing features may include those of images, video, text, audio, or any combinations. Most systems use individual features or multiple features independently. Integration of features have been shown in a few systems but has not been fully explored.
Adaptability	Most systems use static indexing features, which are extracted in advance. Selection of features involves trade-offs of indexing cost and the search flexibility. However, due to the subjective nature of visual search, there exist needs of dynamic indexing features which adapt to changing user needs and application contexts.

Abstraction	Images may be indexed at various levels,including feature (e.g., colour, texture, and shape), object (e.g., moving foreground object), syntax (e.g., video shot), and semantics (e.g., image subject). Most automatic systems aim at low-level features, while the high-level indexes are usually done manually. Interaction among different levels is an exciting but unsolved area.
Generality	The indexing schemes and database content may be customized to incorporate specific domain knowledge, such as those in medical and remote-sensing applications. Other systems may aim at unconstrained types of visual content such as those on the Internet.
Content Collection	The population of content could be achieved by software robots, which roam freely over the World-Wide Web and automatically download visual content according to some heuristics. Or visual content may be manually prepared by domain operators, such as the on-line news archives and photo stocks. Or in the future, visual content may be submitted autonomously like the way people submit documents to the Usenet today.
Categorisation	When the database size grows, subject taxonomies will be very useful in providing hierarchical categories where users may freely navigate and browse through the entire database archive. Some systems simply provide browsing tools for users to interactively view images or video of interest, which can then be followed by a detailed query.

Table 2.1: Taxonomy of **MIR** systems based on functionality and mode of operation (Chang, Smith, Beigi, and Benitez, 1997)

The taxonomy presented in Table 2.1 is quite broad. The rest of this section attempts to present a more narrow classification of MIR. As computer technologies evolve towards the new era of user-centric model, it therefore becomes pertinent to classify MIR systems/techniques based on users' information need. The taxonomy of MIR techniques is categorised in three levels based on varying aspects of users' information need. The first level is the feature level, which deals with multimedia data at a fairly low-level. Low level multimedia features like shape, colour histogram, texture, camera motion, and others are modelled. Most automatic feature extraction techniques reviewed in section 2.1 fall under this category. A typical query at this level could be: "*Find me a shot that shows a colour composition with at least 30% of blue*". The second level is the content level. At this level, objects represented in multimedia content are identified using object detection and matching techniques (Donderler, Ulusoy, and Gudukbay, 2004; Fei-Fei, Fergus, and Perona, 2006; Quelhas *et al.*, 2007). The Query by visual example (QBVE) technique falls under this category. Content-based multimedia indexing and retrieval systems based on this classification include WebSeek and VideoCue. Possible query at this level include: "*Find me all shots with the New York twin towers*".

Level 3 is the knowledge level and consists of high level knowledge retrieval based on the semantics and context of the multimedia content. User queries at this level are inseparable with the human knowledge level. Such example queries include:

- *"Find me all Rooney's goals for England"*
- *"Find me all Beckham's free kick"*
- *"Find me all clips where Beckham missed penalty kicks in Euro 2004 games"*

The knowledge level query is certainly easier to use, especially for naive web users. The taxonomy of multimedia information retrieval based on users' information need is

presented in Figure 2.5.

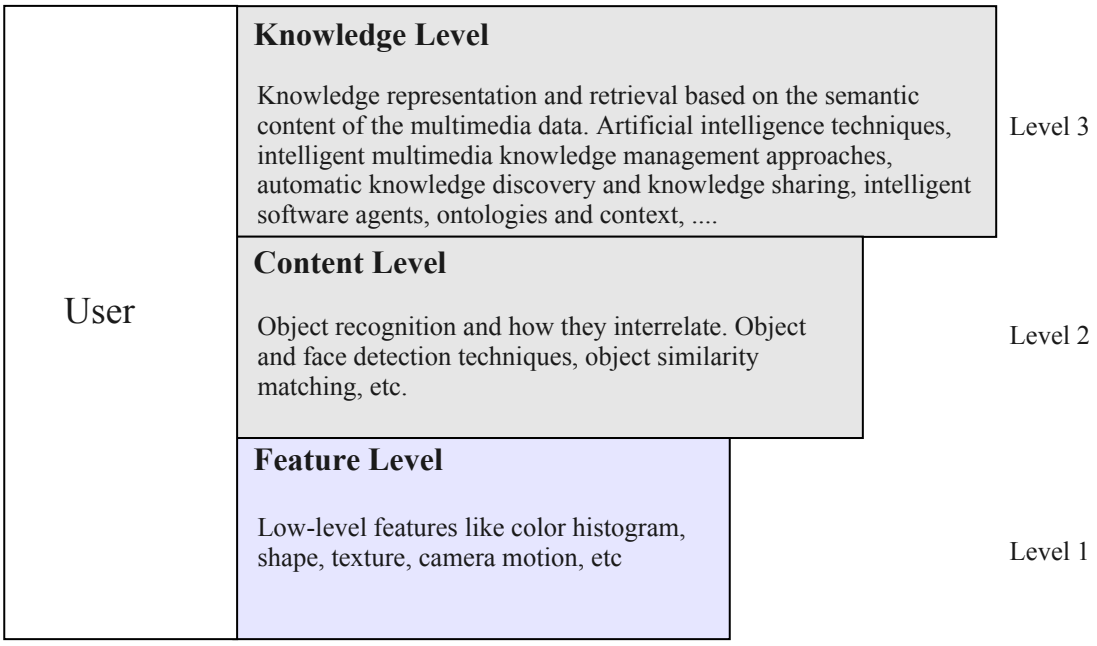

Figure 2.5: A taxonomy of multimedia information retrieval approaches

While there are on-going efforts by the research community at various levels of abstraction involved (see Figure 2.5), there are no concrete existing MIR systems that are fully at the knowledge level (Hu, 2011). However, some of the visual information retrieval systems reviewed in sections 2.1.3 and 2.2.4 aim to operate at the knowledge level but are currently somewhere between content and knowledge level. The majority of web users will be more interested in multimedia information retrieval at the knowledge level; since IR users naturally express their information need at the semantic level and would not understand low-level multimedia details (Hare *et al.*, 2006a; Hider, 2006; Lew *et al.*, 2006; Enser, 2008a; Chung and Yoon, 2010; Fauzi and Belkhatir, 2013). This thesis is more concerned with visual information retrieval at the knowledge level. Thus, the rest of this chapter focuses on approaches to semantic multimedia retrieval at the knowledge level.

2.4 Context in Visual Information Retrieval

Humans are quite successful in conveying ideas to each other and reacting appropriately. This is due to some common knowledge they share and their innate ability to draw knowledge from other factors surrounding them, in order to understand situations better. Computers are unable to imitate humans in this respect since they lack common knowledge and the ability to sense or understand *context*. Various areas of Computer Science have been investigating the concept of context over the years (Belotti *et al.*, 2004; Loyola, 2007). Context is a key concept in natural language processing and more generally in artificial intelligence, human-computer interaction, and mobile computing. However, there are many diverse interpretation of the notion of context in these various areas of Computer Science which lead to multifarious definitions of the concept (Loyola, 2007). Context is seen in mobile computing as "any information that can be used to characterise the situation of an entity" (Dey, 2000). In artificial intelligence, context is seen as a "means of partitioning knowledge into manageable sets" (Theodorakis, Analyti, Constantopoulos, and Spyratos, 2002; Theodorakis and Spyratos, 2002). In linguistics, context is the part of a discourse surrounding a word or passage that helps make its meaning clear[25].

The normal experience with web search engines is that various hyperlinks are retrieved in response to user queries and the burden is left to the user to navigate the search results and identify the relevant ones. This is a very tedious task since some of these web search engines lack context and user profiling. Consider searching for "News apps" at the Google Play store for Android phones and tablets. If the user performed such a search from London, the retrieval results will likely reflect news apps for UK and possibly other news sites within Europe. Similarly, if the user performed the search from the USA, the retrieval result will include news apps for USA news sites. This is because Google tracks the location of the user and uses that information to filter search

25 http://www.sil.org/lingualinks/literacy/ReferenceMaterials/GlossaryOfLiteracyTerms/WhatIsContext.htm

results. So much depends on the context of the user and the query. The notion of context as used in IR is closely related to the concept of user modelling. IR researchers have recognised the need to consider representing the user in an IR system (Torres and Parkes, 2000; Ruthven, 2011). Having prior knowledge about users' information need, undoubtedly aids in accurate information retrieval. User modelling was defined by McTear (1993) as "the process of acquiring knowledge about a user in order to provide services or information adapted to their specific requirements". Approaches to user modelling are described by Belkin (1997) and Biswas (2012). User modelling is expected to improve retrieval precision in information retrieval. However, the focus may involve better modelling of the user's background, personalisation, the context in which information access occurs, etc.

Aside from applying context to users and queries in multimedia retrieval systems, context can play a key role at the multimedia semantic annotation level (Wiliem, Madasu, Boles, and Yarlagadda, 2012; Yi, Peng, and Xiao, 2012). Information retrieval researchers, Myrhaug and Goker (2003) defined context as a description of the aspects of a situation. Goker, Myrhaug, and Bierig (2009) identified three theoretical models for interaction and context within Information Retrieval as Cognitive, Episode, and Stratified interaction models. However, this thesis approaches the notion of *context* from a multimedia semantic representation perspective, and defines it as ***any information about the multimedia content that enhances the semantic understanding of the multimedia content***. A formal definition of context is presented in section 3.4.1.

Hoogs *et al.* (2003) applied contextual information from transcribed commentary in the automated annotation of visual information content. Huang *et al.* (2003), introduced contextual knowledge to facilitate knowledge communication in multimedia e-Learning environment. Mylonas *et al.* (2008) proposed the use of context within their taxonomic knowledge representation model, to interpret the meaning of multimedia documents. They regard context as the common meaning that different concepts share. For example, basketball is considered the context of ball, referee and basket since it is a

common antecedent for all three concepts. Zhang and Izquierdo (2011) adopts a Bayesian network model using the K2 algorithm for semantic context learning and inference. A further review of context models in visual information semantic understanding is presented in section 3.4.4. While there are various efforts towards applying context to improving multimedia semantic understanding, most of the reviewed works have either focused on a specific domain or limited contextual knowledge to the application domain. There is need for a generic and domain-independent context model.

The possible sources of context information in a video clip include audio transcript, text annotation, textual description from video source, etc. In information retrieval, context is seen to use users' situational information (Gross and Klemke, 2003; Ruthven, 2011). Users' current situation is analysed, compared to available information, and information that is considered of most value in the situation is provided to the user. While a lot of research efforts have gone into improving information retrieval results by making use of contextual knowledge about users and queries (Gross and Klemke, 2003; Ingwersen and Jarvelin, 2005; Natsev *et al.*, 2007; Borlund *et al.*, 2008; Goker, Myrhaug, and Bierig, 2009), there is also the need to improve the metadata created at the indexing stage. Applying the best retrieval methodologies or algorithms to a poorly indexed repository will still yield unsatisfactory retrieval results. It is argued that context can be used at the production (i.e. content representation and indexing) stage to capture improved and clearer description of multimedia content so as to provide support for improve IR precision. This is very challenging as it is difficult to recognise and model all contextual dimensions of multimedia data.

Context can be formalised and represented using first-order logic, set theory, directed graph, or other artificial intelligence techniques (McCarthy and Buvac, 1998). But since the focus is on applying context to the web in order to facilitate richer multimedia semantic understanding, it is necessary to represent it into generic markup languages

such as eXtensible Markup Language (XML)[26] and Resource Description Framework (RDF)[27]. This would facilitate context reuse and sharing.

2.5 Analysis and Problem Justification

Visual information is rich in content and usually connote different meanings to different individuals owing to their different backgrounds and circumstances (Quelhas *et al*., 2007). While multimedia low-level features are easily measured and computed, users of MIR systems process visual information with their cognitive and perceptive abilities (Hare *et al.*, 2006a; Enser, 2008a; Chung and Yoon, 2010). Even though computers can perform logical and mathematical computations many times faster than humans can ever attempt, humans can perform cognitive and perceptual analysis a lot faster than computers. There is need to design models that will allow computers understand the high-level semantic meaning in multimedia documents by possibly translating computable low-level multimedia features (like colour histogram, shape, texture etc.) into high-level semantic concepts which should be naturally understandable to human users. This is commonly referred to in the literature as the "semantic gap" between multimedia low-level features and the high-level concepts that the multimedia resources express (Hare *et al.*, 2006a; Zhang, 2007; Duygulu and Bastan, 2011; Hu *et al.*, 2011).

Content-based multimedia information retrieval techniques approach the "semantic gap" problem from the bottom-up by extracting low-level features (like colour, texture, shapes) and allowing queries at such a low-level. They are very useful in specific application areas like surveillance systems or medical images where sample image

26 http://www.w3.org/XML/
27 http://www.w3.org/RDF/

queries are available and exact or near matches are desired based on the image/video features (Yang, Lovell, and Dadgostar, 2009; Patel and Meshram, 2012). However, content-based techniques may not be applicable to the web where general visual data are shared and most users lack specific MIR domain expertise. It can also be argued that having humans annotate visual data can be time-consuming and subject to errors due to human bias (Kender and Naphade, 2005; Duygulu and Bastan, 2011). When posing query to content-based visual retrieval systems, the user often has an incomplete idea of the concept to search for. Even where the user has a good idea of the concept to search for, it may not be easy to clearly represent the concept in terms of low-level features (Enser, 2008b). For example, users cannot pose a query of "golden goal" in terms of low-level features like colour and shape nor could they produce a sample image. The "golden goal" rule, is a concept in football that states that the first team to score during the 30 minutes of extra-time (after a stalemate in the full 90 minutes of play) would immediately be declared the winner.

Approaching multimedia information retrieval from the top-down (concept-based multimedia information retrieval), high-level concepts like "golden goal" can be identified and assigned a symbolic label. While there are research efforts to propose theoretical models or principles for indexing visual information such as those proposed by Panofsky (1962), Shatford (1986), and Jaimes and Chang (2000); labelling or describing visual data does not always capture all the semantics (Hare *et al.*, 2006b; Enser, 2008b). It may not be possible to identify and label semantic concepts at the exact level of specificity based on visual features (Enser *et al.*, 2007). According to Hare *et al.* (2006b), "the relationships between the objects as depicted in the image, and the variety of connotations invoked, the implied relationship with the world at large, implied actions, and the broader context, all contribute to the rich high-level full semantic representation of the image". While concept-based retrieval techniques are

more expressive (Hare *et al.*, 2006a) compared to content-based techniques for domain-independent multimedia retrieval on the web, there are problems associated with the annotation of concepts relating to abstract concepts or emotions such as "happiness" (Hanjalic and Xu, 2005).

A picture is said to worth a thousand words. But how best can these words be arranged to give an exact description of the picture? In an experiment by Rorissa, Clough, and Deselaers (2008) in their investigation of the correlation between low-level visual image features and human similarity perception, and whether it is possible to find a combination of features that closely resembles human similarity perception showed a positive correlation. It is argued that a combination of extracted features and *context*, will aid in better semantic annotation. Combining the indexing of semantic concepts alongside their *context* will facilitate the retrieval of the same object under different circumstance or user perception. Automated annotation is required to overcome the shortcomings of manual annotation which is mainly subjective and time-consuming (Bloehdorn *et al.*, 2005; Kender and Naphade, 2005; Duygulu and Bastan, 2011).

Liu *et al*. (2007) in their survey of content-based image retrieval with high-level semantics, assert the need for an integrated content-based image retrieval framework that includes low-level feature extraction, effective learning of high-level semantics, friendly user interface, and efficient indexing tool. According to Liu *et al.* (2007) most existing systems limit their contributions to one or two of such components. In addition, the need to have a robust method and format of describing metadata prior to storage in repositories by using standard schema that ensures interoperability between different systems, was identified (Pino and Di Salvo, 2011). Although MPEG-7 metadata are widely used in MIR, Semantic Web technologies, such as RDF and OWL have been shown to be more suitable to represent and adapt multimedia semantics to

specific contexts and achieve interoperability and reuse (Tous and Delgado, 2010).

Basic multimedia annotation on its own does not guarantee semantic understanding until properly represented and organised in a machine-readable form such that knowledge can be inferred, shared, and discovered (Dalakleidi *et al.*, 2011). Towards bridging the semantic gap between multimedia low-level features and high-level concepts, this research approaches the MIR issue from a knowledge representation (refer to section 2.3) perspective by investigating the application of *context*.

2.6 Summary

This chapter provided a survey of several research efforts in multimedia semantic indexing and retrieval. Due to the multimodal nature of visual data, where the author can express semantic ideas, using more than one information channel (Snoek and Worring, 2005), it has become necessary to perform searches and retrieval of visual content at such a semantic level (Enser, 2008a; Fauzi and Belkhatir, 2013). However, multimedia low-level features are easily measured and computed, but users process the information contained in visual data with high cognitive abilities (Hare *et al.*, 2006a; Chung and Yoon, 2010) that computing power is incapable of (Duygulu and Bastan, 2011). This is referred to as the "semantic gap" which Smeulders *et al.* (2000) defines as the "*lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation*". The two main approaches for visual information semantic retrieval were reviewed: Content-based and Concept-based multimedia information retrieval. Content-based multimedia information retrieval adopts a bottom-up approach and relies on multimedia low-level feature extraction to index and match multimedia data, while

concept-based adopts a top-down approach and uses textual annotation to index and retrieve the high-level semantic concepts in multimedia data. Refer to sections 2.1 and 2.2 for a detailed review of these approaches and various existing systems.

A taxonomy of multimedia information retrieval was presented and the notion of *context* as it relates to information retrieval (Ingwersen and Jarvelin, 2005; Goker, Myrhaug, and Bierig, 2009), and specifically semantic multimedia content representation, was reviewed. Context was defined as *any information about the multimedia content that enhances the semantic understanding of the multimedia content*. An analysis of the literature survey was presented and it was identified that this thesis approaches the "semantic gap" problem by investigating the application of *context* to bridging it, through the creation of a framework and models that facilitate the development of tools for context-based semantic visual information annotation and retrieval. The next chapter presents the context-based approach to the research problem.

Chapter 3

Development of Context-based Multimedia Management Framework

An analytical review of the state of the art in visual information indexing and retrieval, and other related technologies has been presented in Chapter 2. The need to bridge the semantic gap between multimedia low-level features and high-level concepts was identified. This chapter presents the development of the context-based semantics integration framework for heterogeneous multimedia sources. This framework seeks to provide visual information annotation and retrieval at the knowledge level (see figure 2.5). Techniques for the use and representation of context information in visual data from heterogeneous sources are presented. While there are components in the framework which are common to other frameworks in the literature, the unique component in the framework is the context model. This led to further development and formalisation of the context model for semantics understanding and representation.

This chapter has four main sections. Section 3.1 presents the research methodology that was undertaken in the development of the framework and model; section 3.2 presents a review of existing multimedia annotation and retrieval frameworks; section 3.3 presents the limitation of the state of the art that will inform the decisions in formulating the conceptual framework; while section 3.4 presents the context approach.

3.1 Methodology

An exploratory research methodology was adopted to develop the framework. Various existing multimedia annotation and retrieval frameworks were reviewed and their

shortcomings identified. The conceptual framework was proposed based on the identified shortcomings from existing frameworks. Since the framework is complex, and semantic multimedia retrieval starts with multimedia semantic understanding and representation, it was necessary to formalise the context-based semantic understanding and representation model through rigorous specification. The next section presents a review of existing multimedia annotation and retrieval frameworks.

3.2 Review of Existing MIR Frameworks

A review of existing content-based and concept-based multimedia retrieval systems was presented in sections 2.1.3 and 2.2.4 respectively. A taxonomy of multimedia retrieval systems presented in section 2.3. This section presents a review of some known frameworks from the literature that impacted on the proposed conceptual framework. The reviewed frameworks are some of the known recent frameworks in the literature that stand out and are closest to the proposed framework in this thesis. They are selected across both content-based and concept-based multimedia retrieval approaches. The researcher acknowledges that all possible frameworks in the literature may not have been reviewed. However, the analysis in this section is not limited to only the reviewed MIR frameworks in this section, but extends to all the reviewed multimedia retrieval systems in chapter 2.

Juan and Cuiying (2010) proposed the framework in Figure 2.1 (presented in section 2.1), which comprises mainly of a video preprocessing subsystem and the video query subsystem. The video preprocessing subsystem provides some low-level video processing such as key frame extraction, feature extraction, and feature storage. The video query subsystem interfaces between the user and the index and displays the video

content that meets various users' search requirements. In the framework, the extraction of semantic concepts are based on the features extracted from the key frames following the lens boundary detection. A good feature of the framework is its grouping of the key frames into scenes to form a story in order to facilitate better semantic description. However, the framework is still limited by what can be extracted from the multimedia low-level features. Rashid, Niaz, and Bhatti (2009) proposed an improved framework that featured a three-layered framework comprising mode, model, and modality. The framework is presented in Figure 3.1.

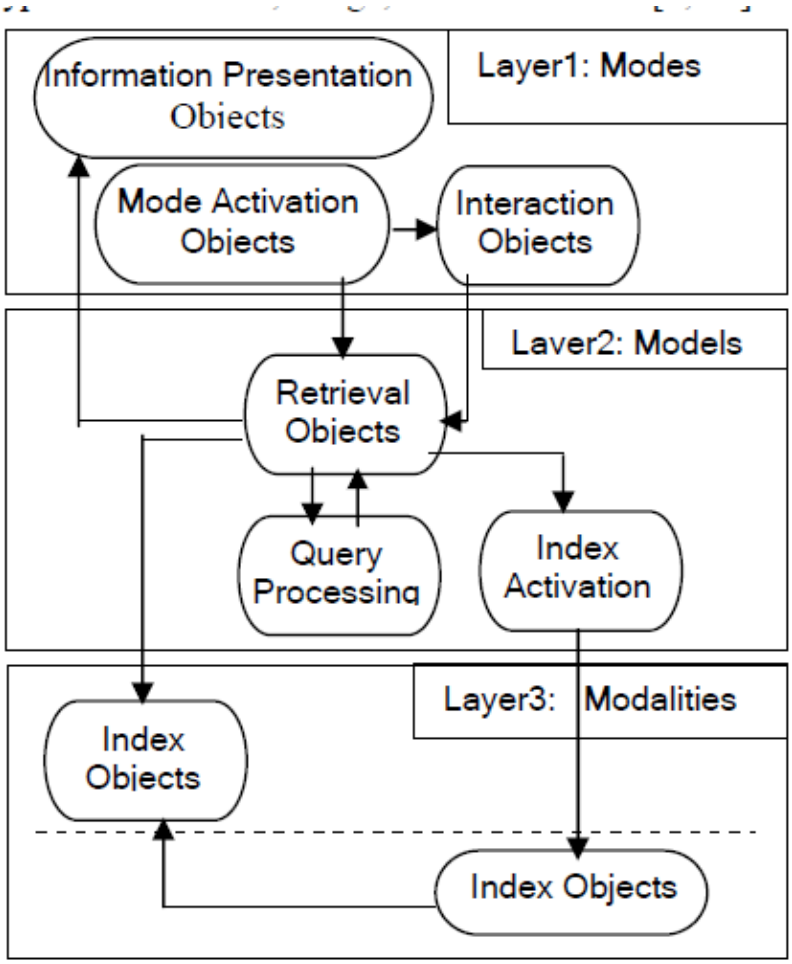

Figure 3.1: Proposed framework for multimedia information retrieval by Rashid, Niaz, and Bhatti (2009)

Mode represents the user interface and provides the means to interact with multimedia objects. Model co-ordinates the activities of the modality and mode layers. It takes user queries passed on from the mode layer and performs a search within index objects in the modality layer, returning the query result to the upper layer. Rashid, Niaz, and Bhatti (2009) argues that their proposed modular approach can satisfy users' information need when information is searched within all three proposed information modalities. The challenge of concept disambiguation in multimedia retrieval frameworks has resulted in the emergence of various frameworks that rely on ontologies for modelling and representing semantic concepts. One of such frameworks proposed by Sokhn *et al.* (2011), is a multimedia information retrieval system adapted to conference media. Its main unique features are the metadata management module and the ontology-based model that describes and structures the information conveyed within a conference life cycle. The framework is presented in Figure 3.2

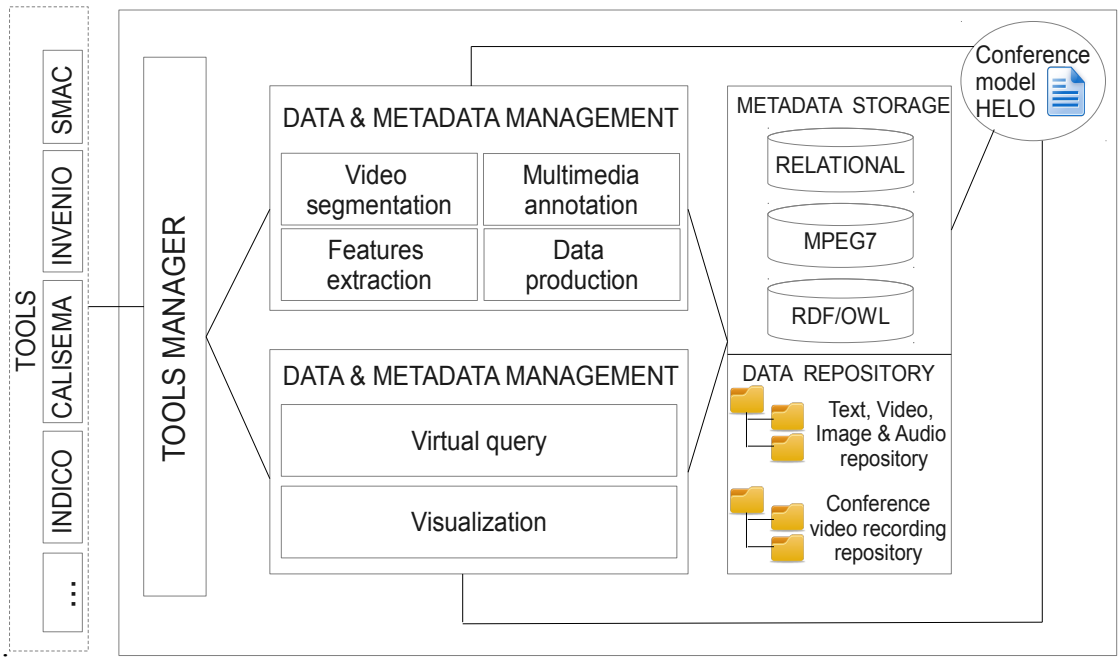

Figure 3.2: Proposed MIR framework by Sokhn *et al.* (2011)

The proposed framework by Sokhn *et al.* (2011) provides support for different multimedia content types which are indexed as MPEG-7 metadata. There are several other researchers that propose an ontology-based approach to multimedia annotation (Khan and McLeod, 2000; Schreiber *et al.*, 2001; Paz-Trillo *et al.*, 2004; Meland *et al.*, 2003; Petridis *et al.*, 2006; Polydoros, Tsinaraki, and Chirstodoulakis, 2006; Gómez-Romero *et al.*, 2011). While most of the existing frameworks and systems differ in their approach to multimedia annotation and retrieval, they all have the multimedia feature extraction and retrieval subsystem, and the data repository for storing indexed objects or metadata. The core differences amongst the existing frameworks lie in how they approach the mapping between multimedia low-level features and semantic concepts, and how they are eventually represented. Whilst some MIR frameworks such as S-CREAM (Handschuh, Staab, and Studer, 2003), KSPACE (Izquierdo et al., 2007), BOEMIE (Paliouras, Spyropoulos, and Tsatsaronis, 2011), and VisionGo (Luan, Zheng, Wang, and Chua, 2011) provide significant improvement in achieving varying levels of automated semantic concept detection from visual features; they have not been able to achieve semantic concept detection that matches the human perceptual level (Enser et al., 2007; Duygulu and Bastan, 2011).

Another difference among existing MIR frameworks is the processing of queries issued by users and the mapping to the appropriate stored metadata in order to satisfy users' search interest. MIR frameworks should provide support for users to pose knowledge level queries (see Figure 2.5, in section 2.3) that reflect human perceptual expectations based on their search needs and be assured of good retrieval results.

Several components have been identified from the review of existing frameworks as necessary parts of a multimedia retrieval system. However, among all reviewed frameworks there is lack of the provision of an integrated non-domain specific

framework that includes effective learning of high-level semantics, friendly user interface, efficient indexing tool, and semi-automated multimedia information retrieval at the knowledge level. The next section presents the limitation of the state of the art which helps to formulate the framework and identify its key components.

3.3 Limitations of the State of the Art

Research in multimedia information retrieval have over the years been tackling the problem of mapping computable visual information features to high-level semantic concepts (Antani, Kasturi, and Jain, 2002; Uren *et al.*, 2005; Datta *et al.*, 2008; Dasiopoulou *et al.*, 2011; Hu *et al.*, 2011). Various frameworks and models (reviewed in Chapter 2 and section 3.2) have been proposed to address the issue. While they have made some worthy progress, the success is limited and necessitates further investigation (Dasiopoulou *et al.*, 2011). The major limitation of the state of the art is the lack of semantic labelling of the multimedia features at the expected level of human perception for a given situation (Duygulu and Bastan, 2011; Hu *et al.*, 2011; Wang and Hua, 2011).

Several frameworks have proposed manual annotation of visual information. Such frameworks include Virage Media Management System (Bach *et al.*, 1996) and M-OntoMat Annotizer (Petridis *et al.*, 2006), which is part of the CREAM (Handschuh, Staab, and Studer, 2003) framework. The annotation is usually done by domain experts who provide textual descriptions to the visual information. It is important that an annotation system that supports manual annotation should provide a simple and intuitive user interface. Such a framework should provide a user interface that supports posing of queries at the human perceptual level and also render multi-user interaction

for collaborative annotation. However, the rate at which multimedia content is generated on the web is so high, making it unrealistic to apply manual annotation to the huge visual information content on the web. In addition to the inefficiency of manual annotation, semantic description from humans can be error-prone and subject to individual bias (Bloehdorn *et al.*, 2005; Kender and Naphade, 2005; Duygulu and Bastan, 2011; Fauzi and Belkhatir, 2013).

An automated annotation framework is an important direction to achieving the annotation of the huge multimedia content on the web. Frameworks like BOEMIE[28] (Paliouras, Spyropoulos, and Tsatsaronis, 2011), VIDI-Video[29] (Papadopoulos *et al.*, 2009), RUSHES (Zhang and Izquierdo, 2011), MediaMill[30], CuZero[31], and VisionGo (Luan, Zheng, Wang, and Chua, 2011) all provide varying levels of automated semantic annotation. Semantic objects are generally identified using different methods to train detectors to match multimedia low-level features to semantic objects (Mylonas *et al.*, 2008; Papadopoulos *et al.*, 2009; Smeaton, Over, and Kraaij, 2009; Zhang and Izquierdo, 2011; Wiliem e*t al.*, 2012). However, Duygulu and Bastan (2011) observed that the supervised training of the semantic detectors can be subjective and error-prone. There is still need to provide frameworks and models that are able to achieve visual information semantic concept understanding that closely match the human perceptual level. Also, the variety of these semantic concepts means that they have to be built according to different domains of use (Duygulu and Bastan, 2011) and lack interoperability. A large-scale framework such as a multimedia annotation framework on the web, will require a uniform standard for sharing domain knowledge (Tzouvaras, Troncy, and Pan, 2007; Tous and Delgado, 2010; Pino and Di Salvo, 2011).

28 http://www.boemie.org
29 http://www.vidivideo.info
30 http://mediamill.cla.umn.edu/mediamill
31 http://www.ee.columbia.edu/ln/dvmm/cuzero/

The emergence of social media sites like Facebook[32], Flickr[33], and Youtube[34] has made collaborative video or image description on the web more prominent by means of social tagging. Users of these social media while tagging the same video, often provide varying comments, which is an expression of their individual perception of the video (Quelhas *et al.*, 2007). Such a collaborative annotation framework requires a scalable multimedia framework that can perform well under heavy usage.

Towards addressing the semantic gap between the visual features and high-level semantic perception, *context* has recently been used in multimedia retrieval to model user's search requirements (Ingwersen and Jarvelin, 2005; Borlund *et al.*, 2008; Goker, Myrhaug, and Bierig, 2009). Nunes, Santoro, and Borges (2009) posits that an implementation of context information management in a computational environment, may act as a filter that defines, at any given moment, which piece of knowledge will be taken into consideration in helping perform tasks. The next section investigates the *context* approach to building an efficient and effective integrated framework capable of managing the annotation and retrieval of multimedia at the semantic level.

3.4 The Multimedia Context Model

The generic multimedia context model is presented to facilitate the integration, understanding, and discovery of multimedia semantics. This section disambiguates the notion of "*context*" and provides a definition with regards to multimedia semantics. The design requirements and principles are presented in section 3.4.2; section 3.4.3 presents the conceptual framework; section 3.4.4 presents the background and motivation leading to the formalisation of the context model; while section 3.4.5 presents the

32 http://www.facebook.com
33 http://www.flickr.com/
34 http://www.youtube.com

formalisation process of the context model.

3.4.1 What is Context?

Context has been identified from the literature (refer to section 2.4) to play a crucial role in human knowledge representation, reasoning, and perception (Huang, Eze, and Webster, 2006; Eze and Ishaya, 2007; Nunes, Santoro, and Borges, 2009). While context has been widely applied in user and query modelling (Ingwersen and Jarvelin, 2005; Natsev *et al.*, 2007; Borlund *et al.*, 2008; Goker, Myrhaug, and Bierig, 2009; Ruthven, 2011; Biswas, 2012) to improve retrieval results, this thesis approaches context from a knowledge perspective (Huang and Tao, 2004; Huang, Eze, and Webster, 2006; Dalakleidi *et al.*, 2011) towards semantic multimedia understanding and representation. It is expected that applying context at both the content representation and retrieval phases will greatly enhance semantic multimedia retrieval. Context is regarded as any information about the multimedia that enhances the semantic understanding of the multimedia content.

Definition:

> ***The context of an entity (i.e. an object, an event, or a concept) is a collection of semantic situational information that characterises the entity's internal features or operation and external relations under a specific situation***.
>
> (Huang, Eze, and Webster, 2006)

Context can be derived either from the internal features of the multimedia content, *implicit context*; or from resources external to the multimedia itself, *explicit context* (Belotti *et al.*, 2004; Goker, Myrhaug, and Bierig, 2009).

Typical multimedia contextual sources on the web include:

- literal statements, such as free semantic annotation of multimedia resources
- knowledge sources, such as information from webscraping (Fauzi, Hong, and Belkhatir, 2009; Alcic and Conrad, 2010), like surrounding text associated with where the visual data is found, filename, page title, ALT tag description from web pages, etc.
- transcription of the audio part of video files (Heryanto, Akbar, and Sitohang, 2011)
- entity's properties and general descriptive metadata (Dublin Core Metatdata[35]), such as author, title, date of publication, etc.
- inference and deductions based on a well-defined ontology
- multimedia features such as objects that can be detected using object recognition techniques.

Multimedia information retrieval systems need the ability to represent, utilise and reason about context to help improve semantic representation and management of multimedia resources (Eze and Ishaya, 2007).

3.4.2 Design Requirements

This section presents various aspects of the framework developed in this study. Other architectural support and integration of components towards multimedia semantic understanding, representation, and retrieval are discussed. The requirements and principles are drawn from the analysis (refer to sections 2.5 and 3.3) and review of literature on semantic multimedia indexing and retrieval, presented in chapter 2.

35 http://dublincore.org

3.4.2.1 Semantic Understanding

A common problem that cuts across all the reviewed frameworks and systems is the lack of semantic understanding at the human perceptive level (Hare *et al.*, 2006a; Hider, 2006; Enser *et al.*, 2007; Natsev *et al.*, 2007). They are restricted to semantic concepts that can be gleaned from the visual features and are limited in their ability to recognise abstract concepts or emotions (Enser *et al.*, 2007; Mylonas *et al*., 2008; Snoek and Worring, 2009; Duygulu and Bastan, 2011). Concepts like "Love" or "Beauty" cannot be easily matched to any set of visual information. The lack of semantic understanding at the human perceptual level is the key requirement that the context model seeks to address. Other requirements are discussed.

3.4.2.2 Interoperability

In this present world of semantic web and collaboration, interoperability has become a fundamental requirement for the automatic processing and discovery of resources from heterogeneous sources. Annotations and metadata should be represented in a standard format that can interoperate with other systems thereby fostering knowledge discovery and sharing. Also, the framework should be able to discover semantic knowledge from other systems either through software agents, web service, or direct interface. The multimedia semantics incubator group[36] have identified interoperability among metadata, ontologies and other services as a key issue among multimedia annotation frameworks (Tzouvaras, Troncy, and Pan, 2007).

3.4.2.3 Automation

Annotation and indexing for multimedia content on the web is almost unrealistic or unusable if attempted to be done manually. The process is highly demanding and will take years for humans to manually annotate after which the data would have become

36 http://www.w3.org/2005/Incubator/mmsem/

obsolete due to the dynamic nature of the web and new content being generated by the second. Manual annotation is not only inefficient but could be error-prone and subjective (Kender and Naphade, 2005; Hare *et al.*, 2006a; Duygulu and Bastan, 2011). The system should be able to support an automated approach to semantic understanding and organisation in order to ensure efficiency and minimise errors (Bloehdorn *et al.*, 2005; Duygulu and Bastan, 2011; Hu *et al.*, 2011). There is also the design implication to provide support for ontologies or concept detectors as a means to achieving automation (Hollink, Worring, and Schreiber, 2005; Uren *et al.*, 2005; Dalakleidi *et al.*, 2011; Dasiopoulou *et al*., 2011).

3.4.2.4 Usability

Good software solutions may not be used to achieve their desired objective if the user interface is poorly designed. Usability is therefore very key in a multimedia retrieval framework like in any other system design (Grudin and Bargeron, 2005). There should be support for good, intuitive, unambiguous interface to both humans (where manual processing is supported) and machines (where processing is automated) in order to ensure usability, reliability, and efficiency. The interface should provide support for concurrent visual data annotation by users, while ensuring a harmonised annotation metadata storage.

3.4.2.5 Adaptability

The systems should be agile enough to accommodate different behaviours and respond appropriately based on various dynamic information acquired in the course of time about its users and environment. The automatic multimedia learning and discovery component should be dynamic enough to learn and reason about context (Uren *et al.*, 2005; Eze and Ishaya, 2007; Dalakleidi *et al.*, 2011).

3.4.2.6 Scalability

Performance relative to size, is always very important in any system design. Multimedia processing is a resource-intensive operation in terms of processing power and volume of multimedia resources. It is crucial that the framework performs well with an increase in multimedia resources (Snoek and Worring, 2009). Ruger (2011) and Pino and Di Salvo (2011) also identified the need to apply more scalable technologies to improve multimedia semantic indexing and retrieval systems.

3.4.3 The Conceptual Framework

The framework is divided into two major parts - content representation and retrieval subsystems. The content representation subsystem entails the extraction and representation of knowledge from video resources using the context model, a set of defined ontologies, and inferences or rules. The extracted metadata is represented in XML/RDF format, organised and indexed in a database. The retrieval phase relies on the context model to return the most appropriate result to the user query based on pre-indexed metadata.

The framework in Figure 3.3 supports semi-automated semantic visual information annotation and retrieval; and has an interface layer which comprises the user interface for human users and a web service for agents to access and interoperate.

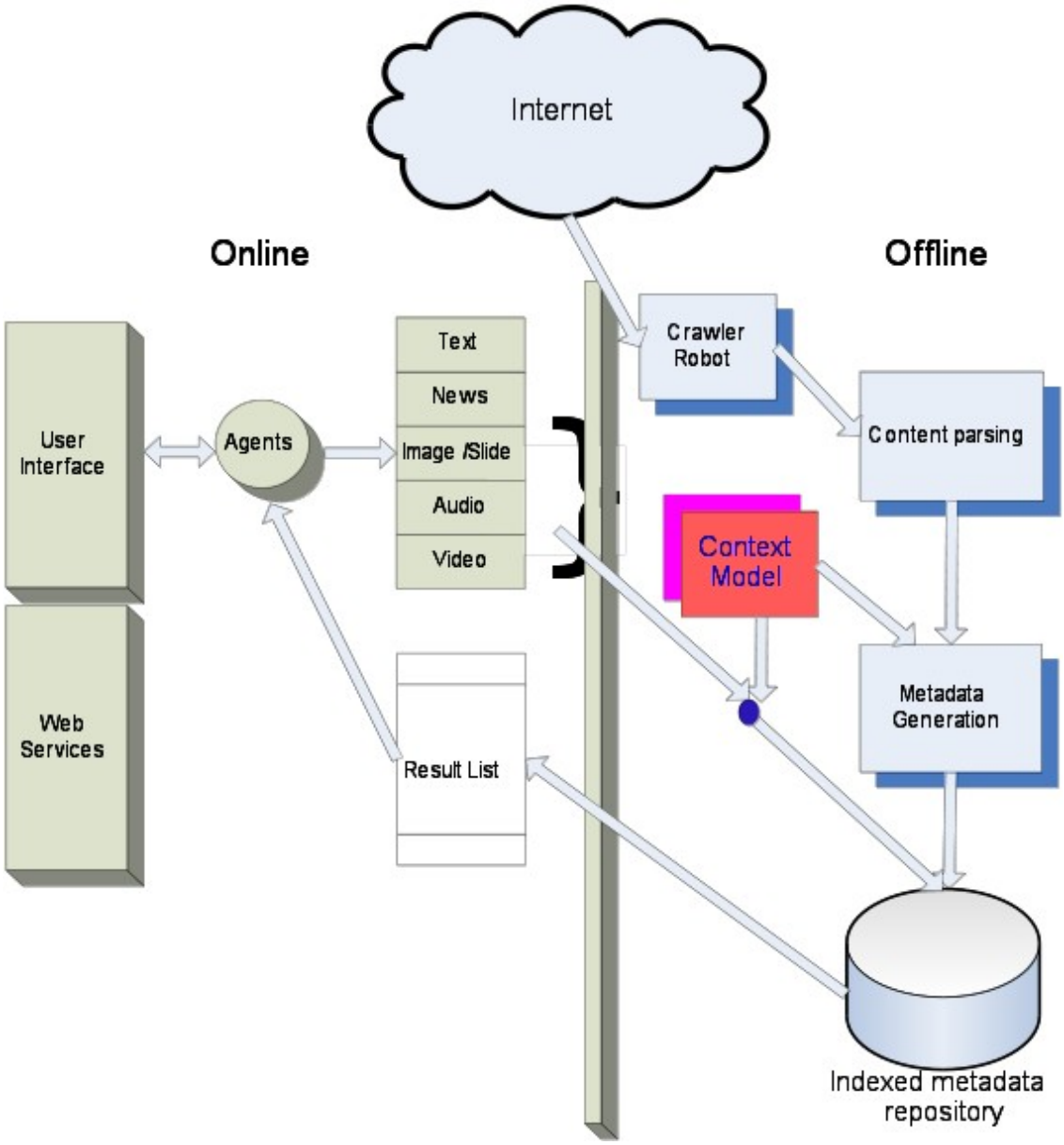

Figure 3.3: Framework for Context-based Multimedia Management

Within the content representation subsystem, visual data from various sources on the web are identified, parsed, and metadata generated and indexed with support from the context model. There are techniques for extracting descriptive texts for visual information on web pages such as the page segmentation technique proposed by Zhai and Liu (2005); Petasis *et al.* (2008); and Fauzi, Hong, and Belkhatir (2009). Alcic and

Conrad (2010) evaluated various extraction techniques and provided a benchmark. The identification of the visual data on the web is achieved with the *crawler robot* component, which passes the visual data location to the *content parsing* component. The *content parser* performs some low-level feature extraction (see section 2.1.1) and also generates metadata in the form of XML/RDF about the resource by applying the *context model*. The metadata is indexed in the *metadata repository*, which is essentially a data store. The *context model* is at the centre of the framework providing support to both the content representation and retrieval subsystems.

The basic idea about the *context model* is to help provide a well-rounded and consistent view of visual data despite varying levels of users' perception (Spink and Cole, 2005) about such visual data. This is like the proverbial blind men and the elephant, where a group of blind men felt different parts of an elephant's body and came up with different recognitions of their perception of the elephant. The context model strives to extract semantic information from various sources. The background to context modelling and formalisation of the model are presented in sections 3.4.4 and 3.4.5 respectively.

The retrieval subsystem relies on software agents to manage the interaction between the client and the *indexed metadata repository* so as to transform the stored metadata into a presentable and understandable format to users. Software agents can generally be described as computer software capable of flexible autonomous actions in a dynamic, unpredictable and open environment (Luck *et al.*, 2003). There are formalisms and architectures for building agents which any piece of software that implements an agent must comply to as expressed by Wooldridge and Jennings (1995). Research in this area is already advanced and the focus of this thesis with regards to agents is on the application rather than on dealing with issues associated with agent architecture or framework. Sarangi and Panda (2010) has detailed background information, issues, and review of the state of the art on agents. With the emergence of new technologies like web services[37] (which allow application to application communication over the

37 http://www.w3.org/2002/ws/

Internet) and the semantic web[38] (which provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries), the web is being transformed into an environment where software components can freely interact. Intelligent information agents are already successfully being used for information filtering and gathering as a way to manage the issue of information overload on the web as presented by Klusch (2001). There are already many existing software tools or languages for building agents[39]. The goal of agents in this research is to facilitate knowledge sharing by automating knowledge discovery and multimedia knowledge representation. One key challenge with modelling agents from this research point of view is the issue of defining a uniform and interoperable communication pattern. It is expected that the use of ontology will provide a means to standardise the agent communication (refer to section 2.2.2). Ontology is very important for agent communication since it explicitly specifies the terms they must understand and use during communication.

The context-based multimedia management framework relies on the identification and use of contextual information about the multimedia resources to enhance semantic understanding of such multimedia content. Most of the components of the framework are common to other frameworks reviewed in section 3.2. However, the unique feature of this framework lies in the integrated, domain-independent approach to visual data from heterogeneous sources on the web; and the *context model*, which impacts on both content representation and retrieval subsystems. It is expected that the injection of the *context model* will improve the overall semantic annotation and retrieval experience. The next section presents the background and motivation for the *context model*.

38 http://www.w3.org/2001/sw/
39 http://www.agentcities.org/Resources/Software/ or http://www.agentlink.org/resources/agent-software.php

3.4.4 Towards Context Modelling

One of the main problems with context modelling in computer systems is lack of simple representation of context, and efficient algorithms for the extraction of contextual information. Humans generally apply the notion of context in their everyday activity as part of their cognitive processing (Belotti *et al.*, 2004). For example, if a man is overlooking the Scarborough Sea from the top of the Castle and he sees a seemingly tiny object moving consistently on the surface of the water, he can possibly recognise this object as a ship or boat. He could make this assumption based on the knowledge he has about his surrounding, circumstance and perspective. Similarly, the multimedia context model would provide an improved semantic description of multimedia content based on other sources of information about the multimedia content. This is referred to as *context*. The man in the analogy knows that a sea animal will not move consistently on the surface of the water for a long time. He recognised the object as a ship, even though he could not see it very clearly, based on other knowledge he has about the sea. Such knowledge is referred to as *common knowledge* (Loyola, 2007). It can be inferred from this analogy that context relies on common knowledge to attain better and clearer semantic description of objects or events.

Multimedia resources can be regarded as a symbolic system. Context in multimedia resources may be likened to *Semiotics*, the study of symbolic systems as presented by Singh (2002). Semiotics has three parts:

1. *Syntax*, or structure
2. *Semantics*, or structured-based meaning
3. *Pragmatics*, or context-based meaning

The concept of semiotics can be applied to multimedia semantic content modelling because it allows for a systematic modelling of the constituent attributes. Syntax can be likened to the multimedia low-level features like colour, texture, and shape. Semantics

(or meaning as captured in syntax), refers to identified objects within the multimedia content and how they interrelate, while pragmatics can be likened to context-based semantic meaning in a human perceptual form. Goker, Myrhaug, and Bierig (2009) identified the need for theory about context and its structures that will facilitate the development of improved frameworks and systems, as a major motivation for modelling context. A good discussion and review of context models in IR is presented by Belotti *et al.*, (2004), Goker, Myrhaug, and Bierig (2009), and Ruthven (2011). While there are research efforts in modelling context from a query and user interaction perspective (Gross and Klemke, 2003; Belotti *et al.*, 2004; Ingwersen and Jarvelin, 2005; Borlund *et al.*, 2008; Ruthven, 2011), this thesis limits the discussion on context modelling as it relates to multimedia semantics.

Mylonas *et al.* (2008) proposed the taxonomic context model, to interpret the meaning of multimedia documents. Context was modelled as a fuzzy relational algebra and regarded as the common meaning that different concepts share. For example, basketball is considered the context of ball, referee and basket since it is a common antecedent for all three concepts. The provision of more concepts into the taxonomic context model results in the narrowing of the context. The taxonomic context model proposed by Mylonas *et al.* (2008) seem to focus more on implicit context (refer to 3.4.1 for the definition of implicit and explicit context) to help disambiguate multimedia document meaning. Yi, Peng, and Xiao (2012) proposed a temporal context model as a probability function to mine information between video shots towards improving video annotation. There are four key context formalisation approaches in the literature: Key-Value, Markup-based, Logic-based, and Ontology-Based models. Key-Value models are the most basic form of modelling context and represents context as a collection of key-value pairs. The key represent a unique identifier for the context, while the value represent the actual context in that instance. The Markup-based model improves on the key-value model through the use of markup languages (like XML) to represent context and separates the context structure (schema) from the content (document instance).

Markup-based context models are highly portable and can be extended. However, they are hardly adaptive to new contextual relationships due to the complexity of extending the schema (Pérez *et al.*, 2009). The more recent Logic-based models (McCarthy and Buvac, 1998; Theodorakis and Spyratos, 2002; Mylonas *et al.*, 2008; Wiliem, Madasu, Boles, and Yarlagadda, 2012) provide a sound formality in the definition and representation of the context through well-formed propositions that are grounded in mathematical logic. Logic-based models support reasoning and inference about context, based on defined rules. Ontology-based models (Nunes, Santoro, and Borges, 2009; Gómez-Romero *et al.*, 2011) involve an explicit definition of the relationship among concepts in a knowledge-base accessible to reasoning engines or interpreters. Ontologies provide a framework for sharing precise meaning of concepts and supports reasoning about context based on the relationship among context data. While ontology-based models disambiguate meaning, they are usually domain specific. Refer to section 2.2.2 for a discussion about ontology. Loyola (2007) and Pérez *et al.* (2009) provide a good analysis of approaches towards formalising context.

Context modelling in video semantics is usually viewed as a concept classifier or detector problem (Wiliem, Madasu, Boles, and Yarlagadda, 2012). There are generally two types of context specific to video data: spatial and temporal context (Jiang *et al.*, 2009; Wiliem, Madasu, Boles, and Yarlagadda, 2012; Yi, Peng, and Xiao, 2012). Spatial context represents visual concept relationship within a single shot, while temporal context describes visual concept relationship and dependency between continuous shots (Yi, Peng, and Xiao, 2012). However, this thesis focuses on another form of context which is referred to as: *semantic context*. Semantic context unlike spatial or temporal context, are general-purpose representation that describes the visual concept relationships in visual data.

The possible sources of contextual information in a football video include text annotation (teams playing, match statistics, score line, etc.), speech (match commentary), video source (a lot of information about a video clip can be gathered

from the web page where it is found), etc. This is not exhaustive as the context model is open and extensible, and therefore can accommodate the addition of further relevant contextual dimensions in video data. The context data is stored in such a way that higher level facts can be inferred from the individual pieces of context data. An ontology knowledge base of explicit concepts is maintained to facilitate mapping and inference deduction from the recognised contexts. The contextual information is represented as metadata using XML and RDF. RDF has a specific declarative semantics, which is specified independently of any RDF processor. This made it more suitable for modelling context. XML is used for basic resource content description, since its meaning is only determined by the actions that programs undertake on it. The RDF/XML combination helped to facilitate interoperability of the model.

3.4.5 Context Formalisation

The previous section reviewed different approaches towards modelling context. This thesis approaches the context modelling by combining two context modelling approaches namely: Markup-based and Logic-based models. Markup is required for the metadata semantic representation; while logic is required to define the relations and rules and also concept understanding (refer to the S-Space model presented in Chapter 4). Even though ontology-based models have been considered the most advanced in modelling context in the literature (Loyola, 2007; Pérez *et al.,* 2009), the focus of the context model is on semantic understanding and representation rather than just disambiguation. The uniqueness of this model lies with the integrated approach of combining visual information representation with context harnessed from various knowledge sources (see section 3.4.1). This is unlike other context models (Jiang *et al.*, 2009; Wiliem, Madasu, Boles, and Yarlagadda, 2012; Yi, Peng, and Xiao, 2012), in the video semantic retrieval domain which focus only on semantic concept classification or detection. Whilst other works focus only on either explicit context (Gross and Klemke, 2003; Belotti *et al.*, 2004; Ingwersen and Jarvelin, 2005; Borlund *et al.*, 2008; Ruthven,

2011) or implicit context (Mylonas *et al.*, 2008; Jiang *et al.*, 2009; Nunes, Santoro, and Borges, 2009; Gómez-Romero *et al.*, 2011; Wiliem, Madasu, Boles, and Yarlagadda, 2012), this work derives from both implicit and explicit contexts. The implicit context (e.g. semantic objects based on extracted features, audio transcription, some Dublin Core Metatdata set that can be extracted from the video properties) come from features of the visual data itself, while the explicit context (e.g. literal statement from annotation or tags, surrounding descriptive text associated with where the visual data is found, filename, page title, ALT description from web pages) are gathered from the web.

A generic context mediation model is presented to facilitate the integration, understanding, and discovery of multimedia semantics. This requires rigorous specification to decompose the complexity and ensure unambiguous specification that could lead to system error or wrong implementation. The multimedia semantic context descriptive model (CON) is formalized using Feature Notation (Scheurer, 1994, pp. 410 – 431), which is a mathematically-based technique. Feature Notation (FN) is an extension of traditional mathematical notation that combines basic models (such as sets, relations, functions, etc.) to form more complex models through a modular, systematic, and rigorous specification process. FN can be applied throughout the development and specification of a system to precisely and rigorously describe the system. It involves validation and verification at each stage to ensure the completeness, correctness, and consistency of the specification (Scheurer, 1994; Kaur, Gulati, and Singh, 2012).

In general, a theoretical model does not seek to describe a single object but a whole class of objects. Such models have many features and often involve not only a certain class of objects but also certain operations on its members each of which is individually named. FN format is similar to Z schemas but highly expressive and simple. FN extends the traditional notation of set theory by a few simple rules, found necessary for the development of large-scale models. These rules are called Feature notation, used for systematic construction of objects and their features (Scheurer, 1994). The Feature Notation is preferred to other traditional mathematical notations due to its established

modular nature, which therefore makes reusability and extensibility easier. Although, the Feature Notation format is similar to the schemas of the Z Notation Language, there are important differences between both notations. A Feature Notation format gives a name to a class of objects, as a Z schema does, but also makes a clear distinction between a) the name of the class and b) the names of variables ranging on the specified class, of which there may be many (Scheurer, 1994). A variable is always symbolically distinct from the class over which it ranges. This distinction is not so clearly emphasized in Z case studies, where a schema name is interpreted as both a certain class and an object in the class (Scheurer, 1994).

The multimedia semantic context descriptive (CON) model consists of the various members defined below:

Given a set of multimedia objects, X; let s represent semantically isolated segments in the object $x \mid x \in X$ (i.e. shot/track in video/audio, parts in image or text).

EP is the set of all possible Extracted Properties. Various extractable media information like media type, media format, number of frames, and file size are all members of this set. Only two conditions are imposed on *EP*: i) this set may not be empty. This constraint was designed to rule out the possibility of not having some basic media information; ii) this set is a disjoint subset of *RD*.

<u>*EP*: Set</u>	set of extracted properties
$EP \neq_{\mathrm{d}} \varnothing$	
IsDisjoint(EP, RD)	$EP \subset_{\mathrm{d}} RD$

ARD is the set of all possible annotator's resource description. "Annotator" here can be a human user or a software agent. Members of this set include basic descriptive metadata such as author, title, URI, date of publication, etc. This set is also constrained to be non-empty to avoid the possibility of not having at least a description for the multimedia object. This set like the *EP* set, is also a disjoint subset of *RD*.

<u>*ARD*: Set</u> set of annotator's resource description

$ARD \neq_{\mathrm{d}} \varnothing$

IsDisjoint(*ARD*, *RD*) $ARD \subset_{\mathrm{d}} RD$

KS is the set of all possible contextual or knowledge sources. Members of this set include information from webscraping, audio transcription for video resource, ontology knowledge base, multimedia features, etc. This set is constrained to be non-empty and also a disjoint subset of *DS*, as it adds uniqueness to the model and enforces clearer automated semantic discovery and organisation.

<u>*KS*: Set</u> set of all contextual or knowledge sources

$KS \neq_{\mathrm{d}} \varnothing$

IsDisjoint(*KS*, *DS*) $KS \subset_{\mathrm{d}} DS$

ASD is the set of all possible annotator's semantic description. This set comprises free textual annotation of multimedia objects. *ASD* is a disjoint subset of *DS* and can be empty.

<u>*ASD*: Set</u>	set of annotator's semantic description
IsDisjoint(*ASD*, *DS*)	$ASD \subset_d DS$

DS is the set of semantic description scheme. This set is a strict union of the disjoint sets *KS* and *ASD*. Therefore the set inherits all constraints imposed on *KS* and *ASD*. *DS* is clearly non-empty since one of its disjoint subsets (*KS*) is non-empty.

<u>*DS*: Set</u>	set of all semantic description scheme
$KS \cup_d ASD$	union of the disjoint sets *KS* and *ASD*

RD is the set of all resource descriptors. This set is a strict union of *EP* and *ARD*, and thus inherits all constraints imposed on both sets. Hence this set is also non-empty.

<u>*RD*: Set</u>	set of all resource descriptors
$EP \cup_d ARD$	union of the disjoint sets *EP* and *ARD*

The following relations are defined:

1. y RelRD $x \Leftrightarrow_d x \in X \wedge y$ is a resource descriptor for x based on $\forall$ s | s $\in$ *RD*

2. y RelDS $x \Leftrightarrow_d x \in X \wedge y$ is a semantic description scheme for x based on $\forall$ s | s $\in$ *DS*

Finally, the semantics of multimedia object, *x* in context as *CON(x)* can be represented as:

$$CON(x) =_d Ran(RelRD) \cup \sum_{s=1}^{n} Ran(RelDS)_s \qquad where\, n = total\ number\ of\ s$$

The context description model could be represented graphically as shown in Figure 3.4

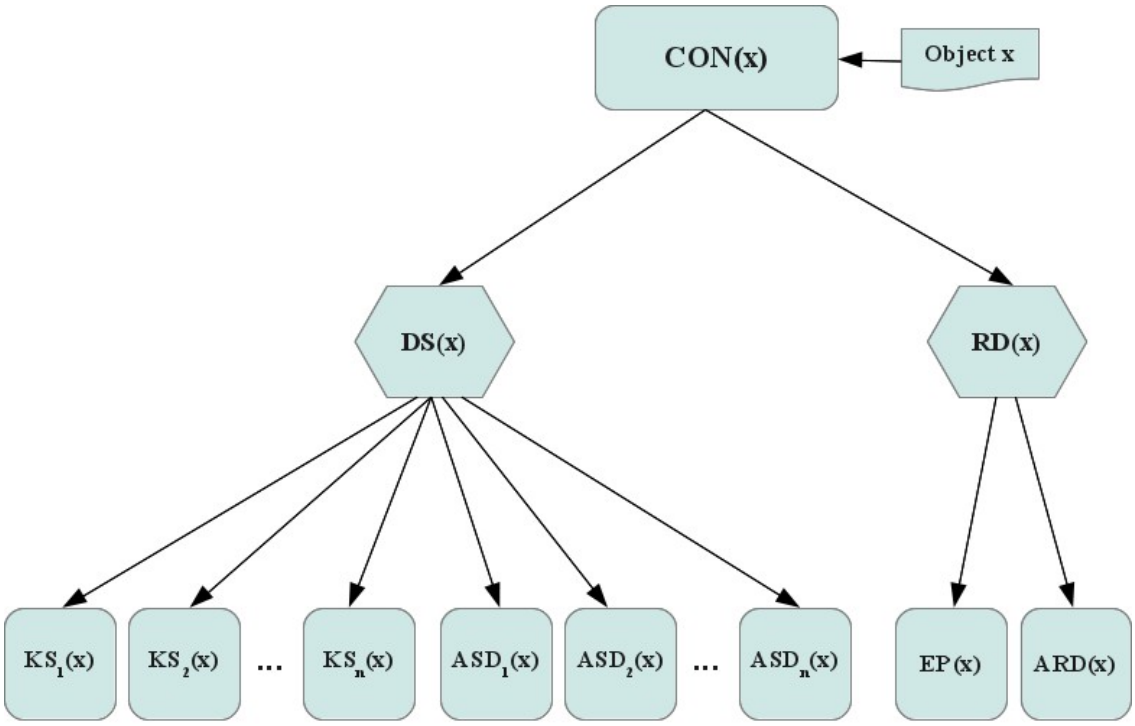

Figure 3.4: Context model for multimedia semantics

Observe that *KS* and *ASD* are segmented. This is necessary since a single multimedia object could have various semantic descriptions based on its various segments. It is therefore necessary that the timeline is captured in the model in order to facilitate contextual retrieval. *RD* does not require segmentation as its members are fixed description for the entire multimedia object *x*, without any reference to the segments. The model is not restricted to bind context description into specific data representation format like RDF or XML. A sample XML schema implementation of the model is presented below.

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema"
targetNamespace="http://cafe.cic.hull.ac.uk/~icr03ee/research/contextmodel"
   xmlns="http://cafe.cic.hull.ac.uk/~icr03ee/research/contextmodel"
elementFormDefault="qualified">
   <xs:simpleType name="stringtype">
      <xs:restriction base="xs:string"/>
```

```
</xs:simpleType>

<xs:simpleType name="inttype">
   <xs:restriction base="xs:positiveInteger"/>
</xs:simpleType>

<xs:simpleType name="datetype">
   <xs:restriction base="xs:date"/>
</xs:simpleType>

<xs:simpleType name="uritype">
   <xs:restriction base="xs:anyURI"/>
</xs:simpleType>

<xs:simpleType name="dectype">
   <xs:restriction base="xs:decimal">
      <xs:minInclusive value="0"/>
   </xs:restriction>
</xs:simpleType>

<xs:complexType name="descriptiontype">
   <xs:sequence>
      <xs:element name="description" type="stringtype" />
   </xs:sequence>
   <xs:attribute name="segmentid" type="inttype" use="required"/>
</xs:complexType>

<xs:simpleType name="KSourceType">
   <xs:restriction base="xs:string">
      <xs:enumeration value="Webscrapping"/>
      <xs:enumeration value="AudioTrans"/>
      <xs:enumeration value="AutoDetect"/>
   </xs:restriction>
</xs:simpleType>

<xs:complexType name="EPtype">
   <xs:sequence>
      <xs:element name="mediatype" type="stringtype" fixed=""/>
      <xs:element name="mediaformat" type="stringtype"/>
      <xs:element name="no_of_frames" type="inttype" default="1"/>
      <xs:element name="filesize" type="inttype"/>
      <xs:element name="mediaduration" type="dectype"/>
   </xs:sequence>
</xs:complexType>

<xs:complexType name="ARDtype">
   <xs:sequence>
      <xs:element name="title" type="stringtype"/>
      <xs:element name="creator" type="stringtype" minOccurs="0"/>
      <xs:element name="subject" type="stringtype"/>
      <xs:element name="description" type="stringtype"/>
      <xs:element name="publisher" type="stringtype" minOccurs="0"/>
```

```
        <xs:element name="contributor" type="stringtype" minOccurs="0"/>
        <xs:element name="date" type="datetype" minOccurs="0"/>
        <xs:element name="identifier" type="stringtype" minOccurs="0"/>
        <xs:element name="source" type="uritype"/>
        <xs:element name="language" type="stringtype" />
        <xs:element name="relation" type="stringtype" minOccurs="0"/>
        <xs:element name="coverage" type="stringtype" minOccurs="0"/>
        <xs:element name="rights" type="stringtype" minOccurs="0"/>
      </xs:sequence>
    </xs:complexType>

    <xs:complexType name="KStype">
      <xs:sequence>
        <xs:element name="KSdescription" type="descriptiontype"
maxOccurs="unbounded"/>
      </xs:sequence>
      <xs:attribute name="source" type="KSourceType" use="required"/>
    </xs:complexType>

    <xs:complexType name="ASDtype">
      <xs:sequence>
        <xs:element name="ASDdescription" type="descriptiontype"
maxOccurs="unbounded"/>
      </xs:sequence>
      <xs:attribute name="id" type="stringtype" use="required"/>
    </xs:complexType>

    <xs:complexType name="DStype">
      <xs:sequence>
        <xs:element name="KS" type="KStype" maxOccurs="unbounded">
          <xs:unique name="testUnique">
            <xs:selector xpath="KS/KSdescription"/>
            <xs:field xpath="@segmentid"/>
          </xs:unique>
        </xs:element>
        <xs:element name="ASD" type="ASDtype" maxOccurs="unbounded"/>
      </xs:sequence>
    </xs:complexType>

    <xs:complexType name="RDtype">
      <xs:sequence>
        <xs:element name="EP" type="EPtype"/>
        <xs:element name="ARD" type="ARDtype"/>
      </xs:sequence>
    </xs:complexType>


    <xs:complexType name="CON">
      <xs:sequence>
        <xs:element name="RD" type="RDtype"/>
        <xs:element name="DS" type="DStype"/>
      </xs:sequence>
```

```
    <xs:attribute name="mediaid" type="uritype" use="required"/>
  </xs:complexType>

  <xs:element name="ContextMediaDescription" type="CON"/>

</xs:schema>
```

The model is designed as a generic model and thus is quite flexible and easily extensible.

3.5 Summary

The research methodology towards developing the framework and model was presented. It involved a review of existing multimedia annotation and retrieval frameworks and the identification of limitations. Necessary components were identified which informed the decisions in formulating the conceptual framework. A rigorous specification process using Feature Notation (as presented in section 3.4.5) was undertaken to formalise the model. Mathematics is essential to formalising models. It ensures consistency, and helps in validating possible implementation in an actual system. Intrinsically, a model of information retrieval serves as a blueprint for the development of an actual information retrieval system.

A context-based framework for the management of multimedia resources have been developed. The key design consideration and principles leading to the development of the framework and model were presented. The formalised integrated model supports multimedia semantic discovery and representation from heterogeneous sources and relies on the identification and use of contextual information about multimedia resources to enhance the organisation and management of multimedia semantics.

The development and formalisation of the context model is a clear demonstration that it is possible to develop a generic formalised model for video semantic management

and representation; by implication validating and verifying the second research hypothesis, H_2. One important aspect of the context model, CON(x) is the knowledge source (KS) which represents the set of all possible contextual or knowledge sources. KS is key to achieving automated semantic understanding which is one of the key requirements (see section 3.4.2.3) in the framework design. The next chapter examines it in more detail and presents the semantic recognition model which is crucial towards achieving automated multimedia semantic annotation.

Chapter 4

Towards Automated Multimedia Semantic Understanding

The development and formalisation of the context-based multimedia management framework was presented in chapter 3. One key aspect of the context model, CON(x) is the knowledge source (KS) which represents the set of all possible knowledge sources. The challenge therefore, is how to computerise these sources in an efficient and effective manner. It was identified that one of the members of the KS set which is very important in automating semantic extraction (see section 3.4.2.3) from multimedia content required further development and formalisation. Thus, in order to demonstrate how KS is represented, an automated semantic recognition model will be developed and formalised.

This chapter is organised in three sections. Section 4.1 sets the background to automated multimedia semantic understanding; section 4.2 presents the design, development and formalisation of the S-Space model; while section 4.3 provides a summary of the chapter and draws conclusion.

4.1 Automating Multimedia Semantic Understanding

It has been shown in the literature (as presented in sections 3.3 and 3.4.2.3) that automation is very important in visual information annotation systems (Snoek *et al.*, 2006; Duygulu and Bastan, 2011; Hu *et al.*, 2011) since it can remove the inefficiency

and subjectivity (Kender and Naphade, 2005; Hare *et al.*, 2006a) of manual annotation. The context model presented in chapter 3 derives from both implicit and explicit context (refer to section 3.4.5) to achieve semantic annotation and representation. The implicit context (e.g. semantic objects based on extracted features, audio transcription, some Dublin Core Metatdata set from the video properties) originates from the internal features of the visual data itself, while the explicit context are extracted from sources (e.g. literal statement from annotation or tags, surrounding descriptive text associated with where the visual data is found on the web, filename, page title, ALT description from web pages) external to the visual data itself.

Context modelling in video semantics is usually viewed as a concept classifier or detector problem (Wiliem, Madasu, Boles, and Yarlagadda, 2012). There are other research efforts (see section 3.4.4) that focus on modelling context towards automated semantic recognition in visual data (Hollink, Worring, and Schreiber, 2005; Dalakleidi *et al.*, 2011; Dasiopoulou *et al*., 2011). Uren *et al.* (2005) provides a survey of related work on automated semantic understanding. Yi, Peng, and Xiao (2012) proposed a temporal context model as a probability function to mine information between video shots towards improving video annotation. While mining information across shots can improve semantic understanding, there is still the question of how the semantic objects are modelled in relation to the visual information features. Mylonas *et al.* (2008) attempts to address the issue by modelling context as a fuzzy relational algebra in their taxonomic context model. Context is used to interpret meaning in multimedia documents and is regarded as the common meaning that different concepts share. They posit that the provision of more concepts into the taxonomic context model results in the narrowing of the context.

However, this thesis has provided for both temporal and spatial context in the context model already presented and summarised as:

$$CON(x) =_d Ran(RelRD) \cup \sum_{s=1}^{n} Ran(RelDS)_s \qquad where\, n = total\ number\ of\ s$$

As shown from the mathematical representation of the context model above, temporal shot segmentation is considered across various knowledge sources and contextual dimensions (i.e. $Ran(RelDS)_s$). It is argued from the perspective that this thesis views context, that more concepts will result in more contexts rather than narrow the context as Mylonas *et al.* (2008) posits. The relationships amongst the objects in the visual information all culminates to context information. For example in a football shot, the distance of a ball from the penalty area, the position of the goal keeper and that of other players, will determine whether the shot represents a penalty or free-kick event.

Similarly, the automated semantic recognition model, S-Space that is presented in the next section, attempts to model the interrelationship of the objects within the video to detect high-level semantics. The design considerations leading to the development of the S-Space model are discussed and the formalised Semantic Space (S-Space) model for automated semantic recognition presented.

4.2 Semantic Space (S-Space) Model

Automation of multimedia content semantic extraction and representation requires a learning stage in which the system learns the relationship between the contextual features and semantic entities in multimedia objects. This section explores the design considerations for automating multimedia semantic recognition and organisation. The

formalised model is presented.

4.2.1 Design Considerations and Motivation

Human ideas are not only conveyed by language but they are actually formed by the available language. This is why objects or drawings that cannot be easily linked with an individual's language will have absolutely no meaning to that individual. Hence, a child who is seeing a personal computer for the first time may likely call it a television set. This is because the word “personal computer” does not exist in the child’s word set. The human mind is inseparable from the functioning of signs. Therefore, machines will have to emulate humans in order to recognise or understand multimedia semantics.

The formal relation of signs to one another can be regarded as *syntactic*; the relation of signs to the objects to which the signs are applicable can be referred to as *semantic*; while the relation of signs to interpreters can be referred to as *pragmatic* or *contextual* (Singh, 2002). A sign has a semantic dimension in so far as there are semantic rules (whether formulated or not is irrelevant), which determines its applicability to certain situations under certain conditions. In emulating humans, the automated semantic recognition model, henceforth referred to as Semantic Space (S-Space) model, should have the following features:

- can remember things
- can learn new things
- can reason and deduct meaning

The formalisation of the S-Space Model is presented in the next section.

4.2.2 Semantic Space (S-Space) Formalisation

A model for multimedia semantic organisation and discovery is introduced and formally defined. The model is termed *S-Space* (Semantic Space). It got its name from its unique ability to clearly decompose, manage, discover, and represent multimedia semantics as though they are enclosed within a space; akin to the human brain enclosed in the skull constantly processing and interpreting thought forms or signals. The model is highly flexible and extensible. The model is formalised using Feature Notation (as presented in section 3.4.5). The next section presents the formal definition of a Semantic Concept (*S-Con*) which is a base class that must be specified towards the definition of S-Space. Section 4.2.2.2 presents the features of *S-Con*; section 4.2.2.3 defines the various operations that can be performed on *S-Con*; section 4.2.2.4 formally defines *S-Space*; while section 4.2.2.5 defines all the operation that can be performed on *S-Space*.

4.2.2.1 Formal Definition of S-Con

A concept on its own has no clear semantics. In order to disambiguate concepts, there is need to clearly identify other contextual information that are related to the concept for a clearer semantic meaning. Hence, the word "kick" has no clear semantic meaning but "free-kick", "Beckham's free-kick", or "David Beckham's free-kick at the FIFA World Cup 2006 in Germany" has clearer semantics. The more information that is available to disambiguate the concept, the clearer the semantics. This additional information is henceforth referred to as *contextual information*. Contextual Information is formally denoted as a function *ConInf()*.

In order to proceed to the formal definition of *S-Space*, the class *S-Con* (Semantic Concept) is introduced and formally defined.

The Class *SC:S-Con* is an object with the following primary variable features: a base set SCNodes and a relation *SCParentOf*. There exists *ConInfo(SCNodes)* for all nodes such that it returns all contextual information that relates to *SCNode*. This is expressed by:

$\forall i$: *SCNodes* $\exists$ SC*ConInfo(i: SCNodes)*

For any two nodes *i*, *j: SCNodes*, '*i SCParentOf j*' means that *i* is the parent of *j* in *SC*. The set of nodes *j* which have a parent is the range *of SCParentOf* that is *SCParentOfRan*. A node *j* without a parent is called a root. The set of roots is denoted by *SCRoots*. It is determined formally by the condition:

$SCRoots =_d SCNodes - SCSParentOfRan$

Also, for any point *j: SCParentOfRan*, there is exactly one *i* such that *i SCParentOf j*. This means that *SCParentOf* is injective. This is expressed by:

IsInjective(*SCParentOf*)

where for any relation R, the predicate IsInjective(*R*) is true iff

$\forall y$: *RRan* $\exists$ 1 *x*: *RDef* $\cdot$ *x* R *y*

SC satisfies a certain inductive principle. This condition ensures that any node *j:SCNodes* is reachable from some root by a 'path' determined by *SCParentOf*. This implies that there must exist a sequence

$(r = i_0, i_1, i_2, \ldots, i_n = j)$

where r is a root ($r \in$ *SCRoots*), $i_0, i_1, i_2, \ldots, i_n \in$ *SCNodes* and i_0 *SCParentOf* i_1, i_1 *SCParentOf* i_2, …, i_{n-1} *SCParentOf* i_n.

This condition is formally expressed as follows:

For any relation R and any set A, let 'A ClosedBy R' be the predicate true iff A is closed by R, that is, for any points x, y such that $x\ R\ y$, if $x \in A$ then $y \in A$.

A ClosedBy R' iff

$$\forall x: A \cap RDef, y: RRan \cdot x\ R\ y \Rightarrow y \in A$$

Any set A is *inductive* with respect to *SCRoots* and *SCParentOf* iff (a) $SCRoots \subseteq A$ and (b) A ClosedBy *SCParentOf*. Therefore the induction principle for *SC*: S-Con reads:

for any inductive subset $A \subseteq_d SCNodes$, $A = SCNodes$. This implies that *SCNodes* is the smallest inductive set.

The formal definition of *S-Con* is thus given in the following definition where '*' denotes a primary variable feature. A graphical illustration is given in Figure 4.1.

<u>*SC: S-Con* (Semantic Concept)</u>

*	*SCNodes*: Set	Base set of *SC*
	$SCRoots \subseteq_d SCNodes$	Set of roots of *SC*
*	$SCParentOf: SCNodes \leftrightarrow SCNodes$	Parent relation
	i SCParentOf j: Bool	True iff *i* is the parent of *j* in *SC*
	(*i*, *j*: *SCNodes*)	

$SCCo1\ \therefore\ SCRoots =_d SCNodes - SCSParentOfRan$

$SCCo2\ \therefore$ IsInjective(*SCParentOf*)

$SCCo3\ \therefore\ \forall i: SCNodes\ \exists$ SC*ConInfo(i: SCNodes)*

SCCo4 $\therefore$ *Inductive Principle*

$\forall A \subseteq_d SCNodes \cdot$

if (a) $SCRoots \subseteq A$

(b) A ClosedBy *SCParentOf*

then A = *SCNodes*

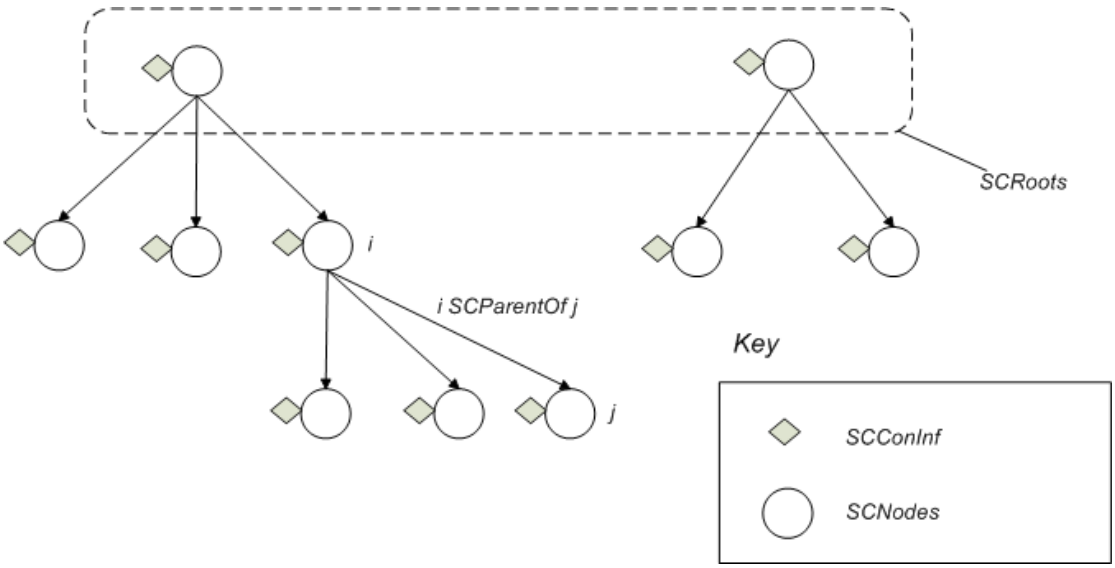

Figure 4.1: Graphical model of *SC:S-Con*

4.2.2.2 Features of S-Con

SC: S-Con has been defined. The following features of *SC*: S-Con that determine the state of SC at any given time are defined:

SCParent: The opposite relation of *SCParentOf* which returns *i: SSNodes* as the parent of *j: SCNodes* where *i SCParentOf j* is denoted by *SCParent* and can be formally defined as:

$SCParent =_d OPP(SCParentOf)$

This feature ensures that any node *i*: *SCNodes – SCRoots*, has a parent which is written as *SCParent*(*j*).

SCChis: For each *i: SCNodes*, *SCChis(i)* is the set of 'children' of *i*. It is defined by:

$SCChis(i) =_d \{ j: SCNodes \mid i\ SCParentOf\ j \}$

SCAtomic: An *atomic* of *SC: S-Con* is a childless node, which denotes some basic concepts and represents the lowest unit towards high-level semantic understanding. All scenes, events and high-level semantic descriptions are derived from a combination of atomics. However, it should be noted that an atomic on its own does not have any clear semantics.

An atomic is formally defined as a point *i: SCNodes* such that $SCChis(i) = \varnothing$.

$SCAtomic =_d \{ i: SCNodes \mid SCChis(i) = \varnothing \}$

There exists *ConInfo(SCAtomic)* for all atomics such that *ConInfo(i: SCAtomic)* returns all contextual information that relates to i: *SCAtomic*. Thus for any atomic *i: SCAtomic* $\exists$ SC*ConInfo(i: SCAtomic)*.

SCAncs: For any node *i: SCNodes*, there exists several nodes *SCParent(i)*, *SCParent(SCParent(i))*, *SCParen*(*SCParent(SCParent(i)))* and so on. These elements are the ancestors of *i*. It can be seen that if i is not a root, they form a finite set, which contains exactly one root of SC. This set is denoted by *SCAncs(i)*. It is a total function

$SCAncs: SCNodes \rightarrow \mathrm{Pow}(SCNodes)$

and is defined recursively by:

(1) $\forall\ r: SCRoots \cdot\ SCAncs(r) =_d \varnothing$

(2) $\forall\ i: SCNodes, j{:}SCChis(i)\ \cdot\ SCAncs(j) =_d SCAncs(i) + \{i\}$

Two related elementary properties of *SCAncs* are stated:

For any *SC: S-Con* and any *i: SCNodes*,

(1) There exists exactly one root *r: SCRoots* such that either $r = i$ or $r \in SCAncs(i)$

(2) $i \notin SCAncs(i)$

SCDecs: For any *i: SCNodes*, the set of 'descendants' of *i* is defined as:

$SCDescs(i) =_d \{ j: SCNodes \mid i \in SCAncs(j) \}, \forall i: SCNodes$

Thus, any *j* is a descendant of *i* iff *i* is an ancestor of *j*, and *SCDesc* is a total function

$SCNodes \rightarrow \mathrm{Pow}(SCNodes)$

SCComposite: A composite of SC is a node that denotes higher-level concepts. This is defined as a node *i: SCNodes* such that $SCChis(i) \neq \emptyset$. *SCComposite* is clearly the opposite of *SCAtomics* and can be formally defined as:

$SCComposite =_d OPP(SCAtomic)$

$SCComposite =_d \{ i: SCNodes \mid SCChis(i) \neq \emptyset \}$

The contextual information of a composite is the aggregation of the contextual information of all descendant nodes. This is formally defined as:

$\mathrm{SC}ConInfo(i: SCComposite) =_d \Sigma\, \mathrm{SC}ConInfo(SCDescs(i))$

The features of S-Con are summarised and illustrated in Figure 4.2.

SC: S-Con Features: Definitions

SCParent(*j*): *SCNodes* Parent of node *j* in *SC*

(*j*: *SCParentOfRan*)

SCDef1 $\therefore$ *SCParent* $=_d$ *OPP(SCParentOf)*

SCChis(i) $\subseteq_d$ *SCNodes* Set of children of node *i*

(*i: SCNodes)*

SCDef2 $\therefore$ $\forall$i: *SCNodes* · *SCChis(i)* $=_d$ { *j: SCNodes* | *i SCParentOf j* }

SCAtomic $\subseteq_d$ *SCNodes* Set of atomic concepts in *SC*

SCDef3 $\therefore$

(1) *SCAtomic* $=_d$ { *i: SCNodes* | *SCChis(i)* = $\varnothing$ }

(2) $\forall$*i: SCAtomic* $\exists$ SC*ConInfo(i: SCAtomic)*

SCAncs(i) $\subseteq_d$ *SCNodes* Set of ancestors of node *i* in *SC*

(*i: SCNodes)*

SCDef4 $\therefore$

(1) $\forall$ *r: SCRoots* · *SCAncs(r)* $=_d$ $\varnothing$

(2) $\forall$ *i: SCNodes*, *j:SCChis(i)* · *SCAncs(j)* $=_d$ *SCAncs(i)* + *{i}*

SCDesc(i) $\subseteq_d$ *SCNodes* Set of descendants of node *i*

(*i: SCNodes)*

SCDef5 $\therefore$

$\forall$i: *SCNodes* · *SCDescs(i)* $=_d$ { *j: SCNodes* | *i* $\in$ *SCAncs(j)* }

SCComposite $\subseteq_d$ *SCNodes* Set of composite semantics in *SC*

SCDef6 $\therefore$

(1) $SCComposite =_d \{ i: SCNodes \mid SCChis(i) \neq \varnothing \}$

(2) $\forall i: SCComposite \; \exists \; SCConInfo(i: SCComposite) =_d$

$\Sigma \; SCConInfo(SCDescs(i))$

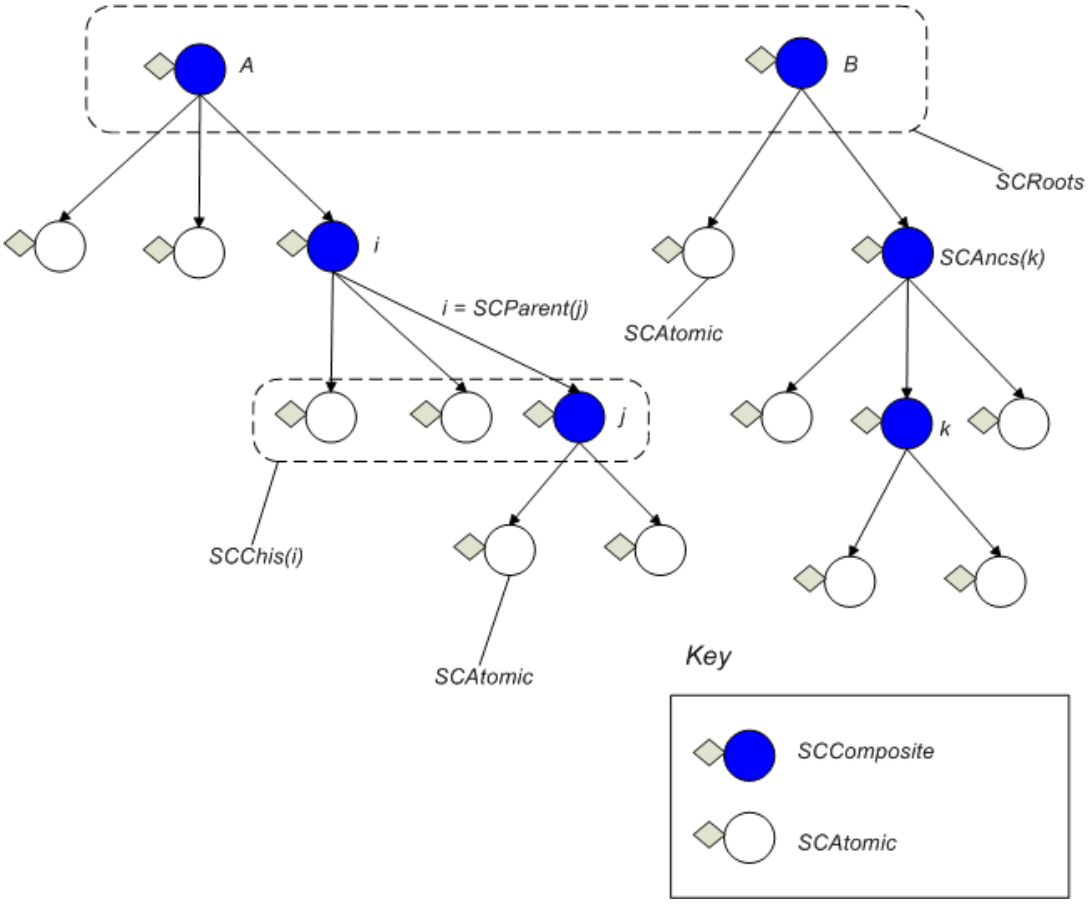

Figure 4.2: Features of SC:S-Con

Figure 4.2 depicts an *SC: S-Con* highlighting most of its features. Nodes A and B are the roots, *SCRoot*. Any number of roots is permitted in an *S-Con.* All the nodes in blue, i.e. nodes with child nodes attached to them, are composite nodes, *SCComposite*. *SCComposite* denote higher-level semantics and as can be seen in the figure, a root, *SCRoot* is also a composite node. The nodes in white are the atomic nodes, *SCAtomic* that represent the basic unit of semantics. *SCAtomic* has no child nodes attached to

them. The node i is the parent of j, i = *SCParent*(j) since the node j is a direct descendant of i as shown. Similarly, the root node, A and the node i are ancestors of j since there is clearly a path from the top that link them to node j. Hence, {A: *SCNodes*, i: *SCNodes*} ∈ *SCAncs*(j). The set of 'children' of node i, SCChis(i) as shown, are the direct descendant nodes of the composite node i. SCChis() is usually applied to composite nodes. An atomic node does not have any descendant node and thus SCChis() on an atomic node will return ∅.

4.2.2.3 Operations on S-Con

Operations on S-Con which is a feature of S-Space are defined.

Emp

Emp returns an empty *S-Con* without a root or any nodes.

SCNodes = ∅

The formal definition of *Emp* is given:

Emp

Return the empty *S-Con*

Emp $=_d$ *SC′* where *SC′ : S-Con* and

(1) *SC'Nodes* $=_d$ ∅

Ins

Given a semantic concept *SC*: *S-Con*, a node *n: SCNodes*, and a composite object C: *SCComposite*, this operation returns a new *SC′*:*S-Con*, which is the result of inserting *C* in *SC* immediately above *n*, that is, with *n* a child of *C*. Thus, Ins(*SC*, *n*, *C*) $=_d$ *SC′* where *SC′*:*S-Con* .

The result *SC'* of Ins(*SC*, *n*, *C*) is well defined iff *SCNodes* and *CNodes* are disjoint. For if *SCNodes* and *CNodes* had some nodes in common, the union of *SCParentOfGr* and *CParentOfGr* would not in general be the graph of a parent function. Therefore, a precondition is imposed on the Ins(*SC*, *n*, *C*) operation: *C: SCComposite* must not already exist in *SC*, i.e. IsDisjoint(*SCNodes*, *CNodes*). This restriction is adopted to avoid redundancy and it is explicit in the model. This restriction also reflects the assumption that seeking to 'insert' a composite object *C: SCComposite* which already exists in *SC* will probably result in an error. Another restriction imposed on this operation is that if $SC = \varnothing$, then this operation reduces to: Ins(*SC*, *C*) and the resulting *SC'* will basically have C as its root. In order to allow for a systematic bottom-up construction of semantic concepts, a condition is imposed on node *n; n* must not have an ancestor nodes in *SC*, i.e. $SCAncs(n) =_d \varnothing$. This condition ensures that integrity is maintained during the insertion of *C*. An attempt to insert *C* to any other nodes in *SC* but a node without an ancestor will necessitate an update of the *contextual information*, *ContInfo()* of every node in *SC*. This is clearly an enormous task to perform at every single insert; hence the need for the restriction.

The base set *SC'Nodes* must be the union of the set *SCNodes* and *CNodes*. Also, the graph *SC'Gr* is defined as the union of *SCGr* and { *CNodes* }. The full specification of Ins is given as follows:

Ins

Insert a composite object, *C*: *SCComposite* below a node *n: SCNodes* in a semantic concept *SC*:*S-Con*, where *SCNodes* and *CNodes* are disjoint.

Given *SC*:*S-Con*, *n: SCNodes*, and *C*: *SCComposite* with

(Pre 1) IsDisjoint(*SCNodes*, *CNodes*)

(Pre 2) if $SC = \varnothing$, then $n = \varnothing$, and Ins(*SC*, *n*, *C*) $=_d$ Ins(SC, C) = *SC'*

with C as the root of *SC'*

(Pre 3) $SCAncs(n) =_d \varnothing$

Ins(SC, n, C) $=_d$ SC' where SC':S-Con and

(1)a. $SC'Nodes =_d SCNodes + CNodes$

b. $SC'Gr =_d SCGr + \{ CNodes \}$

Del

The operation Del reverses the effect of Ins. It takes two arguments: a semantic concept SC': S-Con and a composite object C: $SCComposite$, contained in SC'. Given SC' and C, Del(SC' , C) returns SC: S-Con, which is the result of removing C from SC'. The graph of $SCGr$ is defined as $SC'Gr - \{CNodes\}$ and the base set $SCNodes$ is the relative complement of $CNodes$ in $SC'Nodes$. A precondition imposed on Del is that C: $SCComposite$ must exist in SC' i.e. $C \in SC'$ and $CNodes \subseteq_d SC'\ Nodes$ such that all the nodes in C have a common parent in SC'. Another precondition imposed on this operation is that $CRoot$ must be a root in SC', thus $CRoot \in SC'Roots$. This precondition ensures that high-level concepts can be deleted from SC' without causing any distortion or loss of information, thus maintaining integrity during deletion of a composite object. The definition is given as follows:

Del

Remove a composite object, C: $SCComposite$ from a semantic concept SC': S-Con.

Given SC': S-Con and C: $SCComposite$ where

(Pre 1) $C \in SC'$

(Pre 2) $CNodes \subseteq_d SC'\ Nodes$

(Pre 3) $CRoot \in SC'Roots$

Del(SC' , C) $=_d$ SC where SC: S-Con and

(1)a $SCGr =_d SC'Gr - \{ CNodes \}$

b $SSNodes =_d SC'Nodes - CNodes$

Update

This operation allows for the replacement of a composite object, *C*: *SCComposite* in a semantic concept, *SC*: *S-Con* with a new composite object *C'*: *SCComposite*, immediately above a node *n: SCNodes*, resulting in a new semantic concept, *SC'*: *S-Con.* Thus Update(*SC*, *n*, *C, C'*) $=_d$ *SC'*.

The Update operation can be achieved by first removing *C* from *SC* and then inserting *C'* at the same node *n*, where *C* was attached, in *SC*. This can be expressed as a composition of Ins and Del. The precondition for Update(*SC*, *n*, *C, C'*) is the same as the precondition for Del. Firstly, *C: SCComposite* must exist in *SC* i.e. $C \in SC$ and $CNodes \subseteq_d SCNodes$ such that all the nodes in *C* have a common parent in *SC.* Secondly, *CRoot* must be a root in *SC*, thus $CRoot \in SCRoots$. These preconditions are necessary to avoid distortion or "knowledge gap". Clearly, any *C*: *SCComposite* can be deleted and a new *C'*: *SCComposite* inserted at the node *n*, where *C* existed; but allowing this to happen will result in having *SCAncs*(*n*) "hang loosely" as the update on *SCDesc*(*n*) will naturally affect *SCAncs*(*n*). In order to maintain integrity, one possible solution is to have all ancestor nodes, *SCAncs*(*n*) automatically update themselves. Another solution is to enforce a complete update from $CRoot \in SCRoots$. The later approach is adopted since it is more intuitive and more straightforward to implement than the former. The formal definition of Update is given.

Update

Redefines a composite object *C*: *SCComposite* in *SC*: *S-Con*. Deletes *C* from *SC* and replaces it with *C'* immediately above a node *n*, where *C* existed in *SC*, resulting in *SC'*.

Given SC:$S\text{-}Con$, n: $SCNodes$, C: $SCComposite$, and C': $SCComposite$ with

(Pre 1) $C \in SC$

(Pre 2) $CNodes \subseteq_d SCNodes$

(Pre 3) $CRoot \in SCRoots$

Update(SC, n, C, C') $=_d$ SC' where SC': $S\text{-}Con$ and

(1) $SC' =_d$ Ins(Del(SC, C), n, C')

GetRelations

This function accepts a composite object, C: $SCComposite$ in SC: $S\text{-}Con$ and returns a set of relations which specifies the relationship that exists among the contextual information of the corresponding child nodes, $SCChis(C)$ that constitute C. This function is very useful during the automatic deduction of high-level semantics. One precondition for this function is that C must exist in SC i.e. $C \in SC$ and $CNodes \subseteq_d SCNodes$ such that all the nodes in C have a common parent in SC.

Let n be a positive number representing the cardinality of the set $SCChis(C)$, i.e. $n =_d$ | $SCChis(C)$) | and I $=_d$ {1, 2, 3, …, n}. Given n set of child nodes of C; $A_1, A_2, \ldots, A_{n-1}, A_n$, let $B_i =_d$ SC$ConInfo(A_i \mid i: I)$. Thus, $B_1, B_2, \ldots, B_{n-1}, B_n$ represents the set of the contextual information in $A_1, A_2, \ldots, A_{n-1}, A_n$ respectively. The set of relations that defines the relationship among the contextual information in the child nodes is given as the n-ary cartesian product,

$\prod(B_i \mid i: I) =_d B_1 \text{ X } B_2 \text{ X } \ldots \text{ X } B_{n-1} \text{ X } B_n$

which is the set of all n-tuples $\{(b_1, b_2, \ldots, b_{n-1}, b_n) \mid b_1 \in B_1, b_2 \in B_2,$ and . . . and $b_{n-1} \in B_{n-1}, b_n \in B_n\}$. The formal definition of GetRelations is given.

GetRelations

Returns a set of relations, which specifies the relationship that exists among the contextual information of the corresponding child nodes of *C*: *SCComposite, SCChis*(*C*).

Given *SC*:*S-Con*, *C*: *SCComposite* with

(Pre 1) $C \in SC$

(Pre 2) $CNodes \subseteq_d SCNodes$

GetRelations(*SC*, *C*) $=_d \prod(B_i \mid i{:}\ I)$ where

(1) $n =_d |SCChis(C)|$ and $I =_d \{1, 2, 3, \ldots, n\}$

(2) $A_1, A_2, \ldots, A_{n-1}, A_n$: Set of child nodes of *C*

(3) $B_i =_d SCConInfo(A_i \mid i{:}\ I)$: Set of contextual information in A_i

(4) $\prod(B_i \mid i{:}\ I) =_d B_1 \times B_2 \times \ldots \times B_{n-1} \times B_n,\ \forall i{:}\ I,\ b_i{:}\ B_i$

This marks the end of the definition of *SC: S:Con* and all the operations that can be performed on it. Having fully specified *SC: S:Con,* the focus is now on the Semantic Space, *SS: S-Space* model that aids in video semantic organisation and understanding. *SS: S-Space* consists primarily of *SC: S-Con* arranged within clusters. The formal specification of *SS: S-Space* is given.

4.2.2.4 Formal Definition of S-Space

The semantic space, *SS: S-Space* is an object with the following features which consists of (a) a non-empty set, *SSClu: Set* of clusters of related semantic concepts; (b) a non-empty set of semantic concepts, *SSSC: S-Con* with their corresponding *SSInd: Index* which acts as pointers to *SSSC: S-Con.*

SSInd: Index is the set of all possible entries. One condition is imposed on *SSInd: Index* – this set may not be empty. This restriction is incorporated to rule out the trivial class of only one member, the 'empty' S-Space.

<u>*SSInd: Index*</u> Total set of index, potential or actual

$SSInd: Index \neq_d \varnothing$

SSSC: S-Con is the set of all the possible semantic concepts, *SSSC: S-Con* which may correspond to members of *SSInd: Index*. This set is also restricted to be non-empty for the same reason as *SSInd: Index* is non-empty.

<u>*SSSC: S-Con*</u> Total set of semantic concepts, potential or actual

$SSSC: S\text{-}Con \neq_d \varnothing$

SSClu is the set of all possible clusters of related semantic concepts. The clusters facilitate the classification of the semantic concepts and consequently eases retrieval and improve efficiency.

This is set is equally restricted to be non-empty.

<u>*SSClu*</u> Total set of semantically related clusters

$SSClu \neq_d \varnothing$

Lastly, SS: S-Space is defined as the class of base set of clusters, *SSClu*. *SSClu* has all possibly partial and finite functions *SSCluFn* from *SSInd: Index* to *SSSC: S-Con*. The domain of *SSCluFn* is finite; hence *SSCluFn* is finite. The domain of *SSCluFn, SSCluFnDef* $=_d$ *SSInd: Index* while its range, *SSCluFnRan* $=_d$ *SSSC: S-Con.* The complete specification of S-Space is given and a graphical illustration is presented in Figure 4.3.

<u>SS: S-Space</u>

* SSClu: Set Set of all clusters

* SSCluFn: *SSInd: Index* $\rightarrow$ *SSSC: S-Con*

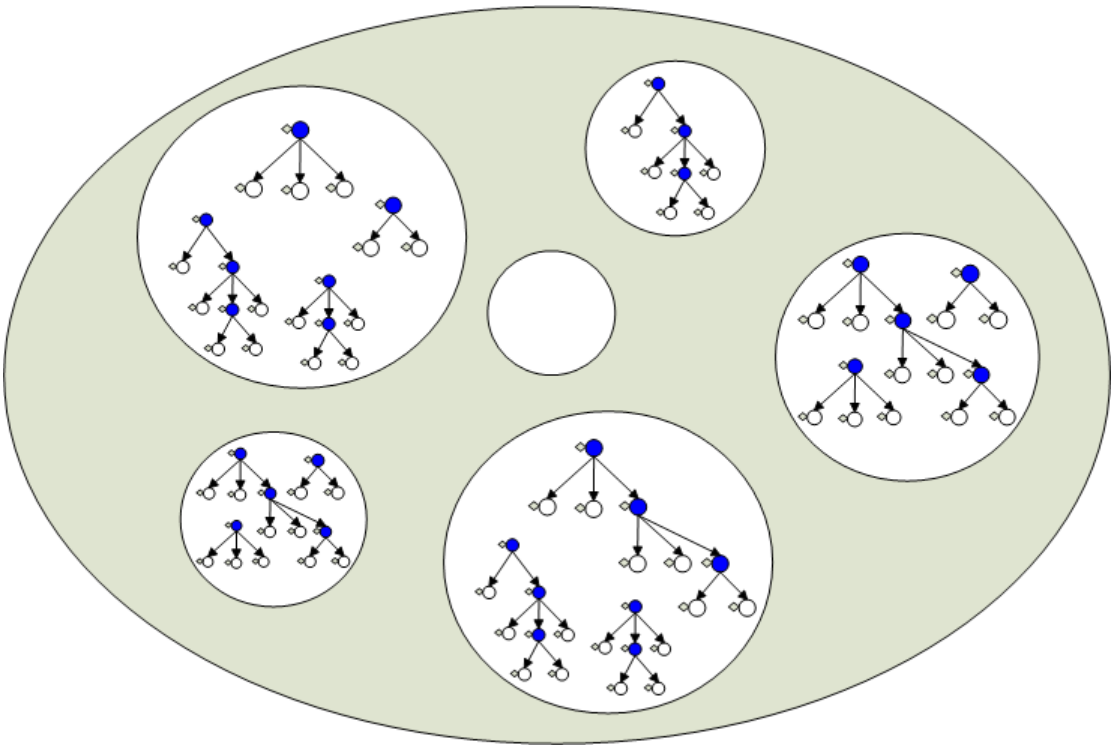

Figure 4.3: Graphical Illustration of the S-Space Model

The S-Space as shown in Figure 4.3 consists primarily of clusters and the couple (*SSInd, SSSC*). The clusters represent groups of semantically related semantic concepts, S-Con.

4.2.2.5 Operations on S-Space

Operations on S-Space are defined.

Emp

Emp returns an empty *S-Space*, that is, the *S-Space* whose *SSClu* is empty and whose *SSCluFnDef* is empty.

<u>Emp</u>

Return the empty *S-Space*

Emp $=_d$ *SS′* where *SS′:* S-Space and

(1) *SSClu* $=_d \varnothing$

(2) *SSCluFnDef* $=_d \varnothing$

Ins

Given a semantic space *SS*: S-Space, a cluster *c*: *SSClu*, an index *SSInd: Index*, and a semantic concept *SSSC: S-Con*, this operations returns a new *SS′*: S-Space, which is the result of inserting (*SSInd, SSSC*) in cluster *c* of SS. Thus, the graph *SS′CluFnGr* is defined as the union of *SSCluFnGr* and {(*SSInd, SSSC*)}.

However, a precondition is imposed on this operation: *SSInd* and *SSSC* must not already exist in SS, i.e. *SSInd* $\notin$ *SSCluFnDef* and *SSSC* $\notin$ *SSCluFnRan*. This restriction is adopted in order to ensure that all entries in *SS* are unique. It also reflects the assumption that seeking to 'insert' a semantic concept *SSSC*, which already exists in the

Semantic Space *SS*, will probably result in an error. Another restriction is that *c* must exist in *SS*, i.e. $c \in SSClu$. This restriction guarantees that the couple (*SSInd, SSSC*) is inserted in a relevant cluster in *SS*. The specification of Ins is given as follows:

Ins

Insert new index-semantic concept couple (*SSInd, SSSC*) within cluster *c*, of Semantic Space, SS.

Given *SS*: S-Space, *c*: *SSClu*, *SSInd: Index*, and *SSSC: S-Con* where

(Pre 1) $SSInd \notin SSCluFnDef$

(Pre 2) $SSSC \notin SSCluFnRan$

(Pre 3) $c \in SSClu$

Ins(*SS*, *c*, *SSInd, SSSC*) $=_d$ SS′ where *SS′*: S-Space and

(1) $SS'CluFnGr =_d SSCluFnGr + \{(SSInd, SSSC)\}$

Del

The operation Del reverses the effect of Ins. It takes three arguments: an *SS′*: S-Space, a cluster *c*: *SSClu*, and a semantic concept *SSSC: S-Con* contained within the cluster *c* in *SS′*. This means that the graph of *SS′CluFn* contains a couple (*SSInd, SSSC*) where *SSInd*: *Index* points to the semantic concept, *SSSC* in *SS′*. Given *SS′*, *c*, and *SSSC*, Del returns *SS*: S-Space, which is the result of deleting (*SSInd, SSSC*) from within cluster *c*, in *SS′*. The graph of *SSCluFnGr* is defined as $SS'CluFnGr - \{(SSInd, SSSC)\}$.

Two preconditions are imposed on Del is that: 1) *c* must exist in *SS′*, i.e. $c \in SS'Clu$; 2) *SSSC* must be defined in *SS′CluFn*, i.e. $SSSC \in SS'CluFnRan$. This implies that $SS'CluFn^{-1}(SSSC)$ will return *SSInd*.

Del

Remove an index-semantic concept couple ($SS'CluFn^{-1}(SSSC), SSSC$) within cluster c, of Semantic Space, SS'.

Given SS': S-Space, c: $SSClu$, and $SSSC$: $S\text{-}Con$ where.

(Pre 1) $c \in SS'Clu$

(Pre 2) $SSSC \in SS'CluFnRan$

Del(SS', c, $SSSC$) $=_d$ SS where SS: S-Space and

(1) $SSGr =_d SS'Gr - \{(SS'CluFn^{-1}(SSSC), SSSC)\}$

Update

This operation allows for the replacement of a semantic concept, i.e. a couple ($SSInd$, $SSCluFn(SSInd)$) within a cluster c in SS: S-Space with a new one, giving SS': S-Space. This can be achieved by first removing the couple ($SSInd$, $SSCluFn(SSInd)$) from cluster c in SS: S-Space and then inserting a new couple ($SSInd$, $SSSC$) back to cluster c'. The choice of not inserting back to the same cluster c but c' is to allow for flexibility as there could be instances where an Update may simply mean a move or update of the $SSSC$ to another cluster. However, c' does not necessarily have to be a different cluster – c and c' may refer to the same cluster, i.e. $c = c'$; which implies that Update is performed within the same cluster.

The preconditions imposed on the Update operation are: 1) c and c' must exist in SS, i.e. $\{c, c'\} \in SSClu$; 2) $SSInd$ must be defined in $SSCluFn$, i.e. $SSInd \in SSCluFnDef$. This implies that $SSCluFn(SSInd)$ will return the corresponding semantic concept, $SSSC$: $S\text{-}Con$ to which $SSInd$ point to.

<u>Update</u>

Redefines a semantic concept pointed to by *SSCluFn*(*SSInd*) in *SS*: S-Space.

Given *SS*: S-Space, *c, c′* : *SSClu*, *SSInd: Index*, and *SSSC: S-Con* where

(Pre 1) $\{c, c'\} \in SSClu$

(Pre 2) $SSInd \in SSCluFnDef$

Update(*SS*, *c*, *c′, SSInd, SSSC*) $=_d$ *SS′* where *SS′*: S-Space and

(1) $SS' =_d$ Ins(Del(*SS*, *c*, *SSCluFn*(*SSInd*)), *c′*, *SSInd, SSSC*)

LookUp

This function looks up a semantic concept, in the form of a composite node, *SSSCComposite* in *SS*: S-Space. It represents the main use of the semantic space, S-Space, which is to look up a semantic concept, *SSSC: S-Con* when presented with a set of nodes, *N*: *SSSCNodes* and a set of relations, *R* which specifies the relationship that exists among the contextual information of the nodes, *N*. Given a semantic space, *SS: S-Space*, a set of nodes, *N*: *SSSCNodes,* and a set of relations R, this function returns a composite node, *SSSCComposite* in *SS* whose child nodes, *SSSCChis*() matches the set of nodes, *N* and the relationship that exist among the contextual information of these child nodes 'closely match' the set of relations, R.

Let *C* represent the set of all composite nodes, *SSSCComposite* in SS. Let *n* be a positive number representing the cardinality of the set *C*, i.e. $n =_d |C|$ and $I =_d \{1, 2, 3, \ldots, n\}$; therefore $C =_d \{c_i \mid i: I\}$.

For any composite node, c_i and the set *N*, let c_i SemanticMatch *N* be the predicate true iff c_i semantically matches *N*, that is, the set of all child nodes in c_i equals *N*.

c_i SemanticMatch N iff

$$SSSCChis(c_i) = N, \forall x: x \in SSSCChis(c_i) \wedge x \in N$$

The definition domain of SemanticMatch, i.e. Def(SemanticMatch) contains all the composite nodes in *SS* whose child nodes equals *N*. Let D denote the set Def(SemanticMatch), thus

$$D =_d \text{Def(SemanticMatch)} \Rightarrow \{d_i \mid 1 \leq i \leq n(\text{Def(SemanticMatch)})\}$$

A function MatchRelation(d_i) is defined which returns the percentage of similarity that exist between the set of relations of contextual information of the composite node, d_i and the set of relations *R* for all the elements of the set *D*.

$$\text{MatchRelation}(d_i) =_d SSSCGetRelations(d_i)\ Matches\ R, \forall i \mid 1 \leq i \leq n(D)$$

Hence the LookUp function, which looks up a composite node in *SS,* returns d_i*: SSSCComposite* if MatchRelation(d_i) has the highest percentage, hence closest to *R* and MatchRelation(d_i) is greater than a threshold value, T; returns null otherwise. LookUp is defined as:

$$\text{LookUp}(SSSC, N, R) =_d \quad d_i \text{ if } (d_i, \text{Max(Ran(MatchRelation))} \in \text{MatchRelationGr} \wedge \text{Max(Ran(MatchRelation))} \geq T$$

$$\text{null otherwise}$$

Note that while the LookUp function looks up a semantic concept, *SSSC: S-Con* in *SS*, it has been defined to return *SSSCComposite* since an *S-Con* could have one or more *SSSCComposite* which represent various higher-level semantic concepts. This approach allows this function to return semantic concepts at any level. The formal specification of LookUp is given.

<u>LookUp</u>

Looks up a composite node in *SS:* S-Space and returns d_i*: SSSCComposite* if MatchRelation(d_i) has the highest percentage, hence closest to the set of relations *R* and MatchRelation(d_i) is greater than a threshold value, T; returns null otherwise.

Given *SS*: S-Space, *N*: *SSSCNodes*, and *R* ;

$$\text{LookUp}(SSSC, N, R) =_d \quad d_i \text{ if } (d_i, \text{Max}(\text{Ran}(\text{MatchRelation})) \in \text{MatchRelationGr} \wedge \text{Max}(\text{Ran}(\text{MatchRelation})) \geq T$$

$$\text{null otherwise}$$

where d_i: *SSSCComposite* and

(1) *C*: *SSSCComposite* Set of all composite nodes in *SS*

(2) $n =_d |C|$ and $I =_d \{1, 2, 3, \ldots, n\} \Rightarrow C =_d \{c_i \mid i: I\}$

(3) c_i SemanticMatch *N* iff

$$SSSCChis(c_i) = N, \forall x: x \in SSSCChis(c_i) \wedge x \in N$$

(4) $D =_d$ Def(SemanticMatch) $\Rightarrow \{d_i \mid 1 \leq i \leq n(\text{Def(SemanticMatch)})\}$

(5) MatchRelation(d_i) $=_d$ *SSSCGetRelations*(d_i) *Matches* R, $\forall i \mid 1 \leq i \leq n(D)$

4.3 Summary

A semantic video recognition model, which is a subset of the generic context mediation model presented in the last chapter was identified as essential towards achieving an automated video extraction. The Semantic Space (S-Space) model, which aids in automated multimedia semantic annotation was developed. The S-Space model is based on the decomposition of multimedia objects to semantically manageable units and the

inclusion of possible contextual information about the objects to facilitate multimedia semantic organization and management. The model is able to clearly decompose, manage, discover, and represent video semantics; it allows for multimedia semantics modelling at multiple levels of granularity. An OWL implementation using Protégé[40] ontology editor and knowledge-base framework from Stanford University is presented.

```
<?xml version="1.0"?>
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns="http://cafe.cic.hull.ac.uk/~icr03ee/research/ontology#"
   xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
   xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
   xmlns:owl="http://www.w3.org/2002/07/owl#"
   xmlns:p1="http://www.owl-ontologies.com/assert.owl#"
  xml:base="http://cafe.cic.hull.ac.uk/~icr03ee/research/ontology" >
  <rdf:Description rdf:about="#hasContextualInformation">
   <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">A
string containing a piece of contextual information</rdfs:comment>
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
   <rdfs:domain rdf:resource="#SCContextInfo"/>
   <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
  </rdf:Description>
  <rdf:Description rdf:about="#hasRelations">
   <rdfs:domain rdf:resource="#SCComposite"/>
   <rdfs:range rdf:resource="#SCComposite"/>
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
  </rdf:Description>
  <rdf:Description rdf:about="#hasAncestor">
   <rdfs:domain rdf:resource="#SCNodes"/>
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#TransitiveProperty"/>
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
   <rdfs:range rdf:resource="#SCNodes"/>
  </rdf:Description>
  <rdf:Description rdf:about="#SCAtomic">
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
   <owl:disjointWith rdf:resource="#SCComposite"/>
   <rdfs:subClassOf rdf:resource="#SCNodes"/>
   <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string"></rdfs:comment>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A0">
   <owl:someValuesFrom rdf:resource="#SCComposite"/>
   <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Restriction"/>
   <owl:onProperty rdf:resource="#hasRelations"/>
  </rdf:Description>
```

40 http://protege.standford.edu

```
  <rdf:Description rdf:about="#SCContextInfo">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
    <rdfs:subClassOf rdf:resource="#S-Con"/>
    <rdfs:subClassOf rdf:nodeID="A1"/>
    <rdfs:subClassOf rdf:nodeID="A2"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">This
class encapsulates the contextual information about an object.</rdfs:comment>
  </rdf:Description>
  <rdf:Description rdf:about="#hasObjectSemantics">
    <rdfs:domain rdf:resource="#SCNodes"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
  </rdf:Description>
  <rdf:Description rdf:about="">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Ontology"/>
  </rdf:Description>
  <rdf:Description rdf:about="#S-Con">
    <rdfs:comment
rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Represents all semantic
concepts.  S-Con comprises primarily of Atomics and Composites in a well-defined
fashion.</rdfs:comment>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A2">
    <owl:onProperty rdf:resource="#hasObjectURI"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Restriction"/>
    <owl:cardinality
rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</owl:cardinality>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A1">
    <owl:onProperty rdf:resource="#hasContextualInformation"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Restriction"/>
    <owl:minCardinality
rdf:datatype="http://www.w3.org/2001/XMLSchema#int">1</owl:minCardinality>
  </rdf:Description>
  <rdf:Description rdf:about="#isContextInfoOf">
    <rdf:type
rdf:resource="http://www.w3.org/2002/07/owl#InverseFunctionalProperty"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <owl:inverseOf rdf:resource="#hasContextInfo"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">The
inverse of hasContext.</rdfs:comment>
    <rdfs:domain rdf:resource="#SCContextInfo"/>
    <rdfs:range rdf:resource="#SCNodes"/>
  </rdf:Description>
  <rdf:Description rdf:about="#SCNodes">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
    <rdfs:subClassOf rdf:nodeID="A3"/>
    <rdfs:subClassOf rdf:resource="#S-Con"/>
  </rdf:Description>
  <rdf:Description rdf:about="#hasObjectURI">
```

```
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
    <rdfs:domain rdf:resource="#SCNodes"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
  </rdf:Description>
  <rdf:Description rdf:about="#SCComposite">
    <rdfs:subClassOf rdf:nodeID="A0"/>
    <rdfs:subClassOf rdf:resource="#SCNodes"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Class"/>
    <owl:disjointWith rdf:resource="#SCAtomic"/>
  </rdf:Description>
  <rdf:Description rdf:nodeID="A3">
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#Restriction"/>
    <owl:onProperty rdf:resource="#hasContextInfo"/>
    <owl:someValuesFrom rdf:resource="#SCContextInfo"/>
  </rdf:Description>
  <rdf:Description rdf:about="#hasContextInfo">
    <rdfs:range rdf:resource="#SCContextInfo"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
    <rdfs:domain rdf:resource="#SCNodes"/>
    <owl:inverseOf rdf:resource="#isContextInfoOf"/>
  </rdf:Description>
  <rdf:Description rdf:about="#hasObjectName">
    <rdfs:range rdf:resource="http://www.w3.org/2001/XMLSchema#string"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#FunctionalProperty"/>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#DatatypeProperty"/>
    <rdfs:domain rdf:resource="#SCNodes"/>
    <rdfs:comment rdf:datatype="http://www.w3.org/2001/XMLSchema#string">Object
Name</rdfs:comment>
  </rdf:Description>
</rdf:RDF>

<!-- Created with Protege (with OWL Plugin 2.2, Build 344)
http://protege.stanford.edu -->
```


The development and formalisation of the Semantic Space (*S-Space*) model demonstrates that it is possible to develop a model for automated video semantic understanding; by implication validating and verifying the third research hypothesis, H_3. The next chapter presents the development of a prototype based on the conceptual framework and context model presented in chapter 3.

Chapter 5

Context-Based Multimedia Management Prototype Development

The second and third research hypotheses, H_2 and H_3 have been evaluated in chapters 3 (section 3.5) and 4 (section 4.2) respectively. Evaluation is a continuous process in a modelling, design, and development research. Various aspects of the model were subjected to strict tests using formal methods before a decision was made. According to Vaishnavi and Kuechler (2004), each decision is followed by a "thought experiment" in which that part of the design is mentally exercised by the designer. This chapter and chapter 6 focus on the formal evaluation of the first research hypothesis, H_1 - "*Semantic description of video from the web will improve with an increase in contextual knowledge than with less contextual knowledge.*"

A robust and efficient multimedia semantic understanding and management framework is key to an effective multimedia semantic retrieval system. The novel multimedia semantic integration framework and a formalised context mediation model for the organisation and representation of the multimedia semantics have been presented. Under the generic multimedia semantics integration framework and based on the context mediation model, a prototype system for multimedia semantics generation and management – CONMAN, which implements aspects of the model is developed. This chapter presents the details of the CONMAN application.

This chapter has five main sections. Section 5.1 gives a system overview of the CONMAN prototype; section 5.2 explains the software process model that was adopted in developing the prototype; section 5.3 presents some key design considerations that are peculiar to the prototype development; section 5.4 discusses the architectural

strategies and presents the system flow; while section 5.5 presents a summary of the chapter.

5.1 CONMAN System Overview

CONMAN aims at semi-automatic semantics generation and authoring for video media resources in an integrated environment. The main focus of the CONMAN prototype is to achieve improved semantic understanding through the implementation of the relevant aspects of the context model (refer to section 3.4.5) and the framework (refer to section 3.4.3). The software application is capable of accepting video data from various sources including local or remote files and web URL. The integrated environment allows for the visual playback and semantic segmentation of visual data. A human annotator can provide various descriptions about the media, based on the various identified segments and some static metadata like title, category, and author.

The CONMAN system implements the framework for context-based multimedia management (presented in Figure 3.3) and by extension the context model, which is an integral part of the framework. Whilst there are two major parts (content representation and retrieval subsystems) within the framework, the CONMAN system focus on implementing the content representation subsystem. The content representation subsystem entails the extraction and representation of knowledge from video resources using the context model. The extracted metadata is represented in XML/RDF format, organised and indexed in a database. The context model is at the centre of the CONMAN prototype providing support to the content representation subsystem. Refer to the graphical representation of the context model presented in Figure 3.4. Each of the panels in the CONMAN user interface (see Figure 5.1) individually map to components within the context model. The *Media URL* panel, *Category* panel, and *Metadata* panel

in the CONMAN user interface, all map to *Annotator's Resource Description (ARD)* within the context model. Similarly, the *Media Information* panel within CONMAN, which displays some of the properties that are automatically extracted from video files maps to *Extracted Properties (EP)* in the context model. Furthermore, the *Semantic Description* panel in CONMAN, which allow users to provide annotations for each of the segmented key frames in the *Semantic Segments* panel maps to *Annotator's Semantic Description (ASD)* in the model. The CONMAN system can be extended to implement the S-Space model (see Chapter 4) which implements a *Knowledge Source (KS)* from the context model towards achieving fully automated video semantic annotation.

Whilst the CONMAN system focused on the implementation of the content representation subsystem, it can be extended to implement the retrieval subsystem and thereby fully implementing the context-based multimedia management framework (see Figure 3.3). The retrieval subsystem, which relies on the context model to return the most appropriate result to the user query based on pre-indexed metadata, is very important to the overall user experience as it assures retrieval results that matches users information need when search queries are presented.

The CONMAN prototype system is developed in Java, and its user interface is shown in Figure 5.1. Video files could be opened either using File – Open menu navigation (open icon or <CTRL>+O are alternatives) or by keying in a local or web URL in the Media URL input field. CONMAN scans the video file, performing shot segmentation and selecting the key frames which are displayed on the panel named *Semantic Segments in Media*. Some properties of the video (such as number of key frames and duration) are extracted and displayed in the *Media Information* panel. Also, the video becomes available in the *Media Preview* panel for possible playback by human

annotators.

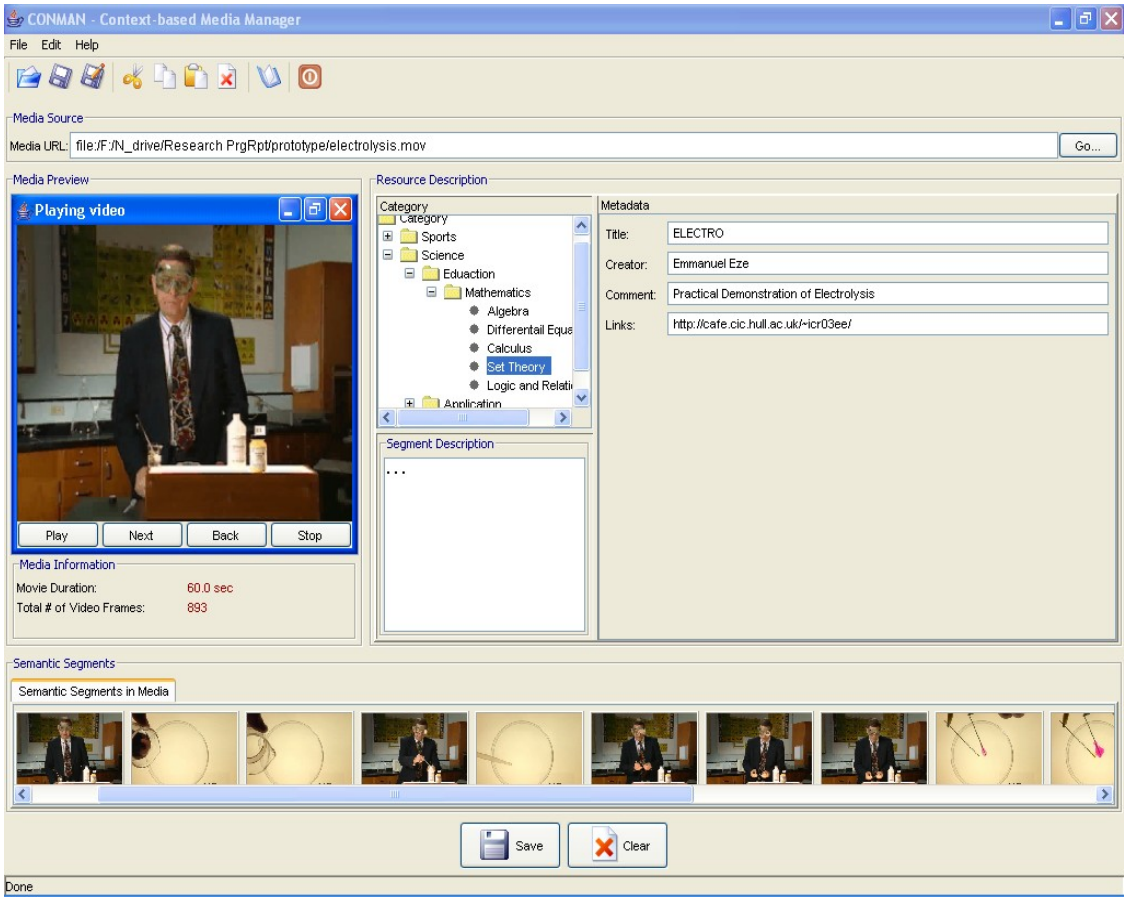

Figure 5.1: CONMAN Graphical User Interface

Users are allowed to provide descriptions based on the various key frames and metadata like creator, title etc. The user can provide various annotation within the *Segment Description* panel for each of the browsable key frames in the *Semantic Segments* panel. Users can also select the the category (genre) the video belongs to and other Dublin Core[41] metadata elements. With reference to other knowledge sources, integrated semantic descriptions of knowledge objects could be saved in the centralised knowledge base as metadata, using XML or RDF formats to allow for interoperability with other tools.

41 http://dublincore.org

5.2 Software Process Model

An Agile Development Paradigm (Black *et al.*, 2009) was adopted in the course of CONMAN development in preference to the more traditional models like the Waterfall (Schach, 1999). Due to the nature of the project, various identified components, and limited time and resources, an agile development approach was more suitable since most of the requirements and solutions evolve.

Agile method focuses and breaks tasks to smaller increments called "iterations" for minimal planning. This was very important as solutions were not readily available and could not have been pre-planned. Take for example, the issue of implementing the video segmentation component which is an open research issue and have researchers dedicated towards evolving efficient and effective video segmentation algorithms. Each iteration usually spans a short period of time and will involve the full software development cycle – requirement analysis, design, coding, and testing. This proved very useful as various components were developed, tested, and integrated over time ensuring easy adaptation to changing circumstances and requirements (Beck *et al.*, 2001).

5.3 CONMAN Design Considerations

CONMAN implements the context model, and by implication the original requirements set out in section 3.4.2. However, some key considerations and constraints which are peculiar to this prototype implementation are further discussed.

5.3.1 Assumptions and Dependencies

The following are the assumptions and dependencies regarding the CONMAN application:

- It should run on a typical mid-range personal computer without requiring extensive processing power and memory.
- It should be able to run on any operating system that supports a Java Virtual Machine (JVM).
- It should be intuitive to use and should require minimal user training.

5.3.2 General Constraints

The limitations that have significant impact on the CONMAN application development include:

- The S-Space model towards automated semantic recognition was not implemented. There are also no other knowledge system that CONMAN could interface to in order to learn and understand semantic concepts while achieving fully-automated annotation.
- Difficulty in getting very good freely available video segmentation routines in Java (see section 5.4) as most of the reviewed options are commercial products and mostly implemented in other Languages.
- The prototype implemented only what is required for the testing of the first research hypothesis, H_1. Since the major focus is on semantic content

description and representation based on the context model, the retrieval subsystem of the framework was not implemented.

5.3.3 Goals and Guidelines

The major goal was to maintain a simple design in creating a user-friendly, light-weight, intuitive software that effectively and efficiently implements the required aspects of the context model presented in chapter 3. A lot of emphasis was placed on evolving a CONMAN application that works and can effectively demonstrate aspects of the model and test the first research hypothesis, H_1.

5.4 CONMAN Architectural Strategies and System Flow

The CONMAN prototype implements the context model. Recall the graphical representation of the context model presented in Figure 3.4. The panels in the CONMAN user interface all map to each of the components of the context model. The *Media URL* panel, *Category* panel, and *Metadata* panel, all map to *Annotator's Resource Description (ARD)*. The *Media Information* panel which displays some of the properties that are automatically extracted from the video file maps to *Extracted Properties (EP)* in the context model. Similarly, the *Semantic Description* panel which allow users to provide annotations for each of the segmented key frames in the *Semantic Segments* panel maps to *Annotator's Semantic Description (ASD)*.

The CONMAN prototype system is developed in Java. The choice of Java is mainly because the researcher has done extensive programming with the Java Language and is very comfortable and expressive with the Java Language. The application relied on the Java Media Framework (JMF), a Java library that enables audio, video and other time-

based media to be added to Java applications and applets. JMF has a limitation of possibly not processing or requiring some plug-in to process some video types on certain computers. JMF facilitates media processing within the prototype thereby imposing its limitations on the prototype. The system work-flow for the CONMAN prototype is presented in Figure 5.2.

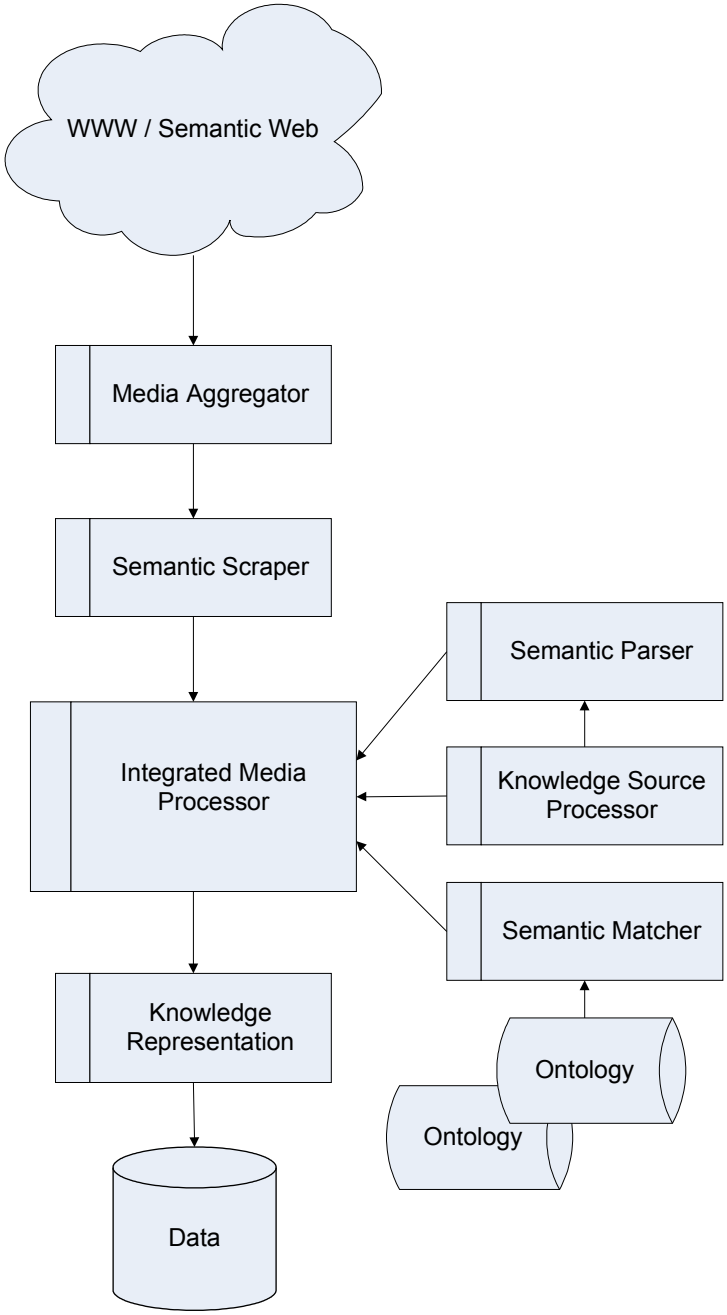

Figure 5.2: CONMAN workflow

The CONMAN system has four important components: media aggregator, semantic scraping, semantic parser, and semantic matcher.

The *Media Aggregator* is normally the starting point for the automatic annotation operation. A base URI is passed to it and it then scans through looking for known media types. If a media object is found, the *Semantic Scraper* component is invoked on the URI before continuing with the annotation. Alternatively, if the media is local or a known specific media on the web is desired, the full URI of the media is simply supplied instead. The *Semantic Scraper* applies some web scraping techniques in order to gather additional information about the media object that could add to better semantic understanding of the media content. The *Semantic Scrapper* is a possible Knowledge Source (KS) represented in the context model.

The *Integrated Media Processor* takes over control form the *Semantic Scrapper* and performs additional media processing involving the *Semantic Parser*, *Semantic Matcher*, and *Knowledge Source Processor* before generating and persisting the metadata. The *Semantic Parser* detects segments (e.g. shots/tracks in video/audio) in the media object. The detected segments are passed to the *Semantic Matcher* for matching against the media knowledge base. Also, the *Knowledge Source Processor* is invoked depending on media type. The *Knowledge Source Processor* identifies and processes other possible semantic knowledge source for the given media type. For example, in the case of video or audio, a possible knowledge source could be the transcription of audio to text. The *Knowledge Representation* component ensures conversion to the appropriate metadata format for storage.

5.5 Summary

The CONMAN prototype system was developed to demonstrate aspects of the framework and context model presented in chapter 3 towards evaluating the research hypothesis, H_1. The CONMAN prototype did not attempt to implement the entire framework but it rather focused on the key component of the framework which is applying the context model to achieving content representation. However, the CONMAN system can be extended to fully implement the context-based multimedia management framework by interfacing it to an implementation of the S-Space model (presented in Chapter 4) in order to achieve automated semantic understanding. Furthermore, the retrieval subsystem can be implemented to ensure that users can pose queries at the human perceptual level and expect good retrieval results based on their search needs. The prototype implements semantic understanding and representation at the knowledge level (refer to Figure 2.5) and aims at semi-automatic semantic generation and authoring for video media resources in an integrated environment. The key design decisions and components were presented and discussed. The next chapter presents the evaluation of the research hypothesis, H_1 based on the CONMAN prototype.

Chapter 6

Experimental Evaluation

This chapter presents the testing of the hypothesis, H_1 based on the CONMAN prototype presented in chapter 5. CONMAN implements aspects of the multimedia context descriptive model (refer to section 3.4.5) necessary to test the first research hypothesis, H_1. There are four main sections in this chapter. Section 6.1 presents the evaluation methodology and procedure; Section 6.2 provides the experimental results from the evaluation; the experimental result is analysed and presented in section 6.3; while section 6.4 presents a summary of the chapter.

6.1 Evaluation Methodology

Testing the first research hypothesis, H_1 requires a reproducible experiment. A naturalistic evaluation (Venable, 2006) approach is taken. Naturalistic evaluation explores the behaviour of a solution technology in its real environment. It is usually conducted using research methods such as field studies, surveys, ethnography, and action research. Venable (2006) supports the naturalistic evaluation approach when he called it "the real proof of the pudding" since it includes all the complexities of human practice in the real world. Fully implementing the frameworks and models proposed in this thesis will demand a lot of resources and time beyond what can be accommodated within the study period. While it is desirable to adopt a naturalistic evaluation approach for the entire framework, the resource and time limitations have necessitated a hybrid evaluation approach. CONMAN - the prototype implementation of the proposed framework for context-based multimedia management is used in a natural setting by humans to perform a controlled experiment. Two sets of video are annotated using the

CONMAN application and data collected for evaluation and analysis. The next section presents the evaluation metric which tests the research hypothesis, H_1.

6.1.1 Metric

Most research in multimedia semantics focus their evaluation metric on the traditional Information Retrieval metrics like precision, recall, fall-out, F-measure, etc. (Ren and Bracewell, 2009; ElAlami, 2011; Gennaro et al., 2011). Grubinger, Clough, Müller and Deselaers (2006) provided a retrieval benchmark for the cross-language image retrieval track (ImageCLEF) which is a sub-track of Cross Language Evaluation Forum (CLEF). However, since the main contribution of this work is not about multimedia retrieval but more about the quality of semantic annotation (refer to section 1.3); the *semantic similarity* of the collected data is measured against a "gold standard" to determine effectiveness. The "gold standard" refers to the description given to the video with all the contextual information by an independent party, who neither had any knowledge about the evaluation nor participated in it. The researcher acknowledges that arriving at a gold standard can be subjective due to possible human bias. The description was reviewed by two other individuals and unanimously accepted as the gold standard after careful consideration by the parties. The "gold standard" description served as a benchmark.

Semantic similarity can be defined as a confidence score that reflects the semantic relation between the meanings of two words and relates to computing the similarity between concepts which are not necessarily lexically similar (Simpson and Dao, 2010). WordNet::Similarity[42] (Pedersen, Patwardhan, and Michelizzi, 2004) is a freely available software package that makes it possible to measure the semantic similarity or relatedness between a pair of concepts (or word senses). It provides six measures of similarity, and three measures of relatedness, all of which are based on WordNet[43].

42 http://search.cpan.org/dist/WordNet-Similarity/
43 http://wordnet.princeton.edu/

WordNet (Miller 1995; Fellbaum 1998), a lexical database which organizes words into synsets, sets of synonymous words, and specifies a number of relationships such as hypernym, synonym, meronym which can exist between the synsets in the lexicon, has been shown to be particularly effective in the calculation of semantic similarity. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept (Varelas *et al.*, 2005). The similarity measures are implemented as Perl modules which take as input two concepts, and return a numeric value that represents the degree to which they are similar or related (Pedersen, Patwardhan, and Michelizzi, 2004).

The similarity measures are grouped into two, based on the path lengths between concepts and on information content, which is a corpus–based measure of the specificity of a concept. The lin (Lin, 1998) and wup (Wu and Palmer, 1994) measures, each belonging to one of the similarity measure group are employed in the evaluation to measure word pairs from the collected data. Pedersen, Patwardhan, and Michelizzi (2004) describe the wup measure as finding the path length to the root node from the least common subsumer (LCS) of the two concepts, which is the most specific concept they share as an ancestor. This value is scaled by the sum of the path lengths from the individual concepts to the root. The measured path is equal to the inverse of the shortest path length between two concepts. The lin measure augments the information content of the LCS of two concepts with the sum of the information content of the individual concepts. The lin measure scales the information content of the LCS by this sum. The lin and wup measures have been used extensively in the literature to measure semantic similarity in different application domains like paraphrase or sentence identification (Fernando and Stevenson, 2008; Simpson and Dao, 2010), image and video retrieval (Aytar, Shah, and Luo, 2008; Ferecatu, Boujemaa, and Crucianu, 2008), and ontology (Lord *et al.*, 2003).

Tintarev and Masthoff (2006) have demonstrated that the lin measure is a viable option for calculating similarity in the context of real-world news headlines. The real-world

news headline is in the general knowledge domain and closely related to the nature of data under evaluation. The evaluation task is described in the next section.

6.1.2 Tasks

Participants in the study were a random selection of 24 volunteer graduates of Computer Science or related discipline. All participants were male and female between the ages of 25 and 60 years old. They were mostly average web users with general knowledge and participated in the evaluation task by annotating two video clips using the CONMAN prototype. The choice of the participants was necessary as video data on the web are generated and described by average web users rather than certain video domain experts. The video clips were selected from the general genre and have the same basic content except that one included the audio track which is a possible context knowledge source while the other did not. The video clips were labelled as Video1 and Video2. Video1 had no audio track, hence less context; while Video2 had the audio track, hence more context. Each participant was asked to view and annotate each video with the CONMAN application, and producing a semantic description. Each participant had to produce semantic descriptions D1 and D2 for Video1 and Video2 respectively. The result is shown in Table 6.5 (presented in section 6.2).

6.1.3 Evaluation Procedure

The null hypothesis for H_1, $H_{1\text{-}0}$ is given as :

Semantic description of video from the web will not improve with an increase in contextual knowledge than with less contextual knowledge.

The alternate hypothesis for H_1, $H_{1\text{-}A}$ is given as:

Semantic description of video from the web will improve with an increase in contextual knowledge than with less contextual knowledge.

Given D1 as the set of sample descriptions taken with less context; D2 as the set of sample descriptions taken with more context; and D3 as the gold standard description:

$H_{1\text{-}0}$: SenSim(D2, D3) ≤ SenSim(D1, D3)

$H_{1\text{-}A}$: SenSim(D2, D3) > SenSim(D1, D3)

where SenSim(X, Y) is a function that returns the sentence semantic similarity of the sentence pair X and Y.

In order to test the null hypothesis $H_{1\text{-}0,}$ the collected data will have to be transformed to a measurable quantitative data.

The sentence semantic similarity measure, SenSim(X, Y) which determines how similar the meaning of the two sentences, X and Y are, will need to be computed on the sample data. The higher the score, the more similar the meaning of the two sentences (Simpson and Dao, 2010).

Simpson and Dao (2010), in addressing the issue of sentence semantic similarity proposed the following steps:

1. First, each sentence is partitioned into a list of tokens.

2. Part-of-speech disambiguation (or tagging).

3. Stemming words.

4. Find the most appropriate sense for every word in a sentence (Word Sense Disambiguation).

5. Finally, compute the similarity of the sentences based on the similarity of the pairs of words.

While ensuring that an automated process for calculating sentence semantic similarity evolves, the following steps were taken in computing the semantic similarity between two sentences:

1. Word tokenization and POS tagging

2. Stop word removal

3. Lemmatization

4. Sentence semantic similarity measure based on semantic similarity measure of word pairs.

The major difference between the proposed approach and the approach proposed by Simpson and Dao (2010) is that effort is made to preserve the semantics of the sentence by avoiding steps like word stemming or having to disambiguate the word sense after word stemming. The various steps are discussed in detail.

Word Tokenization and Part of speech tagging (POS)

Word tokenization splits sentences into words and punctuation marks. This is combined with Part of speech tagging (POS) which is the process of adorning or "tagging" words in a text with each word's corresponding part of speech. Part of speech tagging is based both on the meaning of the word and its positional relationship with adjacent words. For example the word “fly” could mean movement or insect depending on the context of use. Part-of-Speech tagging helps to disambiguate the senses of the words sufficiently to avoid gross errors in determining semantic similarity. A simple list of the

parts of speech for English includes adjective, adverb, conjunction, noun, preposition, pronoun, and verb. The output of the POS tagging also reflect more granular syntactic and morphological structure.

MorphAdorner[44] POS tagger was used in this research study for POS tagging. MorphAdorner is a Java command-line program from Northwestern University, which acts as a pipeline manager for processes performing morphological adornment of words in a text. It supports features such as: lemmatization, sentence detector, part-of-speech tagger, named entity detector, and phrase chunker. MorphAdorner provides different part-of- speech taggers including the MorphAdorner trigram tagger which uses a hidden Markov model and a beam-search variant of the Viterbi algorithm. Another included POS tagger is the MorphAdorner rule-based tagger, which is a modified version of Mark Hepple's rule-based tagger (Burns, 2006). A sample output of the POS tagger on the gold standard, S1 is presented in Table 6.1.

S1 (POS Tag)
The (dt)
boy (n1)
is (vbz)
concerned (vvn)
for (p-acp)
a (dt)
bug (n1)
caught (vvn)

44 http://morphadorner.northwestern.edu/morphadorner/

in (p-acp)
a (dt)
spider's (ng1)
web (n1)
and (cc)
thinks (vvz)
that (cst)
the (dt)
bug's (ng1)
mum (uh-j)
will (vmb)
rescue (vvi)
it (pn31)

Table 6.1: Sample POS tag and word tokens for S1

The full list of POS tags and word tokens for the entire sample data is presented in appendix A. Refer to Appendix B for a presentation of different MorphAdorner NUPOS word classes and parts of speech. MorphAdorner NUPOS Documentation[45] has a detailed explanation of the different possible output from the POS tagger.

45 http://morphadorner.northwestern.edu/morphadorner/documentation/nupos/

Stop Word Removal

Stop words are frequently occurring, insignificant words that appear in a database. They are words that are so common that they are often filtered out prior to, or after, processing of natural language data (text). There is no one definite list of stop words which all tools use. Stop words can be domain specific as words that could frequently occur in one domain may not necessarily frequently occur in another domain. The CLiPS (Computational Linguistics & Psycholinguistics) stop word list[46] which is based on Martin Porter's list[47] and expanded with words that seem to occur frequently in other lists was used in this study. CLiPS is a research centre associated with the Linguistics Department of the Faculty of Arts of the University of Antwerp, Belgium. Table 6.2 presents S1 stripped of stop words.

S1 – No Stop Words
boy
concerned
bug
caught
spider's
web
bug's
mum
rescue

Table 6.2: S1 word token without stop words

46 http://www.clips.ua.ac.be/pages/stop-words
47 http://snowball.tartarus.org/algorithms/english/stop.txt

Lemmatization

Lemmatization is the process of reducing an inflected spelling to its lexical root or lemma form. The lemma form is the base form or head word form that could be found in a dictionary. Other approaches (Simpson and Dao, 2010) for calculating sentence similarity seem to depend on word stemming. However, stemming offers a simpler alternative to lemmatization as it basically attempts to reduce a word to a base form by removing affixes, but the resulting stem is not necessarily a proper lemma. The MorphAdorner's implementation of the Porter and Lancaster stemmers were evaluated against lemmatization and the lemmatization results were better. For example, the stem for the word “curious” returned “curiou” and “cury” by the Porter and Lancaster stemmers respectively; while the lemmatizer returned “curious” which is more consistent with expectations.

Word stemming has useful application in information retrieval applications where semantics exactness is not so important. However, for the purpose of finding semantic similarities in sentences, lemmatization is more appropriate as the resulting lemma preserves semantics and thus can be measured, unlike the resulting stems from stemmers which cannot be measured as they will have no semantic meaning or even exist in the CORPUS. The lemmatization implementation in MorphAdorner was employed in lemmatizing the sample data and the result is given in appendix A, while Table 6.3 presents the result of the lemmatization S1.

S1 – Lemmatized
boy
concerned
bug
catch

spider
web
bug
mum
rescue

Table 6.3: S1 word token without stop words and lemmatized

Sentence Semantic Similarity Measure

Sentence semantic similarity is a measure of the semantic relatedness of two sentences taken into consideration the multidimensional nature of natural language expression. The measure should be an extension of semantic similarity measure of word pairs (as detailed in section 6.1.1). Existing methods for computing sentence similarity have been adapted to long text documents (Mihalcea, Corley, and Strapparava, 2006; Higgins and Burstein, 2007). These methods are more suitable in certain domains and not suitable for short sentences similar to the sample data in this study. When tested with sample data, scores were computed that are inconsistent with human expectations for obvious sample data.

Thus this research has to define and formalise a sentence semantic similarity measure which takes into account the semantic information and word vector. The measure SenSim(X, Y) which is primarily motivated by the Jaccard similarity coefficient (Jaccard, 1901) and the sentence similarity measure proposed by Simpson and Dao (2010), is presented in the next section. One of the key considerations in developing the measure is to facilitate automated computation without human interference.

6.1.4 Formalised Sentence Semantic Similarity Measure

The formalised sentence semantic similarity measure, SenSim(X, Y) is a function that takes two sentence pairs and returns a value between 0 and 1 based on the degree of their relatedness. The closer the value is to 1, the higher their semantic similarity; with 1 denoting the highest similarity and 0 no similarity. The function takes into account the semantic information and word vector. The semantic similarity of two sentences is calculated based on the semantic similarity measures of word pairs from the set of lemmatized word tokens of the sentence pair. This is represented by the function Sim(x, y), where x and y are word pairs. The function Sim(x, y) is generic that it can be of any implementation of any word pair similarity measure like *lch, wup, path, lin, jcn,* and *res* (Pedersen, Patwardhan, and Michelizzi, 2004) using information from a structured lexical database and from corpus statistics.

The SenSem(X, Y) was primarily motivated by the Jaccard similarity coefficient which is a similarity measure that compares the similarity between two sample sets and the measure proposed by Simpson and Dao (2010). The Jaccard similarity coefficient is defined as the size of the intersection divided by the size of the union of the sample sets, given as:

$$J(A,B) = \frac{A \cap B}{A \cup B}$$

When applied to sentence similarity, it is redefined as the size of the intersection of the words in the two sentences compared to the size of the union of the words in the two sentences. This would have been perfect if an exact string match of the individual word tokens is desired.

In order to bridge that gap, a function Sim(x, y) is introduced, which measures the semantic similarity between word pairs, given two sets of tokenized, lemmatized, and POS tagged sentences without stops words, X and Y. The co-domain of the function

Sim(x, y), comprises real numbers in the range of 0 to 1 which represents the level of similarity between the word pair. The value 0, represents no similarity, while the value 1 represent absolute similarity. However, since automation is desired and considering the difficulties in manually identifying all the exact senses of the word pairs and borrowing from the proposed measure by Mihalcea, Corley, and Strapparava (2006), the semantic similarity between all the senses of the word pairs are computed with the maximum similarity score returned. This is denoted by the function Max(Sim(x, y)). The sum of the maximum similarity scores returned by Max(Sim(x, y)) should have been normalised by the cardinality of the union of X and Y as expressed in the Jaccard similarity coefficient. However, the union of the two sets cannot be expressly determined since an exact lexical string matching is not considered. In order to eliminate duplicates as is synonymous with union of sets for the normalised cardinality, A new set, Z is introduced, whose sum of cardinality and that of the set Y achieves the objective. The set comprises elements with Max(Sim(x, y)) ≥ 0.5 and Max(Sim(x, y)) < 1. A threshold of 0.5 signifies the mid point between 0 and 1 where the similarity score is considered significant. Scores of 1, however represents perfect match and are excluded since they would have counted as elements of Y.

Given two sets of tokenized, lemmatized, and POS tagged sentences without stops words, X and Y; the overall sentence semantic similarity of two sentence pairs, X and Y is given as:

$$SenSem(X, Y) =_d \frac{\sum_{i=1}^{|X|} Max(\mathrm{Sim}(w_i, Y))}{|Z| + |Y|}$$

$$\text{where } w_i \in X \text{ and}$$
$$Z = \{ x \mid x \in Max(\mathrm{Sim}(w_i, Y)) \text{ and } 0.5 \leqslant x < 1 \},\ 1 \leqslant i \leqslant |X|$$

This sentence semantic similarity measure was developed due to the need to find automated tools to measure the sentence similarity between the sample data and the

gold standard. The sentence semantic similarity measure, SenSem(X, Y) in this evaluation relies heavily on the tools provided by MorphAdorner[48] for performing morphological adornment of words in a text and WordNet[49] for computational linguistics. The computation of SenSem(X, Y) is a four step approach which starts with word tokenization and POS tagging using the MorphAdorner POS tagger[50]. The second step involves the removal of stop words, which is achieved in this study using the CLiPS (Computational Linguistics & Psycholinguistics) stop word list[51]. In step three, the words are reduced to their lexical root in a process referred to as lemmatization. This can be achieved using the MorphAdorner Lemmatizer[52]. Lastly, a Perl routine adapted from the WordNet-Similarity[53] Perl module was written and used to find the most appropriate WordNet sense of each of the lemmatized words that is most related to the set of lemmatized words from the gold standard (Table 6.3). It is important to note that WordNet sense of each word is determined based on a similarity measure (e.g. *wup* and *lin*). The relatedness between all pairs of the senses are computed and the maximum value returned as defined by the function Max(Sim(w_i, Y)). Table 6.4 presents sample data for all steps towards computing SenSem(X, Y).

POS Tag	**Stop-words removed**	**Lemmatized**	**WordNet Sense (*Lin*)**	**Max(Sim(w_i, Y)) (*lin*)**
The (dt)				
boy (n1)	boy	boy	boy#n#3 and boy#n#3	1
should (vmd)				
be (vbi)				

48 http://morphadorner.northwestern.edu/morphadorner/
49 http://wordnet.princeton.edu/
50 http://morphadorner.northwestern.edu/morphadorner/postagger/
51 http://www.clips.ua.ac.be/pages/stop-words
52 http://morphadorner.northwestern.edu/morphadorner/lemmatizer/
53 http://search.cpan.org/~tpederse/WordNet-Similarity-2.05/

allowed (vvn)	allowed	allow		
to (pc-acp)				
discover (vvi)	discover	discover	discover#v#2 and catch#v#17	0.7126
and (cc)				
ask (vvi)				
questions (n2)	questions	question	question#n#3 and web#n#4	0.3192
himself (px31)				
instead (av)	instead	instead		
of (pp-f)				
being (vbg)				
asked (vvn)				
what (r-crq)				
he (pns31)				
thinks (vvz)				

Table 6.4: Sample data showing result of all four steps for calculating the Sentence Semantic Similarity Measure, SenSem(X, Y)

The computed SenSem(X, Y) for the above example in Table 6.4 is given as:

$$(1 + 0.7126 + 0.3192) / (1 + 8) = 0.2258$$

The derivation of the 0.5 threshold was largely motivated by the Jaccard similarity coefficient and some tests to evaluate the significance and consistency of the result with some random sample data. It is acknowledged that the 0.5 threshold was not rigorously evaluated and will form part of the future research to improve the sentence semantic similarity measure. The researcher acknowledges that all possible existing sentence similarity measures in the literature may not have been reviewed, since this is not the main focus of this thesis. However, future work will thoroughly investigate other approaches to computing sentence similarity measure towards further validating and improving the SenSem(X, Y) measure. The next section presents the experimental result.

6.2 Experimental Result

The sample data presented in Table 6.5 represents the result from the evaluation task. 24 persons participated in the task. Column 1 of the table represents the tag of the participant; column 2 represents the description from the annotation of the video with less context; while the last column represents the description from the annotation of the video with more context. The gold standard annotation of the video clip as described in section 6.1.1, given as D3 is presented thus: "*The boy is concerned for a bug caught in a spider's web, and thinks that the bug's mum will rescue it*".

Participant	D1 (Video1 with less context)	D2 (Video2 with more context)
1	The boy does not understand what you what him to do.	The boy should be allowed to discover and ask questions himself instead of being asked what he thinks.
2	Communication with concentration	Inquisitive mind of communication and gathering facts
3	A little boy carried out an action and was trying to explain his action as well as his observations.	A little boy observed a bug entangled a spider's web and wished that it could free itself from the web.
4	The kid is being asked to identify an object in the darkened room by someone behind the camera.	The kid's parents are responding to his inquiries as regards a fly trapped in a spider web inside the darkened room.
5	A child repeating an action he is seen and been almost told reasons of the decision he is about to make.	A child relating his assurance of protection by his mum due to past event to the bug in the spiders web
6	The boy is being asked to say what he is seeing by some group of people.	A spider is trapped in a cobweb and the boy is asking questions like why its not flying.
7	The child is trying to extinguish the burning fire.	The child is trying to rescue some insect trapped in spider web.

8	the child is trying to kill the insect.	the child is trying to rescue the insect that is trapped in the spider web.
9	The scenario of a child attempting to spray a spot on the wall, but didn't know how to.	The scenario of a child attempting to spray a bug trap by a spider web, but was stopped by the parents
10	Scenario of a child attempting to put out a fire	Scenario of a child who is concerned about an insect trapped in a spider web
11	He is determined to seal the black spot on the window plane with the spray.	He is concerned about the still trapped spider on the window plane and wants to set it free.
12	A little boy who was outdoors and holding his toys sees two candles burning inside a room with empty chairs from the window; he is communicating his observations to a third party but he appears helpless.	A little boy observes a spider trapped in a web at the window pane and he is asked by an adult ,probably his father how the spider can get out,his answer was that the spider's mum will help it out.

13	A Confused boy trying to get rid of an insect but doesn't know how to use the insecticide.	A confused boy trying to understand why the insect is not moving despite making an attempt using the objects in his hand.
14	A clip on how a child can be guided to achieve a task	How to explain in simple terms the situation at hand making the person in question understand and accepting what you are saying.
15	A boy is probably being shown the effect of an action he's been told to perform	A boy is wondering how a bug caught in a spider web is going to be set free.
16	Pesticide should not be left carelessly where children are.	Children heart are, innocent, free, and compassionate; to both human and animals. The boy felt, the baby spider's should have a mum looking after it.
17	A child s decision is being influenced by an adult towards making a fun loving act while having a nice time.	A child is curious on what happens to a spider cut-up in a web and engages his parents in the act
18	Little boy trying to exterminate a bug	Little boy feels empathy for bug

19	The little boy wants to find clarity about the spot on the wall he was gazing at.	In the observations of the child of what was trapped, he believes that the bug can be rescued; but worried that he could do nothing in that direction.
20	The boy sprayed something on the wall while wondering what it could be.	A fly was caught in the spider's web and the boy was wondering how it would be rescued.
21	The little boy stares at a particular spot on the wall without a clue of what to make out of his observing object.	On seeing the bug trapped in a web the little boy was worried that he could do nothing, but that the mummy could do something.
22	Something caught the boy's attention, but he could not make out what about it.	Observing that the web caught the bug, the little boy wondered why there cannot be any rescue for the bug to release itself by flying away.
23	There boy was looking at a dent on the window .	The boy was looking for a way to rescue the fly from the spider's web.
24	Perhaps the boy picked the material he held on hand and could not know what to do next whilst staring at an object of observation.	Apparently, the boy is experimenting in experiencing of what he saw held in a web on the wall.

Table 6.5: Collected data from the evaluation task.

The semantic similarity of both D1 and D2 are each measured against D3 by applying the four-step approach for computing the semantic similarity between two sentences, as outlined and discussed in section 6.1.3. The result of calculating the sentence semantic similarity scores, SenSim(D1, D3) and SenSim(D2, D3) (as presented in section 6.1.4) on the data using the *lin* and *wup* measures are presented in Table 6.6.

S/N	**SenSim(D1, D3) – lin**	**SenSim(D1, D3) – wup**	**SenSim(D2, D3) – lin**	**SenSim(D2, D3) – wup**
1	0.1868	0.2000	0.2258	0.2576
2	0.0284	0.1171	0.0444	0.1337
3	0.3244	0.3694	0.7120	0.6424
4	0.2077	0.2825	0.5081	0.5589
5	0.2100	0.2850	0.7156	0.6539
6	0.3011	0.3030	0.3768	0.4110
7	0.1594	0.2118	0.5306	0.5192
8	0.1940	0.2603	0.5306	0.5192
9	0.2041	0.3168	0.5822	0.5943
10	0.1378	0.2250	0.5129	0.5220
11	0.1508	0.3030	0.4445	0.4105

12	0.5055	0.4896	0.7331	0.7576
13	0.2240	0.3481	0.3980	0.4335
14	0.1499	0.2697	0.3249	0.3589
15	0.2304	0.2958	0.6822	0.6532
16	0.1482	0.2045	0.4253	0.4451
17	0.2384	0.3255	0.4040	0.4260
18	0.2839	0.3222	0.2809	0.3352
19	0.2025	0.3212	0.5081	0.6151
20	0.1596	0.2815	0.6274	0.6904
21	0.3022	0.3860	0.7663	0.7619
22	0.2621	0.3088	0.6788	0.6905
23	0.1678	0.2511	0.5163	0.5380
24	0.3547	0.4591	0.3920	0.4157

Table 6.6: Sentence semantic similarity score for the sample data

Column 1 represent the serial number tag for the samples; column 2 represent the sentence semantic similarity scores with less contextual information using the *lin* measure represented as **SenSim(D1, D3) – lin**; column 3 represent the sentence semantic similarity scores with less contextual information using the *wup* measure represented as **SenSim(D1, D3) – wup**; column 4 represent the sentence semantic similarity scores with more contextual information using the *lin* measure represented as

SenSim(D2, D3) – lin; while, column 5 represent the sentence semantic similarity scores with more contextual information using the *wup* measure represented as **SenSim(D2, D3) – wup**. The result in Table 6.6 is analysed in the next section and conclusion drawn.

6.3 Result Analysis and Discussion

The evaluation seek to examine the efficiency of the prototype (by having people use it to annotate video) and the effectiveness of the context-based approach by validating the null hypothesis, $H_{1\text{-}0}$, given as:

$$H_{1\text{-}0}: \text{SenSim(D2, D3)} \leq \text{SenSim(D1, D3)}$$

The Pearson product-moment correlation coefficient (*r*), which is a measure of the strength of a linear association between two variables was used to compute the relationship between SenSim(D1, D3) and SenSim(D2, D3). Given two variable X and Y, the Pearson product-moment correlation coefficient is given as:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\ \sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

where $\bar{X}$ is the mean of X and $\bar{Y}$ the mean of Y
and n = number of population sample

The result of computing the Pearson product-moment correlation coefficient, *r* on the sample data to assess the relationship between the sentence semantic similarity scores of samples sentence pairs (with less context and more context respectively) using both the *lin* and *wup* measures is presented in Table 6.7.

		SenSim(D1, D3) – lin	SenSim(D1, D3) – wup	SenSim(D2, D3) – lin	SenSim(D2, D3) – wup
SenSim(D1, D3) – lin	**Pearson Correlation**	1.00	.90	.48	.48
	Sig. (1-tailed)		.00	.01	.01
	N	24	24	24	24
SenSim(D1, D3) – wup	**Pearson Correlation**	.90	1.00	.50	.52
	Sig. (1-tailed)	.00		.01	.00
	N	24	24	24	24
SenSim(D2, D3) – lin	**Pearson Correlation**	.48	.50	1.00	.98
	Sig. (1-tailed)	.01	.01		.00
	N	24	24	24	24
SenSim(D2, D3) – wup	**Pearson Correlation**	.48	.52	.98	1.00
	Sig. (1-tailed)	.01	.00	.00	
	N	24	24	24	24

Table 6.7: Result of Pearson product-moment correlation coefficient, r on sample data

The result showed that there was a positive correlation between the two variables (SenSim(D1,D3) and SenSim(D2,D3)) for both *lin* and *wup* measures, ($r = 0.48$, $n = 24$, $p < 0.05$ and $r = 0.52$, $n = 24$, $p < 0.05$), one tail. One-tailed P value was used since the hypothesis is directional and tests for either a difference or no difference in the

direction of the test. Thus a result in the opposite direction is not expected since the sentence semantic similarity scores for the sample with more context is expected to either result in an increase over the sentence semantic similarity scores for the sample with less context, or have no effect.

Overall, there was a significant, positive correlation between SenSim(D1,D3)_lin; SenSim(D1,D3)_wup and SenSim(D2,D3)_lin; SenSim(D2,D3)_wup, indicating that the sentence semantic similarity scores for the sample with more context, (SenSim(D2,D3)_lin and SenSim(D2,D3)_wup) yielded significant improvements to the sentence semantic similarity scores of the corresponding sample with less context, (SenSim(D1,D3)_lin and SenSim(D1,D3)_wup) respectively. This means that SenSim(D2, D3) is **not** less than or equals SenSim(D1, D3), the null hypothesis $H_{1\text{-}0}$ is therefore rejected in favour of $H_{1\text{-}A}$. The mean values for either of the sentence semantic similarity scores with more context (SenSim(D2,D3)_lin; SenSim(D2,D3)_wup) as shown in Table 6.8 is significantly higher than the corresponding scores with less context (SenSim(D1,D3)_lin; SenSim(D1,D3)_wup). The column chart and line graph of the sentence semantics similarity scores in Figures 6.1 and 6.2 respectively presents a visual summary of the result.

Variable	N	Mean	Std Dev	Variance	Minimum	Maximum
SenSim(D1, D3) – lin	24	.22	.09	.01	.03	.51
SenSim(D1, D3) – wup	24	.30	.08	.01	.12	.49
SenSim(D2, D3) – lin	24	.50	.18	.03	.04	.77
SenSim(D2, D3) – wup	24	.51	.16	.02	.13	.76

Table 6.8: Mean, Standard Deviation, and Variance of the sample data

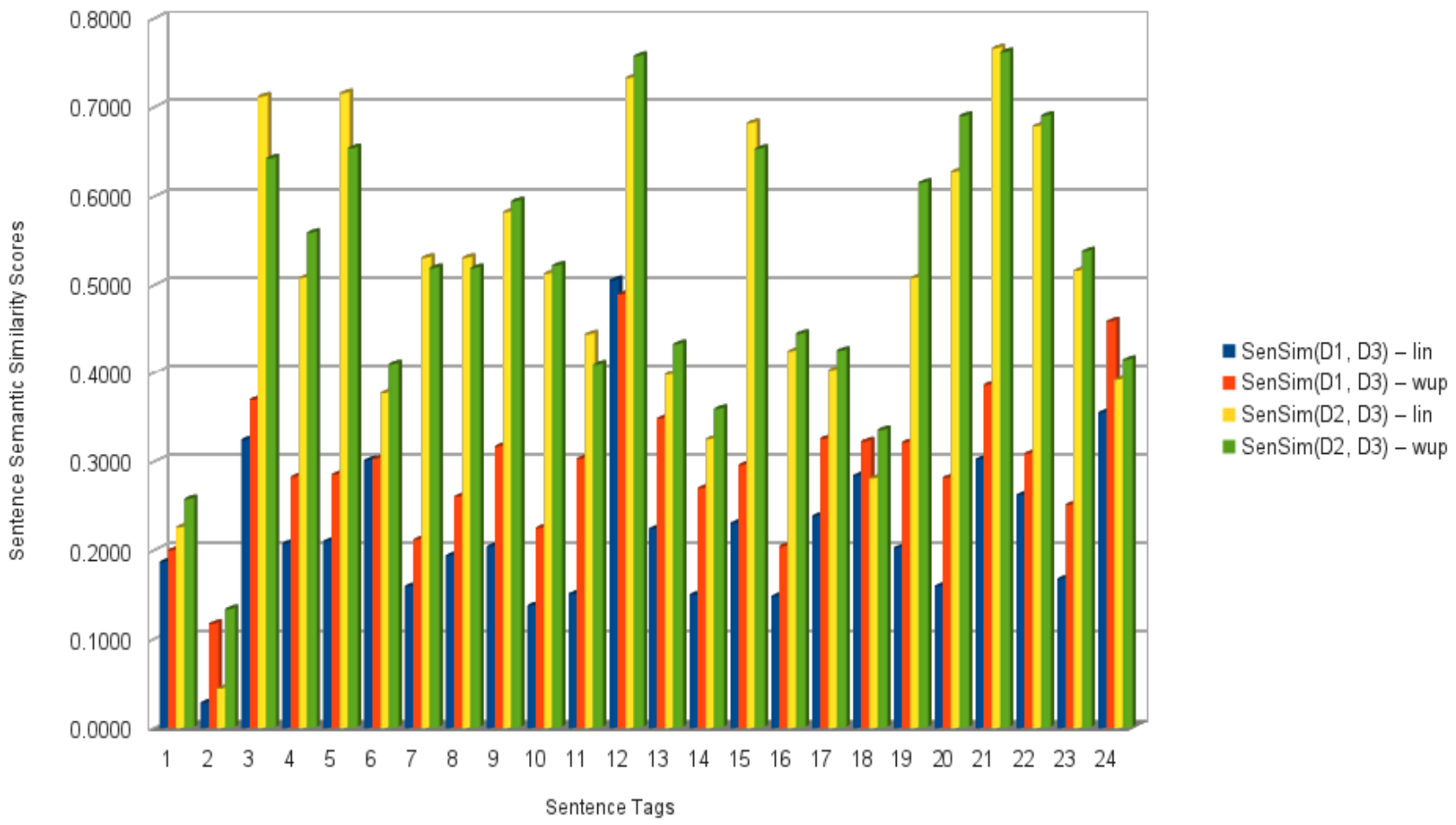

Figure 6.1: Column Chart of the Sentence Semantic Similarity Scores

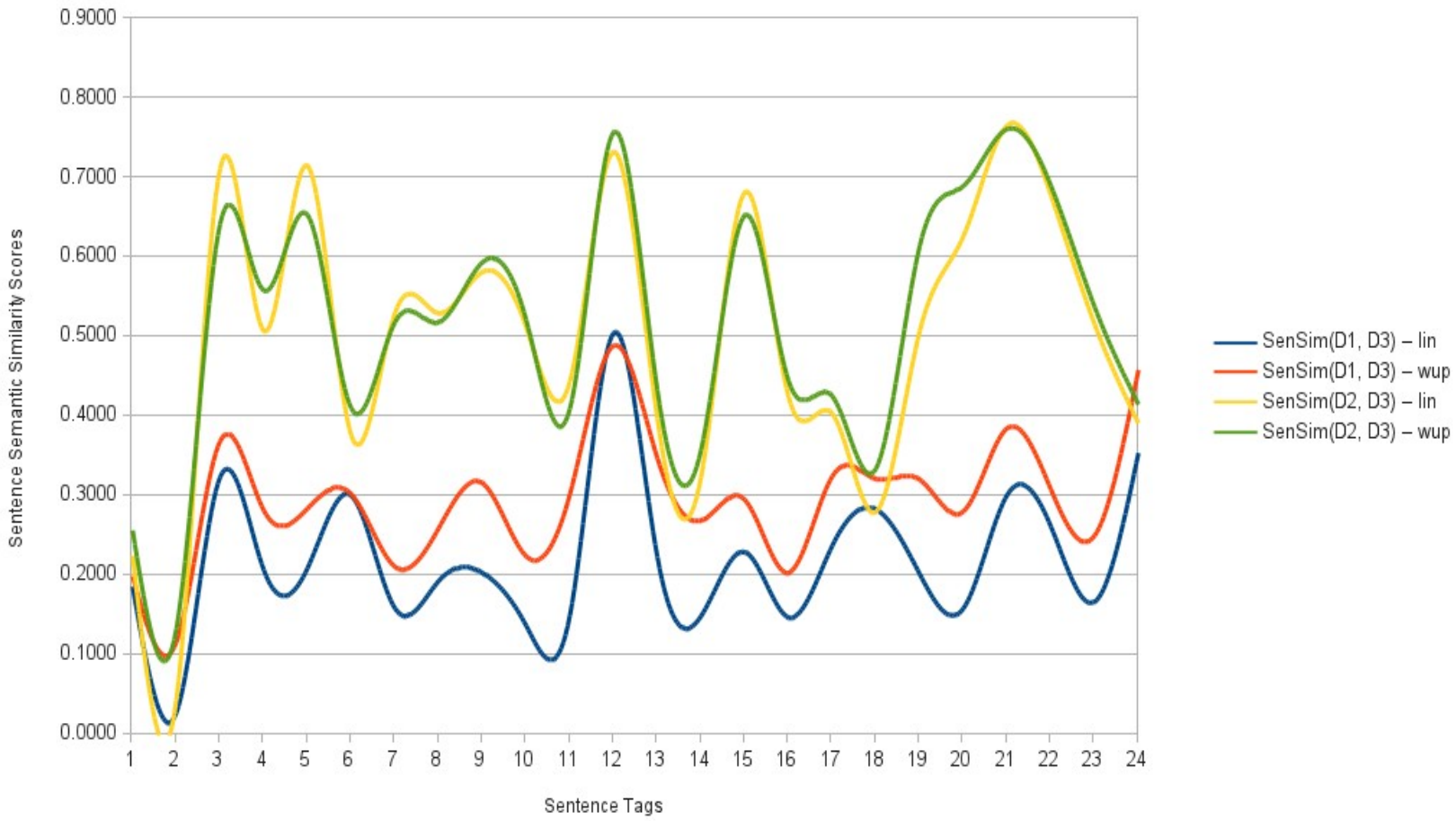

Figure 6.2: Line Chart of the Sentence Semantic Similarity Scores

6.4 Summary

From the experimental evaluation conducted, it was shown that semantic description of video from the web will improve with an increase in contextual knowledge than with less contextual knowledge. Users were asked to use the CONMAN application to describe two video clips with the same content except that one had more contextual information than the other. The basic idea was to evaluate the effectiveness of the CONMAN application and also gather the different descriptions in order to measure the semantic similarity between the two sentence description pairs and thus evaluate the effectiveness of the context model.

A four step approach was proposed for the computation of semantic similarity between sentence pairs. They are: word tokenization and POS tagging; stop word removal; lemmatization; and sentence semantic similarity measure based on the semantic similarity measure of word pairs. A new sentence semantic similarity measure was proposed due to unavailability of a suitable automated measure. The sentence semantic similarity is denoted by SenSim(X,Y) and was primarily motivated by Jaccard's similarity coefficient and the work by Simpson and Dao (2010). It was implemented in Perl for the purpose of the evaluation.

SenSim(X,Y) was computed for the sentence pairs and results presented. The analysis of the result showed that there was a positive correlation between the two variables (SenSim(D1,D3) and SenSim(D2,D3)) for both *lin* and *wup* measures, ($r = 0.48$, $n = 24$, $p < 0.05$ and $r = 0.52$, $n = 24$, $p < 0.05$), one tail. This led to the rejection of the null hypothesis, $H_{1\text{-}0}$ in favour of $H_{1\text{-}A}$. The efficiency of CONMAN was equally validated by virtue of it being used to accomplish the evaluation task.

The researcher acknowledges that the evaluation approach has limitations as the results are not based on large set of participants and video files with varied contextual dimensions. It was difficult to get participants within the time constraint of the research work to focus and annotate the video files since they were not domain experts but general web users. The evaluation task required some efforts and focus from the participants. It might be difficult to have them perform the task on a large set of video as their focus can lessen and result in collection of inconsistent data. Towards addressing the current limitations of the evaluation methodology, future research work will replicate the same evaluation scenario with a large set of video drawn from various subject matter areas in order to validate the initial findings. Future work to implement and integrate the S-Space model into the CONMAN prototype might eliminate the need to evaluate with human participants thereby making it easier to test with a large dataset. The next chapter presents the concluding part of the thesis.

Chapter 7

Discussion and Conclusion

The increase in multimedia content on the web due to the proliferation of computers, mobile devices, and social media sites have resulted in research effort to organise and perform search on the huge multimedia content (Enser, 2008a; Duygulu and Bastan, 2011; Razikin *et al.*, 2011). Computers are better at measuring video features (e.g. colour histogram, texture statistics, shape). However, users mostly express their information need at a high-level with their innate cognitive abilities which is lacking in computers (Hare *et al.*, 2006a; Hider, 2006; Enser, 2008a; Chung and Yoon, 2010; Fauzi and Belkhatir, 2013). This makes it difficult for researchers to model systems that can identify high-level meaning from video data (Smeulders *et al.*, 2000; Enser, 2008a; Dalakleidi *et al.*, 2011). The context approach to semantic visual information annotation presented in this thesis opens up a new dimension in visual information retrieval research. The framework and context model developed in this thesis presents opportunities for further research work.

This chapter presents a summary of the research work presented in this thesis. In reaching a conclusion, the researcher reflects on the research process and examines how well the aims and objectives of this thesis have been achieved. The key contributions of the thesis are presented and direction for future work identified.

7.1 Reflections on the Research

"*... the scientist builds in order to study; the engineer studies in order to build.*"

- Brooks (1996).

This thesis builds on the popular statement by Brooks (1996) and argues that the outcome of the thesis is credible since a consistent scientific research approach has been followed. There are different research methodologies available to researchers in Computer Science to address their research questions. This thesis adopts a mixed method approach, incorporating both Design Science Research (Ulrich, 2006; Gregor and Hevner, 2011; Kuechler and Vaishnavi, 2011) and formal methods (Scheurer, 1994; Sommerville, 2010; Kaur, Gulati, and Singh, 2012) for investigation and evaluation. The context and S-Space models developed in this thesis, were abstracted and rigorously specified using formal methods to evolve into a generalised model and scientific knowledge contribution. This is grounded in the design science research approach which Ulrich (2006), posits that design artefacts or models should emphasise a certain level of abstraction through a rigorous specification using formal language which can possibly prove the adequacy of the artefact and also qualifies it as a scientific knowledge contribution. The models and framework developed in this thesis were published in peer-reviewed conferences and journals (refer to Appendix C) to further demonstrate their validity, contribution, and relevance to the research community.

The aim of this thesis is to investigate how the notion of context can be applied towards bridging the multimedia semantic gap. The research outcome achieved the aim through the creation of a framework and models that facilitate the development of tools for context-based semantic multimedia annotation and retrieval capable of understanding

and managing multimedia semantics at the human perceptible knowledge level (see Figure 2.5). The research questions (RQ) were addressed as follows:

RQ1. **In what form is contextual information expressed in web video data?** Implicit and explicit contexts about web video data were identified (refer to section 3.4.1). The implicit context originates from the video internal features (like audio transcription) while the explicit contexts are derived from external resources to the video itself such as any information that can be gleaned from the website where the video is found. It was necessary to identify various contextual sources for web video in order to properly incorporate them within the context model.

RQ2. **How can context be modelled to represent video semantics from web video data?** Having identified the contextual information expressed in RQ1 and surveyed the state of the art, the context model presented in Chapter 3 was developed and formalised through a rigorous specification process using Feature Notation (refer to section 3.4.5). The context model supports multimedia semantic discovery and representation from heterogeneous sources. It relies on the use of contextual information about multimedia resources to organise and manage multimedia semantics. This also validated the research hypothesis, H_2.

RQ3. **How can context be applied in the automatic semantic description of visual features in web video data?** A key limitation of current state of the art is lack of automated video semantic understanding (Snoek *et al.*, 2006; Duygulu and Bastan, 2011; Hu *et al.*, 2011) at the human perceptual level (refer to sections 3.3 and 3.4.2.3). Automating semantic recognition does not only introduce efficiency but also removes the inefficiency and subjectivity (Kender

and Naphade, 2005; Hare *et al.*, 2006a) associated with manual annotation. The S-Space model presented in Chapter 4 implements a knowledge source (KS) from the context model which aids in automated video semantic annotation. This is based on the decomposition of visual objects to semantically manageable units and the inclusion of possible contextual information about the objects to facilitate visual information semantic understanding. This also validated the research hypothesis, H_3.

RQ4. **Can context information improve the representation of the semantic knowledge in web video data?** Having developed the context model, the CONMAN prototype was developed in Chapter 5 that implements aspects of the model necessary for the experimental evaluation presented in Chapter 6. The objective was to evaluate the effectiveness of the context model through the CONMAN application. Participants produced various annotations D1, (video with less context) and D2 (video with more context) and a gold standard, D3 was defined. The sentence semantic similarity measure, *SenSim(D1,D3)* and *SenSim(D2,D3)* were computed and the analysis of the result showed a positive correlation between the two variables (*SenSim(D1,D3)* and *SenSim(D2,D3)*) for both *lin* and *wup* measures, ($r = 0.48$, $n = 24$, $p < 0.05$ and $r = 0.52$, $n = 24$, $p < 0.05$), one tail. This led to the rejection of the null hypothesis, $H_{1\text{-}0}$ in favour of $H_{1\text{-}A}$.

7.2 Contribution made by the Thesis

Computers are excellent in performing logical and mathematical computations but are found wanting in performing cognitive reasoning which humans utilise to process visual information content (Enser, 2008a; Chung and Yoon, 2010). There are research efforts to design systems that understand the high-level meaning in multimedia documents by possibly translating computable low-level multimedia features (like colour histogram, shape, texture etc.) into high-level semantic concepts which humans can relate to (Hare *et al.*, 2006a; Zhang, 2007; Duygulu and Bastan, 2011; Hu *et al.*, 2011; Fauzi and Belkhatir, 2013). This is referred to in the literature (Hare *et al.*, 2006a; Zhang, 2007; Duygulu and Bastan, 2011; Hu *et al.*, 2011) as the "semantic gap" which Smeulders *et al.* (2000) defined as the "*lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation*".

This thesis addressed the semantic gap problem from a knowledge perspective by adding a context model to a visual information retrieval framework. Current state of the art apply context either at the retrieval phase by modelling user and query context (Ingwersen and Jarvelin, 2005; Natsev *et al.*, 2007; Borlund *et al.*, 2008; Goker, Myrhaug, and Bierig, 2009; Ruthven, 2011; Biswas, 2012), or as concept detector (Jiang *et al.*, 2009; Wiliem, Madasu, Boles, and Yarlagadda, 2012; Yi, Peng, and Xiao, 2012) by attempting to define a mapping between low-level features and high-level concepts. This thesis proposed a different view to context as any information about multimedia data that enhances its semantic understanding (see section 3.4.1) and argues that context can be applied to both the annotation and retrieval subsystems in an MIR framework. Unlike related work in Visual Information Annotation which derives from only implicit context to achieve concept mapping (Fan *et al.*, 2004b; Jiang *et al.*, 2009),

the developed context model derives from both implicit and explicit context to achieve semantic content annotation and representation. In addition, the context model defines the metadata representation of the visual data so as to represent the semantics at different levels based on different knowledge sources.

This thesis lies within the Information Science body of research in Computer Science and the researcher feels that the thesis has contributed to scientific knowledge in the following way:

- to the field of Multimedia Annotation, through the unique context approach and development of the framework which led to the formalised context model for organising and representing the semantics of heterogeneous multimedia data. The model relies on the identification and use of contextual information about multimedia resources to enhance the organisation and management of multimedia.

- to the field of Multimedia Semantics, through the S-Space model for automatic video semantic understanding which is key to managing large scale video as is obtainable on the web. The S-Space model implements a knowledge source (KS) from the context model which aids in automated video semantic annotation. This model relies on the decomposition of visual objects to semantically manageable units and the inclusion of possible contextual information about the objects to facilitate visual information semantic understanding.

7.3 Research Limitations

This thesis took a Design Science Research (DSR) and Formal Methods research approach. Researchers with a different research orientation may view the absence of evaluating the innovative artefacts through a prototype that fully implements the context model as a limitation. However, DSR does not always aim at providing concrete solutions to identified research problems but provides a rigorous methodology for producing novel research artefacts which can be building blocks towards solving both practical and theoretical Computer Science problems. DSR ensures that artefacts are abstracted and generalised (Dodig-Crnkovic, 2002) such that they constitute a new scientific knowledge contribution.

However, a possible limitation is the limited scope of formal methods in modelling user interfaces. This led to the formalisation of only the core components rather than the entire system. There is still an opportunity to research a complete system design so as to assure ease of implementation. The researcher acknowledges that the evaluation approach has limitations as the results are not based on a large set of participants and video files with varied contextual dimensions. There were difficulties in getting a larger group of participants within the time constraint of the research work to focus and annotate the video files, since they were not domain experts but general web users. The evaluation task required some efforts and focus from the participants. It might be difficult to have them perform the task on a large set of video as their focus can lessen and result in collection of inconsistent data. Future research work will replicate the same evaluation scenario with a large set of video (like the TRECVID evaluation benchmark) drawn from various subject matter areas in order to validate the initial findings. Future work to implement and integrate the S-Space model into the CONMAN prototype might eliminate the need to evaluate with human participants

thereby making it easier to test with a large dataset.

7.4 Future Work

The thesis has investigated a context approach to understanding and representing the semantics of video data as an instance of multimedia. This approach has opened up a new dimension that requires further research. While there are many avenues for future work, the most relevant are:

- further evaluation of the completed framework
- Multi-lingual support
- Standard context-aware multimedia representation scheme
- Semantic Retrieval

The identified areas of future work are discussed.

Further evaluation of the completed Framework: In addressing the limitations discussed in section 7.3, future work will be required to fully implement the framework and context model. The developed and formalised S-Space model for automatic video semantic recognition has to be implemented to eradicate the possible subjectivity and bias introduced by human annotators (Kender and Naphade, 2005). Implementing all the components of the framework including the retrieval subsystem, will facilitate evaluation using a large data set such as that available in the TRECVID evaluation benchmark. The evaluation can be based on traditional Information Retrieval metrics like precision, recall, fall-out, F-measure, etc. (Ren and Bracewell, 2009; ElAlami, 2011; Gennaro *et al.*, 2011) in order to validate the initial findings.

Multi-lingual support: The Internet cuts across national boundaries and content are

being generated and accessed in different languages. Each language has its own peculiarities in terms of clarity and consistency in semantic expression. There is need for further research into achieving cross-lingual retrieval. For instance, annotation metadata can be represented and indexed in English, while users can pose their search queries in French and be able to have the content found and delivered to them French. Perhaps, research should focus on extending the context model to provide cross-language support.

Standard context-aware multimedia representation scheme: The limitations of MPEG-7 in describing precise semantics had triggered interest in adopting some machine-understandable semantic languages in order to make semantics reusable and interoperable with other domains (Hunter, 2005; Tous and Delgado, 2010). Considerable progress have been recorded in various research endeavours with different multimedia ontologies (such as COMM[54]) emerging. However, with the context-based approach to multimedia annotation and retrieval presented in this thesis, there is need to investigate the inclusion of context in these ontologies to achieve proper context-based interoperability. The context-aware multimedia representation scheme will ensure that media semantics are not lost but captured and represented at the right level of contextual knowledge.

Semantic Retrieval: The retrieval component of any multimedia annotation and retrieval framework is very important to the overall user experience. A well-managed semantic annotation system may not be complete if users are not able to pose their search queries and get good retrieval results based on their search needs. Research is required to investigate possible combination of text mining techniques with the metadata representation of the visual information. More research is required to

54 http://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/comm/Ontology/

investigate the possibility of identifying other context knowledge sources that can impact and improve user and query modelling towards achieving semantic retrieval. User queries should be matched against the metadata repository to retrieve results that reflects human perceptual expectation.

References

Alcic S. and Conrad S. (2010), 'Measuring performance of web image content extractor', In Proceedings of the *Tenth International Workshop on Multimedia Data Mining*, ACM, New York, USA, pp. 8:1–8:8.

Aly R., Doherty A. R., Hiemstra D., de Jong F., and Smeaton A. F. (2012) 'The uncertain representation ranking framework for concept-based video retrieval' *Information Retrieval 15*, Springer, pp. 1-27. ISSN 1386-4564.

Antani S., Kasturi R., and Jain, R. (2002). 'A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video', *Pattern Recognition,* Vol. 35 pp. 945-965.

Apostoloff N. and Fitzgibbon A. (2006). 'Automatic Video Segmentation using Spatiotemporal T-junctions'. In Proceedings of the *17th British Machine Vision Conference, Edinburgh*, pp. 1089-1098.

Aytar Y., Shah M., and Luo J. (2008). 'Utilizing Semantic Word Similarity Measures for Video Retrieval', *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska.

Bach J. R., Fuller C., Gupta A., Hampapur A., Horowitz B., Humphrey R., Jain R., and Shu C.-F. (1996) 'The Virage image search engine: An open framework for image management', In Proceedings of *SPIE-96, 4th SPIE Conference on Storage and Retrieval for Still Images and Video Databases*, San Jose, US, pp 76 -87.

Baeza-Yates R. and Ribeiro-Neto B. (1999). Modern Information Retrieval, Addison Wesley Publishing, New York.

Baskerville R. and Myers M. D. (2004) 'Special Issue on Action Research in Information Systems: Making IS Research Relevant to Practice-Foreword', *MIS Quarterly* (28:3), pp. 329-335.

Beck K., *et al.* (2001) "Principles behind the Agile Manifesto". Agile Alliance. [Online] Available at: http://www.agilemanifesto.org/principles.html [Accessed 18 November 2004].

Belkin N. J. (1997) 'User modelling in information retrieval'. Tutorial presented in sixth *international conference on user modelling*, Chia Laguna, Sardinia. [Online] Available at: http://www.um.org/um_97/contents.html [Accessed 12 May 2005].

Belotti R., Decurtins C., Grossniklaus M., Norrie M. C., and Palinginis A. (2004), 'Modelling Context for Information Environments', in Baresi L., Dustdar S., Gall H. and Matera M. (Eds.), 'UMICS', Springer, pp. 43-56.

Ben-Gal I. (2007). 'Bayesian Networks'. In Ruggeri F., Kennett R. S., Faltin F. W. (*Eds*), Encyclopedia of Statistics in Quality and Reliability, John Wiley & Sons.

Benbasat I. And Zmud R. (1999). Empirical research in information systems: The practice of relevance. MIS Quarterly, 23(1), 3-16.

Benitez A. B., Smith J. R., and Chang S.-F. (2000). 'MediaNet: A Multimedia Information Network for Knowledge Representation', *SPIE Conference on Internet Multimedia Management Systems (IS&T/SPIE-2000)*, Vol. 4210, Boston, MA.

Benitez A. B., Chang S.-F., and Smith J. R. (2001). 'IMKA: A Multimedia Organization System Combining Perceptual and Semantic Knowledge', *ACM International Conference on Multimedia (ACM MM-2001)*, Canada, Ottawa.

Berners-Lee T., Hendler J., and Lassila, O. (2001). 'The Semantic Web', Scientific American, 284(5), pp 34–43.

Berry M. W, Drmac Z., and Jessup E. R. (1999). Matrices, vector spaces, and information retrieval. SIAM Review 41(2) 335–362.

Bimbo, A. D. (1999). Visual Information Retrieval, Morgan Kaufmann, San Francisco, CA.

Biswas P. (2012). A Brief Survey on User Modelling in Human Computer Interaction. In Tiwary U and Siddiqui T.(*Eds*.), Speech, Image, and Language Processing for Human Computer Interaction: Multi-Modal Advancements, pp. 1-19, IGI Global.

Black S. E., Boca P. P., Bowen J. P., Gorman J., and Hinchey M. G. (2009) 'Formal versus agile: Survival of the fittest'. *IEEE Computer* 49 (9): 39–45.

Bloehdorn S., *et al.* (2005), 'Semantic annotation of images and videos for multimedia analysis', Lecture notes in computer science - The semantic web: research and applications, vol. 3532, Springer, pp. 592–60.

Bock G. W., *et al.* (2005). 'Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate'. MIS Quarterly, 29, 1, pp. 87–111.

Borlund P., Schneider J. W., Lalmas M., Tombros A., Feather J., Kelly D., de Vries A., and Azzopardi L. (2008), 'Information Interaction in Context'. Proceedings of the 1st IIiX *Symposium on Information Interaction in Context*, 14–17 October 2008, London, UK ACM, New York, NY, USA.

Bradford R., (2008) 'An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications', Proceedings of the *17th ACM Conference on Information and Knowledge Management*, Napa Valley, California, USA, pp. 153–162.

Brin S. and Page L. (1998). 'The anatomy of a large-scale hypertextual Web search engine'. Computer Networks and ISDN Systems 30(1–7), pp. 107–117.

Brooks, F. (1996). "The Computer Scientist as Toolsmith II." Communications of the ACM 39(3): 61-68.

Brunelli R., *et al.* (1996). 'A survey on video indexing', *IRST-Technical Report* 9612-06.

Burns P. R. (2006). 'MorphAdorner: Morphological Adorner for English Text'. Available at: http://morphadorner.northwestern.edu/morphadorner/documentation/ [Accessed 28 October 2011].

Burrill V. (1994). 'Group I Report: Authoring Systems', In *Multimedia Systems and Applications*, Encarnacao, J. L. and Foley, J. D. (*Eds*), Springer-Verlag, pp. 11-23.

Caplinskas A. and Vasilecas O. (2004) 'Information systems research methodologies and models'. Proceedings of the International Conference on Computer Systems and Technologies, CompSysTech'2004.

Carrer M., Ligresti L., Ahanger G., and Little, T. D. C. (1997), 'An Annotation Engine for Supporting Video Database Population.', Multimedia Tools Appl. 5 (3) , 233-258.

Celma O., *et al.* (2007). 'MPEG-7 and the Semantic Web'. WSC Incubator Group Editor's Draft. [Online] Available at: http://www.w3.org/2005/Incubator/mmsem/XGR-mpeg7-20070814/ [Accessed 03 January 2012].

Chang S.F., Smith J.R., Beigi M., and Benitez A.B. (1997). 'Visual Information Retrieval from Large Distributed On-line Repositories', *Communications of the ACM*, Vol. 40, No. 12, pp. 63-71.

Chang H. S., Sull S., and Lee S-U. (1999). 'Efficient video indexing scheme for content-based retrieval', *IEEE Trans. Circuits System. Video Technology*. Vol. 9 (8) pp. 1269–1279.

Chi-Chun L. and Shuenn-Jyi W. (2001). 'Video segmentation using a histogram-based fuzzy c-means clustering algorithm'. Computer Standards Interfaces, 3(5), pp.920-923.

Choi J. G., *et al.* (1997). 'Video segmentation based on spatial and temporal information', *IEEE International Conference on Acoustic, Speech, Signal Processing*, ICASSP'97, pp. 2661- 2664, Munich, Germany.

Choroś K. and Gonet M. (2008). 'Effectiveness of Video Segmentation Techniques for Different Categories of Videos'. In Proceedings of the conference on New Trends in Multimedia and Network Information Systems, Aleksander Zgrzywa, Kazimierz Choroś, and Andrzej Siemiński (*Eds*). IOS Press, Amsterdam, The Netherlands, pp. 34 - 45.

Chung E., and Yoon J. (2010). 'An exploratory analysis on unsuccessful image searches'. Proceedings of the American Society for Information Science and Technology, 47(1), 1–2.

Chung R. H. Y., Chin F. Y. L., Wong K.-Y. K., Chow K. P., Luo T., and Fung H. S. K. (2005), 'Efficient Block-based Motion Segmentation Method using Motion Vector Consistency', Proceedings of Conference in Machine Vision Applications, pp. 550-553.

Colombo C., Bimbo A. D., and Pala P. (1999). 'Semantics in visual information retrieval', *IEEE Multimedia,* Vol. 6 (3) pp. 38–53.

Conklin K. R. (1974). 'Knowledge, Proof, and Ineffability in Teaching'. Educational Theory, 24: 61–67.

Corcho O. *et al.* (2003). 'Methodologies, tools, and languages for building ontologies. Where is their meeting point?', *Data & Knowledge Engineering*, Vol 46, pp 41-64.

Cosmin S. (2010). 'A Multimedia Database Server: Implementation and Functions'. International Journal of Computer Science and Applications, Technomathematics Research Foundation, Vol. 7, No. 3, pp. 140 – 155.

Dalakleidi K., *et al.* (2011). 'Semantic Representation of Multimedia Content', Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, pp. 18-49.

Dasiopoulou S., *et al.* (2011). 'A Survey of Semantic Image and Video Annotation Tools', in Book '*Knowledge-driven Multimedia Information Extraction and Ontology Evolution*', 6050/BOEMIE EU Project, G. Paliouras, C.D. Spyropoulos, G. Tsatsaronis (Eds.), Springer Verlang.

Datta R., Joshi D., Li J., and Wang J. Z. (2008), 'Image Retrieval: Ideas, Influences, and Trends of the New Age', ACM Computing Surveys 40 (2), pp. 1-60.

Deardorff T. D. C., *et al.* (1994). 'Video Scene Decomposition with the Motion Picture Parser', In Proceedings of the *IS&T/SPIE Sysmposium on Electronic Imaging Science and Technology (Digital Video compression and Processing on Personal Computers: Algorithms and Technologies)*, San Jose, CA, No 2187 in SPIE, pp. 44-55.

Deb S. (2004). Multimedia systems and content-based image retrieval, Idea Group Publishing, USA.

Denning S. (1998). 'What is Knowledge Management?', A Background Paper to the World Development Report, [Online], Available at: http://www.stevedenning.com/knowledge.htm [Accessed 16 August 2004].

Dey A. K. (2000). 'Providing Architectural Support for Building Context-Aware Applications'. PhD thesis, Georgia Institute of Technology.

Dimitrova N. (1999). 'Multimedia content analysis and indexing for filtering and retrieval applications', Special Issue on *Multimedia Informing Technologies* – Part 1, Vol. 2 No 4.

Dodig-Crnkovic G. (2002). 'Scientific Methods in Computer Science', in Conference for the Promotion of Research in IT at New Universities and at University Colleges in Sweden.

Dodig-Crnkovic G. (2010), 'Constructive Research and Info-computational Knowledge Generation,' in Magnani L., Carnielli W., and Pizzi C. (*Eds.*), *Model-Based Reasoning In Science And Technology Abduction Logic and Computational Discovery Conference*,

Springer Berlin / Heidelberg, Vol. 314, pp. 359-380.

Donderler M. E., Ulusoy O., and Gudukbay U. (2004). 'Rule-base spatiotemporal query processing for video databases', *The VLDB Journal,* Vol. 13, pp. 86-103.

Dong H., Hussain F. K., and Chang E. (2008). 'A Survey in Traditional Information Retrieval Models'. *IEEE Digital Ecosystems and Technologies*, pp.397-402.

Doulamis A. D., Doulamis N. D., and Kollias S. D. (2000). 'A fuzzy video content representation for video summarization and context-based retrieval', *Signal processing*, Vol. 80, pp. 1049-1067.

Doulamis N. D., Doulamis A. D., Avrithis Y. S., and Kollias S. D. (1998). 'Video content representation using optimal extraction of frames and scene, In Proceedings. *ICIP 98*, Vol. 1, pp. 875–879.

Duygulu P. and Bastan M. (2011), 'Multimedia translation for linking visual data to semantics in videos.', Mach. Vis. Appl. 22 (1) , 99-115 .

Eberman B., *et al.* (1999). 'Indexing multimedia for the Internet', Compaq Cambridge Research Laboratory, *Technical Report Series*, CRL 99/2.

Eidenberger H. (2011), Fundamental Media Understanding, atpress, Vienna, ISBN-13: 978-3842379176.

ElAlami M. E. (2011), 'Unsupervised image retrieval framework based on rule base system.', Expert Syst. Appl. 38 (4), 3539-3549.

Enser P. G. B. and Sandom C. J. (2002), “Retrieval of archival moving imagery – CBIR outside the frame?”, in Lew, M.S., Sebe, N. and Eakins, J.P. (Eds), *Proceedings*

of Image and Video Retrieval: International Conference, CIVR 2002, London, UK, July 18-19, 2002, Lecture Notes in Computer Science, Vol. 2383, Springer-Verlag, Berlin, pp. 206-14.

Enser P. G. B., Sandom C. J., Hare J. S., and Lewis P. H. (2007), "Facing the reality of semantic image retrieval", *Journal of Documentation*, Vol. 63 Issue: 4 pp. 465 – 481.

Enser P. G. B. (2008a), 'The evolution of visual information retrieval.', *Journal of Information Science* 34 (4) , 531-546.

Enser P. G. B. (2008b), 'Visual image retrieval'. In Cronin B. (Ed.) Annual Review of Information Science and Technology 42, Information Today, Inc., Medford, NJ, pp. 3–42.

Erdmann M. and Studer R. (2000), 'How to structure and access XML documents with ontologies'. Data and Knowledge engineering – Special issues on Intelligent Information Integration DKE (36) 3:317-335.

Eze E. and Ishaya T. (2007), 'S-Space: a Context-based Ontology Model for Multimedia Semantic Organisation and Discovery'. In proceedings of *1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies*, Genova, Italy.

Fan J., Luo H., and Elmagarmid A. K. (2004a), 'Concept-Oriented Indexing of Video Databases: Towards Semantic Sensitive Retrieval and Browsing', *IEEE Transactions on Image Processing*, Vol.13, No.5, pp. 974-992.

Fan J., Zhu X., Elmagarmid A. K., Aref W. G., and Wu L. (2004b), 'ClassView: Hierarchical Video Shot Classification, Indexing, and Accessing', *IEEE Trans. on Multimedia*, Vol.6, No 1, pp. 70-86.

Fauzi F., Hong J. L., and Belkhatir M. (2009), 'Webpage segmentation for extracting images and their surrounding contextual information', in Gao W., Rui Y., Hanjalic A., Xu C., Steinbach E. G., El-Saddik A., and Zhou M. X. (Eds.), ACM Multimedia, pp. 649-652.

Fauzi F. and Belkhatir M. (2013), 'Multifaceted conceptual image indexing on the world wide web'. Information Processing and Management, 49, 2 , pp. 420-440.

Fei-Fei, L., Fergus, R., and Perona, P. (2006), 'One-shot learning of object categories'. IEEE Trans. Pattern Anal. Mach. Intell. 28(4), 594–611.

Fellbaum C., (1998), 'WordNet, An electronic Lexical Database', MIT Press.

Ferecatu M., Boujemaa N., and Crucianu M. (2008), 'Semantic interactive image retrieval combining visual and conceptual content description'. ACM Multimedia Systems Journal, Vol. 13, No. 5-6, p. 309-322.

Ferman M. A., Tekalp A. M., and Mehrotra R. (2002), 'Robust Color Histogram Descriptors for Video Segment Retrieval and Identification', *IEEE Transactions on Image Processing*, Vol. 11, No 5, pp. 497-507.

Fernando S. and Stevenson M. (2008), 'A semantic similarity approach to paraphrase detection'. Computational Linguistics UK (CLUK 2008) 11th Annual Research Colloquium.

Flinker M., *et al.* (1995), 'Query by image and video content: The QBIC system', *IEEE Computer*, 28(9), 23-32.

García R. and Celma O. (2005), 'Semantic Integration and Retrieval of Multimedia Metadata', in Proceedings of the *5th International Workshop on Knowledge Markup and Semantic Annotation*.

Gargi U., Kasturi R., and Strayer S. H. (2000), 'Performance Characterisation of Video-Shot-Change Detection Methods', *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 10, No 1, pp. 1-13.

Gennaro C., *et al.* (2011), 'A unified multimedia and semantic perspective for data retrieval in the semantic web'. *Information Systems Journal*, 36(2), pp. 174-191. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0306437910000645.

Goker A. and Davies J. (2009) *Information Retrieval: Searching in the 21st Century*. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622.

Goker A., Myrhaug H., and Bierig R. (2009) 'Context and Information Retrieval', In: Goker, A. and Davies, J. (*Eds*). Information Retrieval: Searching in the 21st Century. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, pp. 131-157.

Gómez-Romero J. *et al.* (2011), 'Ontology-based context representation and reasoning for object tracking and scene interpretation in video'. *Expert Systems with Applications*, 38(6), p.7494-7510.

Gregor S. and Hevner A. R. (2011), 'Introduction to the special issue on design science.', *Inf. Syst. E-Business Management* 9 (1) , pp. 1-9.

Griffiths T. L. and Yuille A. (2006), 'A primer on probabilistic inference', Trends in Cognitive Sciences Supplement to special issue on Probabilistic Models of Cognition, 10(7), pp. 1–11.

Gross T. and Klemke R. (2003), 'Context Modelling for Information Retrieval - Requirements and Approaches', *IADIS Intl. Journal on WWW/Internet* 1 (1) , 29-42.

Gruber T. R. (2009), 'Ontology' in the Encyclopedia of Database Systems, Ling Liu and M. Tamer Özsu (*Eds*), Springer-Verlag.

Grubinger M., Clough P., Müller H., and Deselaers T. (2006), 'The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems', In *Proceedings of International Workshop OntoImage'2006 Language Resources for Content-Based Image Retrieval*, held in conjunction with LREC'06, Genoa, Italy, pp. 13 - 23.

Grudin J. and Bargeron D. (2005), 'Multimedia Annotation: An Unsuccessful Tool Becomes a Successful Framework', In Okada K., Hoshi T., and Inoue T. (Eds.), Communication and Collaboration Support Systems, pp. 62-76.

Hada Y., (2012), 'Development of a Mobile Video Learning System for Cellular Phone', IEEE Seventh International Conference on Wireless, Mobile and Ubiquitous Technology in Education (WMUTE), pp. 205 – 207.

Handschuh S., Staab S., and Studer R. (2003), 'Leveraging metadata creation for the Semantic Web with CREAM', KI ’2003 - advances in artificial intelligence, in Proceedings of the Annual German Conference on AI, September 2003.

Hanjalic A. and Xu L.-Q. (2005), 'Affective video content representation and modeling', *IEEE Transactions on Multimedia*, vol. 7, pp. 143–154.

Hare, J. S. *et al.* (2006a), 'Mind the Gap: Another look at the problem of the semantic gap in image retrieval', in *Proceedings of Multimedia Content Analysis, Management*

and Retrieval 2006 SPIE.

Hare J. S. *et al.* (2006b), 'Bridging the Semantic Gap in Multimedia Information Retrieval: Top-down and Bottom-up approaches'. At Mastering the Gap: From Information Extraction to Semantic Representation; *3rd European Semantic Web Conference*, Budva, Montenegro.

Hauptmann A. G. (2004), 'Towards a Large Scale Concept Ontology for Broadcast Video', in Proceedings of the *3rd International Conference on Image and Video Retrieval (CIVR'04)*, pp. 674-675.

Hauptmann A., Yan R., Lin W.-H., Christel M., and Wactlar H. D. (2007), 'Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News.', IEEE Transactions on Multimedia 9 (5), pp. 958-966.

Heggland J. (2005), 'OntoLog: Flexible Management of Semantic Video Content Annotations', Doctoral thesis, Norwegian University of Science and Technology.

Heryanto H., Akbar S., and Sitohang B. (2011), 'Direct access in content-based Audio Information Retrieval: A state of the art and challenges'. *ICEEI*, pp. 1-6.

Hider P. (2006), 'Search goal revision in models of information retrieval.', Journal of Information Science 32 (4) , 352-361.

Hiemstra D. (2009), 'Information Retrieval Models', In: Goker, A., and Davies, J. (Eds). *Information Retrieval: Searching in the 21st Century*. John Wiley and Sons, Ltd., ISBN-13: 978-0470027622, pp. 1-19.

Higgins D. and Burstein J. (2007), 'Sentence similarity measures for essay coherence Introduction: Text coherence in student essays'. Proceedings of the 7th International Workshop on Computational Semantics IWCS, January, pp.1-12.

Hollink L, Worring M, and Schreiber G. (2005), 'Building a visual ontology for video retrieval', In Proceedings of the *ACM Multimedia*, Singapore, November 2005.

Hoogs A., Rittscher J., Stein G. C., and Schmiederer J. (2003), 'Video Content Annotation Using Visual Analysis and a Large Semantic Knowledgebase', in *Computer Vision and Pattern Recognition (CVPR) (2)*, IEEE Computer Society, pp. 327-334.

Hu W., Xie N., Li L., Zeng X., and Maybank S. J. (2011), 'A Survey on Visual Content-Based Video Indexing and Retrieval'. *Systems Man and Cybernetics Part C Applications and Reviews IEEE Transactions,* Vol 41, No 6, pp. 797–819.

Huang T. S., Mehrotra S., and Ramchandran K. (1996), 'Multimedia Analysis and Retrieval System (MARS) Project'. In In Proceedings of Data Processing Clinic.

Huang W., Tao T., Hacid M. S., and Mille A. (2003), 'Facilitate knowledge communications in multimedia e-learning environments', *Proceedings of ACM MMDB2003*, New Orleans, USA.

Huang W. and Tao T. (2004), 'Adding context-awareness to knowledge management in modern enterprises', In Proceedings of the *2004 IEE International Conference on Intelligent Systems*, Varna, Bulgaria.

Huang W., Eze E., and Webster D. (2006), 'Towards integrating semantics of multi-media resources and processes in e-Learning.', Journal on Multimedia Systems, 11 (3), pp. 203-215.

Hunter J., Schroeter R., Koopman B., and Henderson M. (2004), 'Using the semantic grid to build bridges between museums and indigenous communities', in Proceedings of the GGF11—Semantic Grid Applications Workshop, Honolulu, June 10, 2004.

Hunter J. (2005), 'Adding Multimedia to the Semantic Web - Building and Applying an MPEG-7 Ontology', Chapter 3 of "Multimedia Content and the Semantic Web", Eds. Giorgos Stamou and Stefanos Kollias, Wiley.

Idris F. and Panchanathan S. (1997), 'Review of image and video indexing techniques', *Journal of Visual Communication and Image Representation*, Vol. 8(2), pp. 146-166.

Iivari J., (2010), 'Twelve Theses on Design Science Research in Information Systems', Hevner A and Chatterjee S. (Eds). *Design Research in Information Systems*, 22(2004), pp. 43-62.

Ingwersen P. and Jarvelin K. (2005), The Turn: Integration of Information Seeking and Retrieval in Context. The Netherlands, Springer. ISBN:140203850X.

Inoue M. (2009), 'Image retrieval: research and use in the information explosion', Progress in Informatics, No 6, pp.3–14.

Izquierdo E., Chandramouli K., Grzegorzek M., and Piatrik T. (2007), 'K-Space Content Management and Retrieval System', in Proceedings of *14th International Conference on Image Analysis and Processing*, ICIAPW 2007, pp.131 -136.

Jaimes A. and Chang S.-F. (2000), 'A Conceptual Framework for Indexing Visual Information at Multiple Levels', in *IS&T/SPIE Internet Imaging*, Vol. 3964.

Jasinschi R. S., *et al.* (2002). 'A probabilistic layered framework for integrating multimedia content and context information', *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Orlando, Florida, USA.

Jian-Kang W., *et al.* (1995) 'CORE: A Content-Based Retrieval Engine for Multimedia Information Systems', *Journal for Multimedia System*, pp. 25 - 41.

Jiang Y.-G., Wang J., Chang S.-F., and Ngo C.-W. (2009), 'Domain adaptive semantic diffusion for large scale context-based video annotation', in Proceedings of the *12th IEEE International Conference on Computer Vision*, Kyoto, pp. 1420-1427.

Jones D. and Gregor S. (2007), The anatomy of a design theory. *Journal of the Association for Information Systems (JAIS)*, 8(5), Article 19.

Juan K. and Cuiying H. (2010), 'Content-based video retrieval system research', In proceedings of *3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT)*, pp. 701 - 704, Volume: 4.

Kang H-B. (2002), 'Analysis of Scene Context related with Emotional Events', In proceedings of *Multimedia '02,* December 1-6, Juan-les-Pins, France, pp 311-314.

Kaur A., Gulati S., and Singh S. (2012), 'Analysis of Three Formal Methods-Z, B and VDM' *International Journal of Engineering Research & Technology (IJERT)*, Vol. 1 Issue 4, ISSN: 2278-0181.

Kender J. R., and Naphade M. R. (2005), 'Visual concepts for news story tracking: analyzing and exploiting the NIST TRECVID video annotation experiment'. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 1174–1181.

Khan L. and McLeod D. (2000), 'Audio Structuring and Personalized Retrieval Using Ontologies,' in Proceedings of *IEEE Advances in Digital Libraries (ADL 2000)*, Washington DC.

Klien D. and Manning C. D. (2002), 'A Generative Constituent-Context Model for Improved Grammar Induction. In Proceedings of the *40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, pp. 128-135.

Klusch M. (2001), 'Information agent technology for the Internet: A survey', *Data & Knowledge Engineering*, Vol. 36, pp. 337-372.

Kock N. F., McQueen R. J., and Scott J. L. (1997), 'Can Action Research be Made More Rigorous in a Positivist Sense? The Contribution of an Iterative Approach', *Journal of Systems and Information Technology*, (1:1), pp. 1-24.

Koenen R., (*ed.*) (1999). 'MPEG-7: Context, objectives and technical roadmap, V12', *ISOLIEC JTCI/SC29/WG11 MPEG99/N2861*, Vancouver.

Koprinska I. and Carrato S. (2001), 'Temporal video segmentation: A survey', *Signal Processing: Image Communication,* Vol. 16, No. 5, pp. 477-500.

Küçüktunç O., Güdükbay U., and Ulusoy Ö. (2010), 'Fuzzy color histogram-based video segmentation'. Computer Vision and Image Understanding, 114(1), pp.125-134.

Kuechler B. and Vaishnavi V. (2011), 'Promoting relevance in IS research: an informing system for design science research'. *Informing Science: the International Journal of an Emerging Transdiscipline* 14:125–138.

Kuhn R., Chandramouli R., and Butler R., (2002), Cost effective use of formal methods in verification and validation, in: Foundations 02 Workshop on Verification & Validation.

Lashkari A. H., Mahdavi F., and Ghomi V. (2009), 'A Boolean Model in Information Retrieval for Search Engines', In Proceedings of International Conference on Information Management and Engineering (ICIME), 2009, pp. 385 - 389.

Lavee G., Rivlin E., and Rudzsky, M. (2009), 'Understanding video events: A survey of methods for automatic interpretation of semantic occurrences in video', IEEE Trans. Syst., Man, Cybern. C, Appl. Rev., vol. 39, no. 5, pp. 489-504.

Lee H-C. and Kim S-D. (2003), 'Iterative key frame selection in the rate-constraint environment', *Signal Processing: Image Communication*, Vol. 18. pp 1-15.

Lefevre S., Holler, J., and Vincent N. (2003), 'A review of real-time segmentation of uncompressed video sequences for content-based search and retrieval', *Real-Time Imaging*, Vol. 9, Issue 1, pp. 73-98.

Lew M. S., *et al.* (2006), 'Content-based multimedia information retrieval: State of the art and challenges'. ACM Transactions on Multimedia Computing Communications and Applications TOMCCAP, 2(1), pp.1-19.

Liao S-H. (2002), 'Knowledge management technologies and applications—literature review from 1995 to 2002', *Expert Systems with Applications*, Vol. 25, pp. 155–164.

Lin D. (1998), 'An information-theoretic definition of similarity', In Proceedings of the International Conference on Machine Learning.

Liu Y. et al. (2007), 'A survey of content-based image retrieval with high-level semantics'. Pattern Recognition 40(1). pp. 262-282 .

Lo C. and Wang S. (2003), 'A Histogram-based Moment-Preserving Clustering Algorithm for Video Segmentation', *Pattern Recognition Letters,* Vol. 24, pp. 2209–2218.

Lord P. W., *et al.* (2003), 'Semantic similarity measures as tools for exploring the gene ontology'. Pacific Symposium On Biocomputing, 612, pp. 601-612.

Loyola W. (2007), 'Comparison of Approaches toward Formalising Context: Implementation Characteristics and Capacities', *The Electronic Journal of Knowledge Management* Volume 5 Issue 2, pp 203-214.

Luan H., Zheng Y.-T., Wang M., and Chua T.-S. (2011), 'VisionGo: Towards video retrieval with joint exploration of human and computer.', Inf. Sci. 181 (19), 4197-4213.

Luck M., McBurney P., and Preist, C. (2003), Agent Technology: Enabling Next Generation Computing (A Roadmap for Agent Based Computing), *AgentLink*.

Martinez J. M. (2003), 'Coding of moving pictures and audio', [Online] Available at: http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm [Accessed June 12 2005].

McCarthy J. and Buvac S. (1998), 'Formalizing context (expanded notes)', *Technical Report*, Stanford University.

McTear M. (1993), 'User Modelling for adaptive computer system: a survey of recent developments', *Artificial intelligence review* 7, 157-184.

Mech R. and Wollborn M. (1997), 'A noise robust method for segmentation of moving objects in video sequence', *IEEE International Conference on Acoustic, Speech, Signal Processing, ICASSP'97*, pp. 2657-2660, Munich, Germany.

Meier T. and Ngan K. N. (1999), 'Video segmentation for content-based coding', *IEEE Trans. Circuits System. Video Technology*, Vol. 9, pp. 1190-1203.

Meland *et al.* (2003), 'Using Ontologies and Semantic Networks with Temporal Media', In proceedings of the Semantic Web Workshop at the 26th Annual International ACM SIGIR Conference, Toronto, Canada.

Mihalcea R., Corley C., and Strapparava C. (2006), 'Corpus-based and Knowledge-based Measures of Text Semantic Similarity', in Proceedings of AAAI 2006, Boston, July.

Miller G. A. (1999), 'WordNet: A Lexical Database', Communications of the ACM, 38 (11).

Montresor A., *et al.* (2002), 'Messor: Load-balancing a swarm of autonomous agents', *Technical Report UBLCS-2002-11*, [Online], Available at: http://www.cs.unibo.it [Accessed December 21 2003].

Moreno P. J., *et al.* (2002), 'From multimedia retrieval to knowledge management', *Cambridge Research Laboratory Technical Report*.

Moscheni F., *et al.* (1998), 'Spatial temporal segmentation based on region merging', *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 20, pp. 897-915.

Moutselakis E. V. and Karakos A. S. (2009), Semantic Web Multimedia Metadata Retrieval: A Music Approach. 2009 13th Panhellenic Conference on Informatics, pp. 43 - 47.

Mylonas P., Athanasiadis T., Wallace M., Avrithis Y., and Kollias S. (2008), 'Semantic Representation of Multimedia Content: Knowledge Representation and Semantic Indexing', Multimedia Tools and Applications 39 (3) , pp. 293-327.

Myrhaug H. I. and Goker A. (2003), 'AmbieSense – Interactive Information Channels in the Surroundings of the Mobile User'. *2nd International Conference on Universal Access in Human – Computer Interaction*, Crete, Greece, Lawrence Erlbaum Associates, 1158–1162, July 2003.

Nagasaka A. and Tanaka Y. (1991), 'Automatic video indexing and full-video search for object appearances', In Proceedings of *2nd Working Conference on Visual Database Systems*, pp. 119–133.

Nagasaka A. and Tanaka Y. (1992), 'Automatic Video Indexing and Full Video Search for Object Appearances', *IFIP Transactions on Visual Database Systems II*, Knuth, E. and Wegner, L. M. (Eds.), pp. 113-127.

Natsev A., Haubold A., Tesic J., Xie L., and Yan R. (2007), 'Semantic concept-based query expansion and re-ranking for multimedia retrieval', in Lienhart R., Prasad A. R., Hanjalic A., Choi S., Bailey B. P., and Sebe N. (Eds), *ACM Multimedia*, pp. 991-1000.

Nunes V. T., Santoro F. M., and Borges M. R. S. (2009), 'A context-based model for Knowledge Management embodied in work processes.', Information Science 179 (15), pp. 2538 - 2554.

Over P., Ianeva T., Kraaij W., and Smeaton, A. (2005), 'Trecvid 2005 - an overview'. In Proceedings of TRECVID 2005. NIST, USA.

Paliouras G., Spyropoulos C. D., and Tsatsaronis G. (*Eds.*) (2011), Knowledge-Driven Multimedia Information Extraction and Ontology Evolution, Bridging the Semantic Gap, Lecture Notes in Computer Science, Vol. 6050, 1st Edition., IX, ISBN 978-3-642-20794-5.

Panofsky E. (1962), *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*, Harper and Rowe, New York, NY.

Papadopoulos G., Briassouli A., Mezaris V., Kompatsiaris I., and Strintzis M. G. (2009), 'Statistical Motion Information Extraction and Representation for Semantic Video Analysis', *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1513-1528, October 2009.

Patel B. V. and Meshram B. B. (2012), 'Content based video retrieval systems', International Journal of UbiComp (IJU), Vol.3, No.2, pp. 13-30.

Paz-Trillo C., *et al.* (2004), 'Using Ontologies to Retrieve Video Information', submitted to *Workshop on Ontologies and their Applications* in Sao Luis, Brazil.

Pedersen T., Patwardhan S., and Michelizzi J. (2004). 'WordNet::Similarity - Measuring the Relatedness of Concepts'. In the Proceedings of the Nineteenth National Conference on Artificial Intelligence (AAAI-04), pp. 1024-1025, San Jose, CA.

Pérez E., Fortier A., Rossi G., and Gordillo S. E. (2009), 'Rethinking Context Models', in Robert Meersman R., Herrero P., and Dillon T. S. (*Eds.*), OTM Workshops, Springer, pp. 78-87.

Petasis G., Fragkou P., Theodorakos A., Karkaletsis V., and Spyropoulos C. D. (2008), 'Segmenting HTML pages using visual and semantic information', The 4th Web as Corpus Workshop: Can we do better than Google?, at the 6th *International Conference on Language Resources and Evaluation (LREC)*, Marrakech, Morocco, May 2008.

Petrelli, D. and Auld, D. (2008), 'An examination of automatic video retrieval technology on access to the contents of an historical video archive', *Program: electronic library and information systems*, Vol. 42 Iss: 2 pp. 115 - 136

Petridis K. *et al.* (2006), 'Knowledge Representation and Semantic Annotation of Multimedia Content', Proceedings on Vision Image and Signal Processing, Special issue on Knowledge-Based Digital Media Processing, Vol. 153, No. 3, pp. 255-262.

Pino C. and Di Salvo R. (2011), 'A survey of semantic multimedia retrieval systems'. In Proceedings of the 13th WSEAS international conference on Mathematical and computational methods in science and engineering (MACMESE'11), Metin Demiralp, Zoran Bojkovic, and Angela Repanovici (*Eds.*). World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA, 353-358.

Polydoros P., Tsinaraki C., and Chirstodoulakis S. (2006), 'GRAPHONTO: OWL-Based Ontology Management and Multimedia Annotation in the DS-MIRF Framework', in Proceedings of the 4th Special Workshop on Multimedia Semantics, Chania, Crete, Greece. pp. 14-24.

Ponceleon D. B., Srinivasan S., Amir A., Petkovic D., and Diklic D. (1998), 'Key to Effective Video Retrieval: Effective Cataloging and Browsing', in Wolfgang Effelsberg & Brian C. Smith (Ed.), *ACM Multimedia*, ACM, pp. 99-107.

Quelhas P., *et al.* (2007), 'A thousand words in a scene'. IEEE Trans. Pattern Anal. Mach. Intell. 29(9), 1575–1589.

Rafferty P. and Hidderley R. (2005), *Indexing multimedia and creative works: the problems of meaning and interpretation*. Aldershot, Hants, England, Ashgate.

Raghavan V. V. and Wong S. K. M. (1986), 'A critical analysis of vector space model for information retrieval', Journal of the American Society for Information Science, Vol.37 (5), pp. 279-87.

Ramesh V., Glass R. L., and Vessey I. (2004), 'Research in computer science: an empirical study.', *Journal of Systems and Software* 70 (1-2) , 165-176.

Rashid U., Niaz I. A., and Bhatti M. A. (2009), 'M3L: Architecture for Multimedia Information Retrieval'. ITNG'2009. pp.1067-1072.

Razikin K., Keng-Tiong T., Dion H. G., Chua A. Y. K., Chei-Sian L. (2011), 'SPLASH: Perspectives on Mobile Socializing, Playing and Content Sharing,' Eighth International Conference on Information Technology: New Generations (ITNG), pp. 873 - 878.

Ren F. and Bracewell, D. (2009), 'Advanced Information Retrieval'. Electronic Notes in Theoretical Computer Science (ENTCS), 225, p.303-317.

Rocchio J. (1971), Relevance feedback in information retrieval. In G. Salton (ed.), The Smart Retrieval System: Experiments in Automatic Document Processing, pp. 313–323. Prentice Hall.

Rorissa A., Clough P., and Deselaers T. (2008), 'Exploring the relationship between feature and perceptual visual spaces.', Journal of the American Society for Information Science and Technology 59 (5), pp. 770-784.

Ruger S. (2011), 'Multimedia Resource Discovery', In: Melucci M and Baeza-Yates R. (Eds). *Advanced Topics in Information Retrieval*. Springer, New York, pp. 157-186.

Rui Y., Huang T. S., Ortega M., and Mehrotra S. (1998), 'Relevance feedback: A power tool in interactive content-based image retrieval', *IEEE Trans. Circuits and Systems for Video Tech*., Vol.8(5), pp. 644-655.

Ruthven I. (2011), 'Information Retrieval in Context', in Melucci M. and Baeza-Yates R. (Eds.), Advanced Topics in Information Retrieval, Springer Berlin Heidelberg, pp. 195-216.

Sadallah, M., Aubert, O., and Prié, Y. (2011) 'Hypervideo and Annotations on the Web'. Proceedings of Multimedia on the Web Workshop MMWeb2011.

Salton G., Wong A., and Yang C. S. (1975), 'A vector space model for automatic indexing'. *Communications of the ACM*, 18 (11), 613 - 620.

Sarah De Bruyne *et al.* (2007), 'Temporal Video Segmentation on H.264/AVC Compressed Bitstreams'. MMM (1) pp. 1-12.

Sarangi L. and Panda C. (2010), 'A Review on Intelligent Agent Systems', Journal of Computer and Communication Technology, Vol. 2, No. 1, pp. 42-57, ISSN (Print): 0975-7449.

Schach R. (1999), *Software Engineering*, Fourth Edition, McGraw-Hill, Boston, MA.

Scheurer T. (1994), Foundations of Computing: systems development with set theory and logic, Addison-Wisley Publishing Company, England.

Schreiber A. *et al.* (2001), 'Ontology-Based Photo Annotation', *IEEE Intelligent Systems*, Vol 16(3), pp. 66-74, May-June 2001.

Shahrary B. and Gibbon D. C. (1995), 'Automatic generation of pictorial transcript of video programs', In Proceedings of *SPIE Digital Video Compression: Algorithms and Technologies*, San Jose, CA, pp. 512–519.

Shatford S. (1986), 'Analyzing the subject of a picture: a theoretical approach', Cataloguing & Classification Quarterly, Vol. 5 No. 3, pp. 39-61.

Simpson T. and Dao T. (2010), 'WordNet-based semantic similarity measurement'. Available at http://www.codeproject.com/KB/string/semanticsimilaritywordnet.aspx

Singh M. P. (2002), 'The Pragmatic Web', *IEEE Internet Computing*, Vol. 6, No 3, May/June, pp. 4-5.

Smeaton A. F., Over P., and Kraaij W. (2006), 'Evaluation campaigns and TRECVid', In Proceedings of the 8th *ACM International Workshop on Multimedia Information Retrieval* (Santa Barbara, California, USA, October 26 - 27, 2006, MIR '06. ACM Press, New York, NY, pp. 321 - 330.

Smeaton A., Over P., and Kraaij, W. (2009), 'High level feature detection from video in TRECVid: a 5-year retrospective of achievements'. In: Divakaran, A. (ed.) *Multimedia Content Analysis, Theory and Applications*. Springer, Berlin. pp 151-174.

Smeulders A., Worring M., Santini S., Gupta A., and Jain R. (2000), 'Content-Based Image Retrieval at the End of the Early Years.', IEEE Trans. Pattern Anal. Mach. Intell. 22 (12) , 1349-1380.

Smith J. R. and Chang S.-F. (1997), 'Visually Searching the Web for Content.', IEEE MultiMedia 4 (3), 12-20.

Snoek C. G. M and Worring M. (2005), 'Multimodal video indexing: A review of the state-of-the-art'. Multimedia Tools and Applications, 25(1): 5-35.

Snoek C. G., *et al.* (2006), 'The challenge problem for automated detection of 101 semantic concepts in multimedia', in Proceedings of ACM Multimedia, Santa Barbara, CA. pp. 421–430.

Snoek C. G. M. and Worring M. (2009), 'Concept-based video retrieval', F*oundations and Trends in Information Retrieval*, vol. 2, no. 4, pp. 215–322.

Sokhn M., *et al.* (2011), 'End-to-End Adaptive Framework for Multimedia Information Retrieval'. Ifip International Federation For Information Processing, pp.197-206.

Sommerville I. (2010), 'Formal Specification'. In: *Software Engineering. 9th ed*, Addison Wesley. Available at http://www.cs.st-andrews.ac.uk/~ifs/Books/SE9/WebChapters/PDF/Ch_27_Formal_spec.pdf

Spink A. and Cole C. (*Eds*), (2005), New Directions in Cognitive Information Retrieval. Springer, The Information Retrieval Series Volume 19, 2005, DOI:10.1007/1-4020-4014-8.

Su J., *et al.* (2011), 'Efficient Relevance Feedback for Content-Based Image Retrieval by Mining User Navigation Patterns'. IEEE Trans. Knowledge Data Enginerring 23(3): pp. 360-372.

Szummer M. and Picard R. W. (1998), 'Indoor-outdoor image classification', in *IEEE International Workshop on Content-Based Access of Image and Video Databases*, pp. 42 - 51.

Theodorakis M. and Spyratos N. (2002), 'Context in artificial intelligence and information modelling', *Proceedings of 2nd Hellenic Conference on Artificial Intelligence, SETN-2002*, companion volume pp. 27 - 38.

Theodorakis M., Analyti A., Constantopoulos P., and Spyratos N. (2002), 'A Theory of Context in Information Bases', *Information Systems*, Vol. 27, Issue 3, pp. 151-191.

The World Bank (2012), Information and Communications for Development 2012: Maximizing Mobile. Washington, DC: The World Bank, ISBN: 978-0821389911 .

Tintarev N. and Masthoff J. (2006), 'Similarity for News Recommender Systems'. In Proceedings of the AH'06 Workshop on Recommender Systems and Intelligent User Interfaces.

Tjondronegoro D. and Spink A. (2008), 'Web search engine multimedia functionality. Information Processing & Management', 44(1), pp. 340-357.

Torres J. M. and Parkes A. P. (2000), 'User modelling and adaptivity in visual information retrieval systems'. Workshop on *Computational Semiotics for New Media*, University of Surrey, Surrey, UK.

Tous R. and Delgado J. (2010), 'Semantic-Driven Multimedia Retrieval with the MPEG Query Format' , *Multimedia Tools and Applications*, Vol.49, 2010, pp. 213-233.

Turney P. D. and Pantel P. (2010), 'From Frequency to Meaning: Vector Space Models of Semantics.', Journal of Artifcial Intelligence Research (JAIR) 37, pp. 141-188.

Tzouvaras V., Troncy R., and Pan J. (2007), 'Multimedia Annotation Interoperability Framework'. WSC Incubator Group Editor's Draft. [Online] Available at: http://www.w3.org/2005/Incubator/mmsem/XGR-interoperability-20070814/ [Accessed 05 January 2012].

Ueda H., Miyatake T., and Yoshizawa S. (1991), 'An Interactive Natural-Motion-Picture Dedicated Multimedia Authoring System', In Proc. *ACM SIGCHI*, New Orleans, LA, pp. 343-350.

Ulrich F, (2006), 'Towards a Pluralistic Conception of Research Methods in Information Systems Research', Institut fur Informatik und Wirtschaftsinformatik Research Report No 7, December, University of Duisburg-Essen, Germany, available at http://www.icb.uni-due.de/fileadmin/ICB/research/research_reports/ICBReport07.pdf

Uren V., Cimiano P., Iria J., Handschuh S., Vargas-Vera M., Motta E., and Ciravegna F. (2005), 'Semantic annotation for knowledge management: Requirements and a survey of the state of the art', *Journal of Web Semantics*, pp. 14 – 28.

Vailaya A., Figueiredo M. A. T., Jain A. K, and Zhang H. (2001), 'Image classification for content-based indexing', IEEE Transactions on Image Processing 10 (1), pp. 117-130.

Vaishnavi, V. and Kuechler, W. (2004), 'Design Science Research in Information Systems' January 20, 2004, last updated September 30, 2011. URL: http://desrist.org/desrist

Varelas G., Voutsakis E., Raftopoulou P., Petrakis E., and Milios E. (2005), 'Semantic similarity methods in WordNet and their application to information retrieval on the web', in Proceedings of the 7th annual ACM international workshop on Web information and data management, pp. 10 - 16.

Venable J. (2006), A framework for Design Science research activities. In Khosrow-Pour M. (Ed.) Proceedings of the 2006 Information Resources Management Association. Idea Group, pp. 184-187.

Visser R., Sebe N., and Bakker E. M. (2002), 'Object Recognition for Video Retrieval' In Proceedings of International Conference on Image and Video Retrieval (CIVR'02), London, pp. 262-270.

Wang D. (1998), 'Unsupervised video segmentation based on watersheds and temporal tracking', *IEEE Trans. Circuits System. Video Technology*, Vol. 8, pp. 539-546.

Wang H., Divakaran A., Vetro A., Chang S-F., and Sun H. (2003), 'Survey of Compressed-Domain Features Used in Audio-Visual Indexing and Analysis', *Journal of Visual Communication and Image Representation*, Vol. 14, No. 2, pp 150-183.

Wang M. and Hua X. (2011), 'Active Learning in Multimedia Annotation and Retrieval: A Survey', *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 2.

Wiliem A., Madasu V., Boles W., and Yarlagadda P. (2012), 'A Context Space Model for Detecting Anomalous Behaviour in Video Surveillance'. In Proceedings of the 2012

Ninth International Conference on Information Technology - New Generations (ITNG '12). IEEE Computer Society, Washington, DC, USA, pp,18-24.

Wooldridge M. and Jennings N. R. (1995), 'Intelligent Agents: Theory and Practice', *The Knowledge Engineering Review*, Vol. 10 (2), pp. 115-152.

World Wide Web Consortium (W3C). (2009). 'OWL 2 Web Ontology Language Document Overview'. [Online] Available at: http://www.w3.org/TR/owl2-overview/ Retrieved January 18, 2011.

World Wide Web Consortium (W3C). (2011). 'W3C Semantic Web Activity'. [Online] Available at: http://www.w3.org/2001/sw/. Retrieved December 04, 2011.

Wu Z. and Palmer M. (1994), 'Verb semantics and lexical selection'. In 32nd Annual Meeting of the Association for Computational Linguistics, 133–138.

Wu Z. *et al.* (2012), 'GMQL: A graphical multimedia query language', Knowledge-Based Systems, Volume 26, pp. 135-143, ISSN 0950-7051.

Yang Y., Lovell B. C., and Dadgostar F. (2009), 'Content-Based Video Retrieval (CBVR) System for CCTV Surveillance Videos', in '*DICTA', IEEE Computer Society*, pp. 183-187.

Yazbek G., Mokbel C., and Chollet G. (2007). 'Video Segmentation and Compression using Hierarchies of Gaussian Mixture Models'. In the proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP), pp. 1009-1012.

Yi J., Peng Y. X., and Xiao J. G. (2012), 'A temporal context model for boosting video annotation'. Science China Information Sciences, pp. 1-14.

Yuan J. *et al.* (2007), 'A formal study of shot boundary detection', IEEE Transactions on Circuits and Systems for Video Technology 17 (2) (2007) pp. 168–186.

Zavesky E. and Chang S.-F. (2008), 'CuZero: embracing the frontier of interactive visual search for informed users', in Lew M. S., Del Bimbo A., and Bakker E. M. (Eds.), *Multimedia Information Retrieval* , ACM, pp. 237-244.

Zhai G. *et al.* (2005), 'eSports: Collaborative and Synchronous Video Annotation System in Grid Computing Environment'. In Proceedings of the Seventh IEEE International Symposium on Multimedia (ISM '05). IEEE Computer Society, Washington, DC, USA, pp. 95-103.

Zhai Y. and Liu B. (2005), 'Web data extraction based on partial tree alignment' In Proceedings of The 14th International Conference on World Wide Web, pp. 76 - 85.

Zhang H. J., Low C. Y., and Smoliar S. W. (1995), 'Video Parsing and Browsing using Compressed Data', *Multimedia Tools and Applications*, Vol. 1, pp. 91-113.

Zhang H. *et al.* (1993), 'Automatic Partitioning of Video', *Multimedia Systems*, Vol. 1, No 1, pp. 10-28.

Zhang Y. J. (2006), 'An Overview of Image and Video Segmentation in the Last 40 Years'. Advances in Image and Video Segmentation, pp.1-15.

Zhang Y. J. (2007), 'Toward high-level visual information retrieval' In: Zhang Y. J. (*ed*) *Semantic-based visual information retrieval*. IRM Press, Hershey.

Zhang Z., Khan H., and Robertson M. A. (2008), 'A Holistic, In-Compression Approach to Video Segmentation for Independent Motion Detection', EURASIP Journal on Advances in Signal Processing, vol. 2008, Article ID 738158.

Zhang Q. and Izquierdo E. (2011), 'Semantic Context Inference in Multimedia Search', in Domingue J. et al., (Eds.), *Future Internet Assembly*, Springer, pp. 391-402.

Zhu B. *et al.* (1999), 'Support Concept-based multimedia information retrieval: a knowledge management approach', Proceedings of *ICIS'99, 20th Annual International Conference on Information Systems*.

Zhuang Y., Rui Y., Huang T. S., and Mehrotra S. (1998), 'Adaptive key frame extraction using unsupervised clustering', In Proceedings of *ICIP*, Vol. 1, pp. 866–870.

Appendix A – Sample tokenised data with POS tag and lemmatized without stop words.

S/N	**D1 (POS Tag)**	**D2 (POS Tag)**	**D1 – Lemmatized**	**D2 – Lemmatized**
1	The (dt)	The (dt)		
	boy (n1)	boy (n1)	boy	boy
	does (vdz)	should (vmd)		
	not (xx)	be (vbi)		
	understand (vvi)	allowed (vvn)	understand	allow
	what (r-crq)	to (pc-acp)		
	you (pn22)	discover (vvi)		discover
	what (q-crq)	and (cc)		
	him (pno31)	ask (vvi)		
	to (pc-acp)	questions (n2)		question
	do (vdi)	himself (px31)		
		instead (av)		instead
		of (pp-f)		
		being (vbg)		
		asked (vvn)		

		what (r-crq)		
		he (pns31)		
		thinks (vvz)		
2	Communication (n1)	Inquisitive (j)	Communication	Inquisitive
	with (pp)	mind (n1)		mind
	concentration (n1)	of (pp-f)	concentration	
		communication (n1)		communicatio n
		and (cc)		
		gathering (j-vvg)		gather
		facts (n2)		
3	A (dt)	A (dt)		
	little (j)	little (j)	little	little
	boy (n1)	boy (n1)	boy	boy
	carried (vvd)	observed (vvd)	carry	observe
	out (av)	a (dt)		

	an (dt)	bug (n1)		bug
	action (n1)	entangled (vvn)	action	entangle
	and (cc)	a (dt)		
	was (vbd)	spider's (ng1)		spider
	trying (vvg)	web (n1)	try	web
	to (pc-acp)	and (cc)		
	explain (vvi)	wished (vvd)	explain	wish
	his (po31)	that (cst)		
	action (n1)	it (pn31)	action	
	as (c-acp)	could (vmd)		
	well (av)	free (vvi)		free
	as (c-acp)	itself (px31)		
	his (po31)	from (pp)		
	observations (n2)	the (dt)	observation	
		web (n1)		web
4	The (dt)	The (dt)		
	kid (n1)	kid's (ng1)	kid	kid
	is (vbz)	parents (n2)		parent

	being (n1)	are (vbb)		
	asked (vvn)	responding (vvg)		respond
	to (pc-acp)	to (p-acp)		
	identify (vvi)	his (po31)	identify	
	an (dt)	inquiries (n2)		inquiry
	object (n1)	as (c-acp)	object	
	in (p-acp)	regards (vvz)		regard
	the (dt)	a (dt)		
	darkened (vvn)	fly (n1)	darken	fly
	room (n1)	trapped (vvn)		trap
	by (p-acp)	in (p-acp)		
	someone (n1)	a (dt)		
	behind (p-acp)	spider (n1)		spider
	the (dt)	web (n1)		web
	camera (n1)	inside (n1-an)	camera	
		the (dt)		
		darkened (vvn)		darken
		room (n1)		

5	A (dt)	A (dt)		
	child (n1)	child (n1)	child	child
	repeating (vvg)	relating (vvg)	repeat	relate
	an (dt)	his (po31)		
	action (n1)	assurance (n1)	action	assurance
	he (pns31)	of (pp-f)		
	is (vbz)	protection (n1)		protection
	seen (vvn)	by (p-acp)		
	and (cc)	his (po31)		
	been (vbn)	mum (uh-j)		mum
	almost (av)	due (j-jn)		due
	told (vvn)	to (p-acp)	tell	
	reasons (n2)	past (j)	reason	
	of (pp-f)	event (n1)		event
	the (dt)	to (p-acp)		
	decision (n1)	the (dt)	decision	
	he (pns31)	bug (n1)		bug
	is (vbz)	in (p-acp)		
	about (p-acp)	the (dt)		

	to (pc-acp)	spiders (n2)		spider
	make (vvi)	web (n1)		web
6	The (dt)	A (dt)		
	boy (n1)	spider (n1)	boy	spider
	is (vbz)	is (vbz)		
	being (n1)	trapped (vvn)		trap
	asked (vvn)	in (p-acp)		
	to (pc-acp)	a (dt)		
	say (vvi)	cobweb (n1)		cobweb
	what (r-crq)	and (cc)		
	he (pns31)	the (dt)		
	is (vbz)	boy (n1)		boy
	seeing (vvg)	is (vbz)	see	
	by (p-acp)	asking (vvg)		
	some (d)	questions (n2)		question
	group (n1)	like (av-j)		
	of (pp-f)	why (c-crq)		
	people (n1)	its (po31)	people	

		not (xx)		
		flying (vvg)		fly
7	The (dt)	The (dt)		
	child (n1)	child (n1)	child	child
	is (vbz)	is (vbz)		
	trying (vvg)	trying (vvg)	try	try
	to (pc-acp)	to (pc-acp)		
	extinguish (vvi)	rescue (vvi)	extinguish	rescue
	the (dt)	some (d)		
	burning (j-vvg)	insect (n1)	burn	insect
	fire (n1)	trapped (vvn)	fire	trap
		in (p-acp)		
		spider (n1)		spider
		web (n1)		web
8	the (dt)	the (dt)		
	child (n1)	child (n1)	child	child

	is (vbz)	is (vbz)		
	trying (vvg)	trying (vvg)	try	try
	to (pc-acp)	to (pc-acp)		
	kill (vvi)	rescue (vvi)	kill	rescue
	the (dt)	the (dt)		
	insect (n1)	insect (n1)	insect	insect
		that (cst)		
		is (vbz)		
		trapped (vvn)		trap
		in (p-acp)		
		the (dt)		
		spider (n1)		spider
		web (n1)		web
9	The (dt)	The (dt)		
	scenario (n1)	scenario (n1)	scenario	Scenario
	of (pp-f)	of (pp-f)		
	a (dt)	a (dt)		
	child (n1)	child (n1)	child	child

	attempting (vvg)	attempting (vvg)	attempt	attempt
	to (p-acp)	to (p-acp)		
	spray (n1)	spray (n1)	spray	spray
	a (dt)	a (dt)		
	spot (n1)	bug (n1)	spot	bug
	on (p-acp)	trap (n1)		trap
	the (dt)	by (p-acp)		
	wall (n1)	a (dt)	wall	
	but (p-acp)	spider (n1)		spider
	didn't (vddx)	web (n1)		web
	know (vvi)	but (cc-acp)		
	how (c-crq)	was (vbd)		
	to (pc-acp)	stopped (vvn)		stop
		by (p-acp)		
		the (dt)		
		parents (n2)		parent
10	Scenario (n1)	Scenario (n1)	scenario	Scenario
	of (pp-f)	of (pp-f)		

	a (dt)	a (dt)		
	child (n1)	child (n1)	child	child
	attempting (vvg)	who (r-crq)	attempt	
	to (pc-acp)	is (vbz)		
	put (vvi)	concerned (vvn)		concern
	out (av)	about (p-acp)		
	a (dt)	an (dt)		
	fire (n1)	insect (n1)	fire	insect
		trapped (vvn)		trap
		in (p-acp)		
		a (dt)		
		spider (n1)		spider
		web (n1)		web
11	He (pns31)	He (pns31)		
	is (vbz)	is (vbz)		
	determined (vvn)	concerned (vvn)	determine	concern
	to (pc-acp)	about (p-acp)		
	seal (vvi)	the (dt)	seal	

	the (dt)	still (av)		
	black (j-jn)	trapped (vvn)	black	trap
	spot (n1)	spider (n1)	spot	spider
	on (p-acp)	on (p-acp)		
	the (dt)	the (dt)		
	window (n1)	window (n1)	window	window
	plane (n1)	plane (n1)	plane	plane
	with (pp)	and (cc)		
	the (dt)	wants (vvz)		
	spray (n1)	to (pc-acp)	spray	
		set (vvi)		set
		it (pn31)		
		free (j)		free
12	A (dt)	A (dt)		
	little (j)	little (j)	little	little
	boy (n1)	boy (n1)	boy	boy
	who (r-crq)	observes (vvz)		observe
	was (vbd)	a (dt)		

	outdoors (j)	spider (n1)	outdoors	spider
	and (cc)	trapped (vvn)		trap
	holding (vvg)	in (p-acp)	hold	
	his (po31)	a (dt)		
	toys (n2)	web (n1)	toy	web
	sees (vvz)	at (pp)		
	two (crd)	the (dt)		
	candles (n2)	window (n1)	candle	window
	burning (vvg)	pane (n1)	burn	pane
	inside (n1-an)	and (cc)		
	a (dt)	he (pns31)		
	room (n1)	is (vbz)		
	with (pp)	asked (vvn)		
	empty (j)	by (p-acp)	empty	
	chairs (n2)	an (dt)	chair	
	from (pp)	adult (n1)		adult
	the (dt)	probably (av-j)		probable
	window (n1)	his (po31)	window	
	he (pns31)	father (n1)		father

	is (vbz)	how (c-crq)		
	communicating (vvg)	the (dt)	communicate	
	his (po31)	spider (n1)		spider
	observations (n2)	can (vmb)	observation	
	to (p-acp)	get (vvi)		
	a (dt)	out (av)		
	third (ord)	his (po31)	third	
	party (n1)	answer (n1)	party	answer
	but (p-acp)	was (vbd)		
	he (pns31)	that (d)		
	appears (vvz)	the (dt)	appear	
	helpless (j)	spider's (ng1)	helpless	spider
		mum (uh-j)		mum
		will (vmb)		
		help (vvi)		help
		it (pn31)		
		out (av)		

13	A (dt)	A (dt)		
	Confused (vvn)	confused (j-vvn)	confuse	confuse
	boy (n1)	boy (n1)	boy	boy
	trying (vvg)	trying (vvg)	try	try
	to (pc-acp)	to (pc-acp)		
	get (vvi)	understand (vvi)		understand
	rid (vvn)	why (c-crq)	rid	
	of (pp-f)	the (dt)		
	an (dt)	insect (n1)		insect
	insect (n1)	is (vbz)	insect	
	but (cc-acp)	not (xx)		
	doesn't (vdzx)	moving (vvg)		move
	know (vvi)	despite (pp)		
	how (c-crq)	making (vvg)		
	to (pc-acp)	an (dt)		
	use (vvi)	attempt (n1)		attempt
	the (dt)	using (vvg)		use
	insecticide (n1)	the (dt)	insecticide	
		objects (n2)		object

		in (p-acp)		
		his (po31)		
		hand (n1)		hand
14	A (dt)	How (c-crq)		
	clip (vvi)	to (pc-acp)	clip	
	on (p-acp)	explain (vvi)		explain
	how (c-crq)	in (p-acp)		
	a (dt)	simple (j)		simple
	child (n1)	terms (n2)	child	term
	can (vmb)	the (dt)		
	be (vbi)	situation (n1)		situation
	guided (vvn)	at (pp)	guide	
	to (pc-acp)	hand (n1)		hand
	achieve (vvi)	making (vvg)	achieve	
	a (dt)	the (dt)		
	task (n1)	person (n1)	task	person
		in (p-acp)		
		question (n1)		question

		understand (vvi)		understand
		and (cc)		
		accepting (vvg)		accept
		what (r-crq)		
		you (pn22)		
		are (vbb)		
		saying (vvg)		say
15	A (dt)	A (dt)		
	boy (n1)	boy (n1)	boy	boy
	is (vbz)	is (vbz)		
	probably (av-j)	wondering (vvg)	probable	wonder
	being (vbg)	how (c-crq)		
	shown (vvn)	a (dt)	show	
	the (dt)	bug (n1)		bug
	effect (n1)	caught (vvn)	effect	catch
	of (pp-f)	in (p-acp)		
	an (dt)	a (dt)		
	action (n1)	spider (n1)	action	spider

	he's (pns31|vbz)	web (n1)		web
	been (vbn)	is (vbz)		
	told (vvn)	going (vvg)	tell	
	to (pc-acp)	to (pc-acp)		
	perform (vvi)	be (vbi)	perform	
		set (vvn)		set
		free (j)		free
16	Pesticide (n1)	Children (n2)	Pesticide	child
	should (vmd)	heart (n1)		heart
	not (xx)	are (vbb)		
	be (vbi)	innocent (j-jn)		innocent
	left (vvn)	free (j)	leave	free
	carelessly (av-j)	and (cc)	careless	
	where (c-crq)	compassionate (j)		compassionate
	children (n2)	to (p-acp)	child	
	are (vbb)	both (d)		
		human (j)		human
		and (cc)		

		animals (n2)		animal
		the (dt)		
		boy (n1)		boy
		felt (vvd)		
		the (dt)		
		baby (n1)		baby
		spider's (ng1)		spider
		should (vmd)		
		have (vhi)		
		a (dt)		
		mum (uh-j)		mum
		looking (vvg)		look
		after (c-acp)		
		it (pn31)		
17	A (dt)	A (dt)		
	child's (ng1)	child (n1)	child	child
	decision (n1)	is (vbz)	decision	
	is (vbz)	curious (j)		curious

	being (n1)	on (p-acp)		
	influenced (vvn)	what (r-crq)	influence	
	by (p-acp)	happens (vvz)		happen
	an (dt)	to (p-acp)		
	adult (n1)	a (dt)	adult	
	towards (pp)	spider (n1)		spider
	making (vvg)	cut-up (j)		Cut-up
	a (dt)	in (p-acp)		
	fun (n1)	a (dt)	fun	
	loving (vvg)	web (n1)	love	web
	act (n1)	and (cc)	act	
	while (cs)	engages (vvz)		engage
	having (vhg)	his (po31)		
	a (dt)	parents (n2)		parent
	nice (j)	in (p-acp)	nice	
	time (n1)	the (dt)	time	
		act (n1)		act
18	Little (j)	Little (j)	little	little

	boy (n1)	boy (n1)	boy	boy
	trying (vvg)	feels (vvz)	try	feel
	to (pc-acp)	empathy (n1)		empathy
	exterminate (vvi)	for (p-acp)	exterminate	
	a (dt)	bug (n1)		bug
	bug (n1)		bug	
19	The (dt)	In (p-acp)		
	little (j)	the (dt)	little	
	boy (n1)	observations (n2)	boy	observation
	wants (vvz)	of (pp-f)		
	to (pc-acp)	the (dt)		
	find (vvi)	child (n1)		child
	clarity (n1)	of (pp-f)	clarity	
	about (p-acp)	what (r-crq)		
	the (dt)	was (vbd)		
	spot (n1)	trapped (vvn)	spot	trap
	on (p-acp)	he (pns31)		
	the (dt)	believes (vvz)		believe

	wall (n1)	that (cst)	wall	
	he (pns31)	the (dt)		
	was (vbd)	bug (n1)		bug
	gazing (vvg)	can (vmb)	gaze	
	at (pp)	be (vbi)		
		rescued (vvn)		rescue
		but (cc-acp)		
		worried (vvn)		worry
		that (cst)		
		he (pns31)		
		could (vmd)		
		do (vdi)		
		nothing (pix)		
		in (p-acp)		
		that (d)		
		direction (n1)		direction
20	The (dt)	A (dt)		
	boy (n1)	fly (n1)	boy	fly

	sprayed (vvd)	was (vbd)	spray	
	something (pi)	caught (vvn)		catch
	on (p-acp)	in (p-acp)		
	the (dt)	the (dt)		
	wall (n1)	spider's (ng1)	wall	spider
	while (cs)	web (n1)		web
	wondering (vvg)	and (cc)	wonder	
	what (r-crq)	the (dt)		
	it (pn31)	boy (n1)		boy
	could (vmd)	was (vbd)		
	be (vbi)	wondering (vvg)		wonder
		how (c-crq)		
		it (pn31)		
		would (vmd)		
		be (vbi)		
		rescued (vvn)		rescue
21	The (dt)	On (p-acp)		
	little (j)	seeing (vvg)	little	see

	boy (n1)	the (dt)	boy	
	stares (vvz)	bug (n1)	stare	bug
	at (pp)	trapped (vvn)		trap
	a (dt)	in (p-acp)		
	particular (j)	a (dt)	particular	
	spot (n1)	web (n1)	spot	web
	on (p-acp)	the (dt)		
	the (dt)	little (j)		little
	wall (n1)	boy (n1)	wall	boy
	without (p-acp)	was (vbd)		
	a (dt)	worried (vvn)		worry
	clue (n1)	that (cst)	clue	
	of (pp-f)	he (pns31)		
	what (r-crq)	could (vmd)		
	to (pc-acp)	do (vdi)		
	make (vvi)	nothing (pix)		
	out (av)	but (cc-acp)		
	of (pp-f)	that (cst)		
	his (po31)	the (dt)		

	observing (vvg)	mummy (n1)	observe	mummy
	object (n1)	could (vmd)	object	
		do (vdi)		
		something (pi)		
22	Something (pi)	Observing (vvg)		observe
	caught (vvd)	that (cst)	catch	
	the (dt)	the (dt)		
	boy's (ng1)	web (n1)	boy	web
	attention (n1)	caught (vvd)	attention	catch
	but (cc-acp)	the (dt)		
	he (pns31)	bug (n1)		bug
	could (vmd)	the (dt)		
	not (xx)	little (j)		little
	make (vvi)	boy (n1)		boy
	out (av)	wondered (vvd)		wonder
	what (r-crq)	why (c-crq)		
	about (p-acp)	there (pc-acp)		
	it (pn31)	cannot (vmbx)		

		be (vbi)		
		any (d)		
		rescue (n1)		rescue
		for (p-acp)		
		the (dt)		
		bug (n1)		bug
		to (pc-acp)		
		release (vvi)		release
		itself (px31)		
		by (p-acp)		
		flying (vvg)		fly
		away (av)		
23	There (pc-acp)	The (dt)		
	boy (n1)	boy (n1)	boy	boy
	was (vbd)	was (vbd)		
	looking (vvg)	looking (vvg)	look	look
	at (pp)	for (p-acp)		
	a (dt)	a (dt)		

	dent (n1)	way (n1)	dent	
	on (p-acp)	to (pc-acp)		
	the (dt)	rescue (vvi)		rescue
	window (n1)	the (dt)	window	
		fly (n1)		fly
		from (pp)		
		the (dt)		
		spider's (ng1)		spider
		web (n1)		web
24	Perhaps (av)	Apparently (av-j)		apparent
	the (dt)	the (dt)		
	boy (n1)	boy (n1)	boy	boy
	picked (vvd)	is (vbz)	pick	
	the (dt)	experimenting (vvg)		experiment
	material (n-jn)	in (p-acp)	material	
	he (pns31)	experiencing (vvg)		experience
	held (vvd)	of (pp-f)	hold	

	on (p-acp)	what (r-crq)		
	hand (n1)	he (pns31)	hand	
	and (cc)	saw (vvd)		
	could (vmd)	held (vvn)		hold
	not (xx)	in (p-acp)		
	know (vvi)	a (dt)		
	what (r-crq)	web (n1)		web
	to (pc-acp)	on (p-acp)		
	do (vdi)	the (dt)		
	next (ord)	wall (n1)		wall
	whilst (cs)		whilst	
	staring (vvg)		stare	
	at (pp)			
	an (dt)			
	object (n1)		object	
	of (pp-f)			
	observation (n1)		observation	

Appendix B – NUPOS Word Classes and Parts of Speech

(Source: http://morphadorner.northwestern.edu/morphadorner/documentation/nupos/)

B1. Word Classes

In NUPOS, each word part has a "major word class" and a "word class". These concepts provide the coarsest ways to categorize words. There are 17 major word classes:

Major word classes
adjective
adv/conj/pcl/prep
adverb
conjunction
determiner
foreign word
interjection
negative
noun
numeral
preposition
pronoun
punctuation
symbol
undetermined
verb
wh-word

Major word classes are subdivided into a slightly finer categorization by "word class". There are 34 word classes in NUPOS:

Name	Description	Major Class
acp	adverb/conjunction/particle/preposition	adv/conj/pcl/prep
an	adverb/noun	noun
av	adverb	adverb
cc	coordinating conjunction	conjunction
crq	wh-word	wh-word
cs	subordinating conjunction	conjunction
d	determiner	determiner
dt	article	determiner
fo	foreign	foreign word
fr	French	foreign word
ge	German	foreign word
gr	Greek	foreign word
it	Italian	foreign word
j	adjective	adjective
jn	adjective/noun	adjective
jp	proper adjective	adjective
la	Latin	foreign word

n	noun	noun
np	proper noun	noun
nu	numeral	numeral
pf	preposition "of"	preposition
pi	indefinite pronoun	pronoun
pn	personal pronoun	pronoun
po	possessive pronoun	pronoun
pp	preposition	preposition
pu	punctuation	punctuation
px	reflexive pronoun	pronoun
sy	symbol	symbol
uh	interjection	interjection
v	verb	verb
va	auxiliary verb	verb
vm	modal verb	verb
xx	negative	negative
zz	undetermined	undetermined

B2. Parts of Speech

All of the NUPOS parts of speech are displayed below:

Tag	Explanation	Example
a-acp	acp word as adverb	I have not seen him since
av	adverb	soon
av-an	noun-adverb as adverb	go home
av-c	comparative adverb	sooner, rather
avc-jn	comparative adj/noun as adverb	deeper
av-d	determiner/adverb as adverb	more slowly
av-dc	comparative determiner/adverb as adverb	can lesser hide his love
av-ds	superlative determiner as adverb	most often
av-dx	negative determiner as adverb	no more
av-j	adjective as adverb	quickly
av-jc	comparative adjective as adverb	he fared worse
av-jn	adj/noun as adverb	duly, right honourable
av-js	superlative adjective as adverb	in you it best lies
av-n1	noun as adverb	had been cannibally given
av-s	superlative adverb	soonest
avs-jn	superlative adj/noun as adverb	hee being the worthylest constant
av-vvg	present participle as adverb	lovingly

av-vvn	past participle as adverb	Stands Macbeth thus amazedly
av-x	negative adverb	never
c-acp	acp word as conjunction	since I last saw him
cc	coordinating conjunction	and, or
cc-acp	acp word as coordinating conjunction	but
c-crq	wh-word as conjunction	when she saw
ccx	negative conjunction	nor
crd	numeral	2, two, ii
cs	subordinating conjunction	if
cst	'that' as conjunction	I saw that it was hopeless
d	determiner	that man, much money
dc	comparative determiner	less money
dg	determiner in possessive use	the latter's
ds	superlative determiner	most money
dt	article	a man, the man
dx	negative determiner as adverb	no money
fw-fr	French word	monsieur
fw-ge	German word	Herr
fw-gr	Greek word	kurios
fw-it	Italian word	signor
fw-la	Latin word	dominus

fw-mi	word in unspecified other language	n/a
j	adjective	beautiful
j-av	adverb as adjective	the then king
jc	comparative adjective	handsomer
jc-jn	comparative adj/noun	yet she much whiter
jc-vvg	present participles as comparative adjective	for what pleasinger then varietie, or sweeter then flatterie?
jc-vvn	past participle as comparative adjective	shall find curster than she
j-jn	adjective-noun	the sky is blue
jp	proper adjective	Athenian philosopher
js	superlative adjective	finest clothes
js-jn	superlative adj/noun	reddest hue
js-vvg	present participle as superlative adjective	the lyingest knave in Christendom
js-vvn	past participle as superlative adjective	deformed'st creature
j-vvg	present participle as adjective	loving lord
j-vvn	past participle as adjective	changed circumstances
n1	singular, noun	child
n1-an	noun-adverb as singular noun	my home
n1-j	adjective as singular noun	a good
n2	plural noun	children

n2-acp	acp word as plural noun	and many such-like "As'es" of great charge
n2-an	noun-adverb as plural noun	all our yesterdays
n2-av	adverb as plural noun	and are etcecteras no things
n2-dx	determiner/adverb negative as plural noun	yeas and honest kerysey noes
n2-j	adjective as plural noun	give me particulars
n2-jn	adj/noun as plural noun	the subjects of his substitute
n2-vdg	present participle as plural noun, 'do'	doings
n2-vhg	present participle as plural noun, 'have'	my present havings
n2-vvg	present participle as plural noun	the desperate languishings
n2-vvn	past participle as plural noun	there was no necessity of a Letter of Slains for Mutilation
ng1	singular possessive, noun	child's
ng1-an	noun-adverb in singular possessive use	Tomorrow's vengeance
ng1-j	adjective as possessive noun	the Eternal's wrath
ng1-jn	adj/noun as possessive noun	our sovereign's fall
ng1-vvn	past participle as possessive noun	the late lamented's house
ng2	plural possessive, noun	children's
ng2-jn	adj/noun as plural possessive noun	mortals' chiefest enemy
n-jn	adj/noun as noun	a deep blue
njp	proper adjective as noun	a Roman

njp2	proper adjective as plural noun	The Romans
njpg1	proper adjective as possessive noun	The Roman's courage
njpg2	proper adjective as plural possessive noun	The Romans' courage
np1	singular, proper noun	Paul
np2	plural, proper noun	The Nevils are thy subjects
npg1	singular possessive, proper noun	Paul's letter
npg2	plural possessive, proper noun	will take the Nevils' part
np-n1	singular noun as proper noun	at the Porpentine
np-n2	plural noun as proper noun	such Brooks are welcome to me
np-ng1	singular possessive noun as proper noun	and through Wall's chink
n-vdg	present participle as noun, 'do'	my doing
n-vhg	present participle as noun, 'have'	my having
n-vvg	present participle as noun	the running of the deer
n-vvn	past participle as noun	the departed
ord	ordinal number	fourth
p-acp	acp word as preposition	to my brother
pc-acp	acp word as particle	to do
pi	singular, indefinite pronoun	one, something
pi2	plural, indefinite pronoun	from wicked ones
pi2x	plural, indefinite pronoun	To hear my nothings monstered

pig	singular possessive, indefinite pronoun	the pairings of one's nail
pigx	possessive case, indefinite pronoun	nobody's
pix	indefinite pronoun	none, nothing
pn22	2nd person, personal pronoun	you
pn31	3rd singular, personal pronoun	it
png11	1st singular possessive, personal pronoun	a book of mine
png12	1st plural possessive, personal pronoun	this land of ours
png21	2nd singular possessive, personal pronoun	this is thine
png22	2nd person, possessive, personal pronoun	this is yours
png31	3rd singular possessive, personal pronoun	a cousin of his
png32	3rd plural possessive, personal pronoun	this is theirs
pno11	1st singular objective, personal pronoun	me
pno12	1st plural objective, personal pronoun	us
pno21	2nd singular objective, personal pronoun	thee
pno31	3rd singular objective, personal pronoun	him, her
pno32	3rd plural objective, personal pronoun	them
pns11	1st singular subjective, personal pronoun	I

pns12	1st plural subjective, personal pronoun	we
pns21	2nd singular subjective, personal pronoun	thou
pns31	3rd singular subjective, personal pronoun	he, she
pns32	3rd plural objective, personal pronoun	they
po11	1st singular, possessive pronoun	my
po12	1st plural, possessive pronoun	our
po21	2nd singular, possessive pronoun	thy
po22	2nd person possessive pronoun	your
po31	3rd singular, possessive pronoun	its, her, his
po32	3rd plural, possessive pronoun	their
pp	preposition	in
pp-f	preposition 'of'	of
px11	1st singular reflexive pronoun	myself
px12	1st plural reflexive pronoun	ourselves
px21	2nd singular reflexive pronoun	thyself, yourself
px22	2nd plural reflexive pronoun	yourselves
px31	3rd singular reflexive pronoun	herself, himself, itself
px32	3rd plural reflexive pronoun	themselves
pxg21	2nd singular possessive, reflexive pronoun	yourself's remembrance

q-crq	interrogative use, wh-word	Who? What? How?
r-crq	relative use, wh-word	the girl who ran
sy	alphabetical or other symbol	A, @
uh	interjection	oh!
uh-av	adverb as interjection	Well!
uh-crq	wh-word as interjection	Why, there were but four
uh-dx	negative interjection	No!
uh-j	adjective as interjection	Grumio, mum!
uh-jn	adjective/noun as interjection	And welcome, Somerset
uh-n	noun as interjection	Soldiers, adieu!
uh-v	verb as interjection	My gracious silence, hail
vb2	2nd singular present of 'be'	thou art
vb2-imp	2nd plural present imperative, 'be'	Beth pacient
vb2x	2nd singular present, 'be'	thow nart yit blisful
vbb	present tense, 'be'	are, be
vbbx	present tense negative, 'be'	aren't, ain't, beant
vbd	past tense, 'be'	was, were
vbd2	2nd singular past of 'be'	thou wast, thou wert
vbd2x	2nd singular past, 'be'	weren't
vbdp	plural past tense, 'be'	whose yuorie shoulders weren couered all

vbdx	past tense negative, 'be'	wasn't, weren't
vbg	present participle, 'be'	being
vbi	infinitive, 'be'	be
vbm	1st singular, 'be'	am
vbmx	1st singular negative, 'be'	I nam nat lief to gabbe
vbn	past participle, 'be'	been
vbp	plural present, 'be'	Thise arn the wordes
vbz	3rd singular present, 'be'	is
vbzx	3rd singular present negative, 'be'	isn't
vd2	2nd singular present of 'do'	dost
vd2-imp	2nd plural present imperative, 'do'	Dooth digne fruyt of Penitence
vd2x	2nd singular present negative, 'do'	thee dostna know the pints of a woman
vdb	present tense, 'do'	do
vdbx	present tense negative, 'do'	don't
vdd	past tense, 'do'	did
vdd2	2nd singular past of 'do'	didst
vdd2x	2nd singular past negative, verb	Why, thee thought'st Hetty war a ghost, didstna? 0.20
vddp	plural past tense, 'do'	on Job , whom that we diden wo
vddx	past tense negative, 'do'	didn't

vdg	present participle, 'do'	doing
vdi	infinitive, 'do'	to do
vdn	past participle, 'do'	done
vdp	plural present, 'do'	As freendes doon whan they been met
vdz	3rd singular present, 'do'	does
vdzx	3rd singular present negative, 'do'	doesn't
vh2	2nd singular present of 'have'	thou hast
vh2-imp	2nd plural present imperative, 'have'	O haveth of my deth pitee!
vh2x	2nd singular present negative, 'have'	hastna
vhb	present tense, 'have'	have
vhbx	present tense negative, 'have'	haven't
vhd	past tense, 'have'	had
vhd2	2nd singular past of 'have'	thou hadst
vhdp	plural past tense, 'have'	Of folkes that hadden grete fames
vhdx	past tense negative, 'have'	hadn't
vhg	present participle, 'have'	having
vhi	infinitive, 'have'	to have
vhn	past participle, 'have'	had
vhp	plural present, 'have'	They han of us no jurisdiccioun,
vhz	3rd singular present, 'have'	has, hath

vhzx	3rd singular present negative, 'have'	Ther loveth noon, that she nath why to pleyne.
vm2	2nd singular present of modal verb	wilt thou
vm2x	2nd singular present negative, modal verg	O deth, allas, why nyltow do me deye
vmb	present tense, modal verb	can, may, shall, will
vmb1	1st singular present, modal verb	Chill not let go, zir, without vurther 'cagion
vmbx	present tense negative, modal verb	cannot; won't; I nyl nat lye
vmd	past tense, modal verb	could, might, should, would
vmd2	2nd singular past of modal verb	couldst, shouldst, wouldst; how gret scorn woldestow han
vmd2x	2nd singular present, modal verb	Why noldest thow han writen of Alceste
vmdp	plural past tense, modal verb	tho thinges ne scholden nat han ben doon
vmdx	past negative, modal verb	couldn't; She nolde do that vileynye or synne
vmi	infinitive, modal verb	Criseyde shal nought konne knowen me.
vmn	past participle, modal verb	I had oones or twyes ycould
vmp	plural present tense, modal verg	and how ye schullen usen hem
vv2	2nd singular present of verb	thou knowest

vv2-imp	2nd present imperative, verb	For, sire and dame, trusteth me right weel,
vv2x	2nd singular present negative, verb	"Yee!" seyde he, "thow nost what thow menest;
vvb	present tense, verg	they live
vvbx	present tense negative, verb	What shall I don? For certes, I not how
vvd	past tense, verb	knew
vvd2	2nd singular past of verb	knewest
vvd2x	2nd singular past negative, verb	thou seidest that thou nystist nat
vvdp	past plural, verb	They neuer strouen to be chiefe
vvdx	past tense negative, verb	she caredna to gang into the stable
vvg	present participle, verb	knowing
vvi	infinitive, verb	to know
vvn	past participle, verb	known
vvp	plural present, verb	Those faytours little regarden their charge
vvz	3rd singular preseent, verb	knows
vvzx	3rd singular present negative, verb	She caresna for Seth.
xx	negative	not
zz	unknown or unparsable token	n/a

Appendix C – List of Public Output

Eze E., Ishaya T., and Wood D. (2007), 'Contextualizing Multimedia Semantics towards Personalised eLearning', Proceedings of the 4th Special Workshop on Multimedia Semantics, Chania, Crete, Greece. Published in Journal of Digital Information Management, Vol. 5, Issue 2.

Eze E. and Ishaya T. (2007), 'S-Space: a Context-based Ontology Model for Multimedia Semantic Organisation and Discovery'. In proceedings of 1st Workshop on Multimedia Annotation and Retrieval enabled by Shared Ontologies, Genova, Italy.

Ishaya T. and Eze E. (2007), 'Context-Based Multimedia Ontology Model', ICAS, pp.2, Third International Conference on Autonomic and Autonomous Systems (ICAS'07).

Huang W., Eze E., and Webster D. (2006), 'Applying context awareness in learner-centric intelligent e-Learning services towards the semantic web'. Book Chapter in 'E-Service Intelligence - Methodologies, Technologies and Applications', Computational Intelligence Edited by Jie Lu, Da Ruan, Guangquan Zhang, Springer.

Huang W., Eze E., and Webster D. (2006), 'Towards Integrating Semantics of Multi-Media Resources and Processes in e-Learning', *ACM/Springer Journal on Multimedia Systems, Special Issue on Educational Multimedia Systems*, Vol. 11, No 3, pp. 203-215.

Huang W. and Eze E. (2005), 'Multi-Media Semantics Contextualisation for Knowledge-Orientated e-Learning', Proceedings of *ICALT 2005 (The 5th IEEE International Conference on Advanced Learning Technologies)*, Kaohsiung, Taiwan, pp. 623 - 625.